DUAL TRAINING OF ENERGY-BASED MODELS WITH OVERPARAMETRIZED NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Energy-based models (EBMs) are generative models that are usually trained via maximum likelihood estimation. This approach becomes challenging in generic situations where the trained energy is nonconvex, due to the need to sample the Gibbs distribution associated with this energy. Using general Fenchel duality results, we derive variational principles dual to maximum likelihood EBMs with shallow overparametrized neural network energies, both in the active (aka feature-learning) and lazy regimes. In the active regime, this dual formulation leads to a training algorithm in which one updates concurrently the particles in the sample space and the neurons in the parameter space of the energy at a faster rate. We also consider a variant of this algorithm in which the particles are sometimes restarted at random samples drawn from the data set, and show that performing these restarts at every iteration step corresponds to score matching training. Using intermediate parameter setups in our dual algorithm thereby gives a way to interpolate between maximum likelihood and score matching training. These results are illustrated in simple numerical experiments.

1 INTRODUCTION

Energy-based models (EBMs) are explicit generative models which consider Gibbs measures defined through an *energy function* f, with a probability density proportional to $\exp(-\beta f(x))$, where β is the inverse temperature. Such models originate in statistical physics (Gibbs, 2010; Ruelle, 1969), and have become a fundamental modeling tool in statistics and machine learning (Wainwright & Jordan, 2008; Ranzato et al., 2007; LeCun et al., 2006; Du & Mordatch, 2019; Song & Kingma, 2021). Given data samples from a target distribution, the learning algorithms for EBMs attempt to estimate an energy function f to model the samples density. The resulting learned model can then be used to obtain new samples, typically through Markov Chain Monte Carlo (MCMC) techniques.

The standard method to train EBMs is maximum likelihood estimation, i.e. the learned energy is the one maximizing the likelihood of the target samples, within a certain function class. One generic approach for this is to use gradient descent, where gradients may be approximated using MCMC samples from the trained model. However, this is computationally difficult for highly non-convex trained energies, due to 'metastability', ie the presence of large basins in the energy landscape that trap trajectories for potentially exponential time. This has motivated a myriad of alternative losses to learn EBM energies, such as the popular score matching; see (Song & Kingma, 2021) for a review. All in all, such weaker losses come at the expense of a loss of statistical power, which motivates exploring computationally efficient methods for EBM maximum-likelihood estimation.

EBMs also have structural connections with maximum entropy (maxent) models, which have been studied for decades through Fenchel duality. Dai et al. (2019b) was the first work to leverage similar duality arguments for maximum likelihood EBM training. However, their analysis is restricted to energies lying in RKHS balls (i.e. non-parametric linear models). Despite the appealing optimization properties of RKHS, these spaces of functions typically only contain very smooth functions when the dimension is large (Berlinet & Thomas-Agnan, 2004). A recent line of work—originating in supervised learning—has considered an alternative based on shallow neural networks (Bach, 2017),

which admit a linear representation in terms of a measure over its parameters and are able to adapt to hidden low-dimensional structures in the data. The statistical benefits of the obtained $\mathcal{F}_1$ or *Barron* spaces have recently been studied in the context of shallow EBMs by Domingo-Enrich et al. (2021), who show that they may outperform the RKHS models.

In this work, we focus instead on the computational aspects of training such shallow EBMs. Relying on infinite-dimensional Fenchel duality results for the KL-regularized L^2 and L^∞ regression problems over probability measures (App. B), we recast the maximum likelihood training of $\mathcal{F}_1$-EBMs into a min-max problem over measures, and derive a gradient descent-ascent algorithm (Alg. 1) in the associated metric spaces, based on the Wasserstein distance. Crucially, these schemes, defined over idealised parametrisations requiring infinite number of neurons, admit a finite-particle approximation, as in regression or classification. When viewed in terms of particle systems, the dynamics evolve two interacting populations simultaneously: one over the neuron parameters, and the other over the data space (Sec. 3). Moreover, our proposed algorithm naturally interpolates between the primal and dual formualtions, thanks to the relative time-scale between the minimization and maximization steps, and is even able to interpolate between MLE and Score Matching. This dual algorithm is evaluated experimentally in Sec. 5 in a well-calibrated high-dimensional teacher-student environment, which allows us to assess our models against the ground-truth, and test the effect of input dimension. Our experiments confirm that the dual algorithm converges significatively faster than the primal one, suggesting that dual updates might bypass metastability despite high-dimensional and non-convex energy landscapes.

Related work. Our work is based on general Fenchel duality results (App. B) that may be useful in applications beyond the main focus of this paper (see App. D). These theorems are a generalization of results stated in the compact case in Domingo-Enrich et al. (2021) in their Appendix D. Similar duality results have been studied extensively in the area of maximum entropy (maxent) models (reviewed in Ch. 12 of Mohri et al. (2012)). The first maxent duality principle was due to Jaynes (1957). Maxent models have been applied since the 1990s in natural language processing and in species habitat modeling among others, and studied theoretically especially since the 2000s (Altun & Smola, 2006; Dudík et al., 2007).

Recently Dai et al. (2019a) leveraged duality arguments in the context of maximum likelihood EBMs, although in a form different from ours. Their duality result works in the more restrictive setting of "lazy" energies lying in RKHS balls and probability measures with L^2 densities, and they derive it directly from a general theorem that works for reflexive Banach spaces (Ekeland & Temam (1999), Ch. 6, Thm. 2.1). Our Fenchel duality results, which work for Borel probability measures and feature-learning ($\mathcal{F}_1$) energies, are more general because we must rely on measure spaces, which are non-reflexive Banach spaces. Their algorithm is also different: they do not evolve generated samples, but rather use a transport parametrization of the energy. Dai et al. (2019b) expand the work (Dai et al., 2019a) combining it with Hamiltonian Monte Carlo.

A precursor of modern machine learning EBMs were restricted Boltzmann machines (RBMs), first trained via contrastive divergence or CD (Hinton, 2002) - which estimates the gradient of the log-likelihood via approximate MCMC samples of the trained model. It later led to maximum likelihood training of EBMs (see e.g. Xie et al. (2016; 2017); Du & Mordatch (2019) among many others). A popular variant of CD is persistent contrastive divergence or PCD (Tieleman, 2008; Tieleman & Hinton, 2009), in which the MCMC samples are evolved and reused over gradient computations to be progressively equilibrated. Drawing a comparison with our work, our dual $\mathcal{F}_1$-EBM training algorithm resembles PCD in that both evolve a set of samples over training iterations.

A vast array of EBM losses alternative to maximum likelihood have been developed in recent years (Song & Kingma, 2021) with the goal of avoiding the MCMC procedure, which may be costly for non-convex densities. Some successful ones are score matching (Hyvärinen, 2005) and related methods such as denoising score matching (Vincent, 2011). Building on these, recent works have achieved state of the art image generation (Song & Ermon, 2019; 2020; Ho et al., 2020; Song et al., 2021). We derive the score matching algorithm for $\mathcal{F}_1$ energies and show a continuum of algorithms interpolating between dual maximum likelihood and score matching training.

Finally, our work (in particular App. C) also has links with maximum mean discrepancy (MMD) flows. MMDs are probability metrics that were first introduced in (Gretton et al., 2007; 2012) for kernel two-sample tests, and that have enjoyed ample success with the advent of deep-learning-based generative modeling as discriminating metrics (Li et al., 2015; Dziugaite et al., 2015; Li et al., 2017). Among the MMD literature, the closest work to ours is (Arbel et al., 2019), which study theoretically the convergence of unregularized MMD gradient flow (our equation (26) with $\tilde{\beta}^{-1} = 0$). In their experiments, they observe that noisy updates ($\tilde{\beta}^{-1} > 0$) are needed for good generalization. Our work shows that their algorithm is exactly training maximum likelihood EBMs energies in an RKHS ball of radius that depends on the noise level (see App. C).

2 Background and setup

In this section, we provide preliminary background on the functional spaces associated to over-parametrized two-layer networks, and on EBMs and their training losses.

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 Radon (i.e. signed and finite) measures, and $\mathcal{M}^+(K)$ is the set of non-negative Radon measures. If γ is a Radon measure over K, then $\|\gamma\|_{\text{TV}} = \int_K d|\gamma|$ is the total variation (TV) norm of γ, which turns $\mathcal{M}(K)$ into a Banach space. Throughout the paper, and unless otherwise specified, $\sigma : \mathbb{R} \to \mathbb{R}$ denotes a generic non-linear activation function. We use $(\cdot)_+ : \mathbb{R} \to \mathbb{R}$ to denote the ReLU activation, defined as $(z)_+ = \max\{z, 0\}$. We use τ to denote a fixed base probability measure, possibly used with a subindex to specify the space it is defined over. We use $\mathbb{S}^d \subseteq \mathbb{R}^{d+1}$ for the d-dimensional hypersphere and $\log$ for the natural logarithm. We denote the Lebesgue measure by λ. Given two probability measures $\nu, \nu' \in \mathcal{P}(K)$, $D_{\text{KL}}(\nu\|\nu') = \int_K \log \frac{d\nu}{d\nu'}\, d\nu$ denotes the KL divergence from ν' to ν and $H(\nu, \nu') = -\int_K \log(\frac{d\nu'}{d\tau}) d\nu$ is the cross-entropy.

2.1 Overparametrized two-layer neural network spaces

In this work, we will focus on dense function approximation classes generated by overparametrized shallow neural networks. One can distinguish two canonical models, depending on the asymptotic scaling regime. For further background on these regimes, we refer the reader to Chizat et al. (2019).

Kernel regime. Let $\mathcal{X} \subseteq \mathbb{R}^{d_1}$, $\Theta \subseteq \mathbb{R}^{d_2}$, $\varphi : \mathcal{X} \times \Theta \to \mathbb{R}$, and τ_Θ be a fixed base probability measure over Θ. We define $\mathcal{F}_2$ as the reproducing kernel Hilbert space (RKHS) of functions $f : \mathcal{X} \to \mathbb{R}$ such that for some $h \in L^2(\Theta, \tau_\Theta)$, we have that, for all $x \in \mathcal{X}$, $f(x) = \int_\Theta \varphi(x, \theta) h(\theta) d\tau_\Theta(\theta)$. The RKHS norm of $\mathcal{F}_2$ is defined as $\|f\|_{\mathcal{F}_2} = \inf \left\{ \|h\|_{L^2(\Theta)} \mid f(\cdot) = \int_\Theta \varphi(\cdot, \theta)\, h(\theta)\, d\tau_\Theta(\theta) \right\}$ where $\|h\|^2_{L^2(\Theta)} := \int_\Theta |h(\theta)|^2\, d\tau_\Theta(\theta)$ (c.f. Bach (2017)). As an RKHS, the kernel of $\mathcal{F}_2$ is

$$k(x, y) = \int_\Theta \varphi(x, \theta)\varphi(x, \theta)\, d\tau_\Theta(\theta). \tag{1}$$

Feature-learning regime. Set $\mathcal{X}$, Θ and φ as in the previous paragraph, and define $\mathcal{F}_1$ as the Banach space of functions $f : \mathcal{X} \to \mathbb{R}$ such that, for some Radon measure $\gamma \in \mathcal{M}(\Theta)$, for all $x \in \mathcal{X}$ we have $f(x) = \int_\Theta \varphi(x, \theta)\, d\gamma(\theta)$. We define the norm of $\mathcal{F}_1$ as $\|f\|_{\mathcal{F}_1} = \inf \left\{ \|\gamma\|_{\text{TV}} \, | f(\cdot) = \int_\Theta \varphi(\cdot, \theta)\, d\gamma(\theta) \right\}$. This construction was introduced by Bach (2017), who first used the notation $\mathcal{F}_1$ and focused in particular on the case $\mathcal{X} \subseteq \mathbb{R}^d$, $\Theta = \mathbb{S}^d$ and $\varphi(x, \theta) = \text{ReLu}^k(\langle (x, 1), \theta \rangle)$ for some $k \in \mathbb{Z}_+$. This space is also known by the name of Barron space (E et al., 2019; E & Wojtowytsch, 2020) in reference to the classic work (Barron, 1993).

Remark that since $\|h\|_{L^1(\Theta)} = \int_\Theta |h(\theta)|\, d\tau_\Theta(\theta) \leq (\int_\Theta |h(\theta)|^2\, d\tau_\Theta(\theta))^{1/2} = \|h\|_{L^2(\Theta)}$ by the Cauchy-Schwarz inequality, we have $\mathcal{F}_2 \subset \mathcal{F}_1$: in particular finite-width neural networks belong

to $\mathcal{F}_1$ but not to $\mathcal{F}_2$ (Bach, 2017). 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.

2.2 EBMs AND TRAINING LOSSES

Consider a measurable set $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ with a fixed base probability measure $\tau_{\mathcal{X}} \in \mathcal{P}(\mathcal{X})$. If $\mathcal{F}$ is a class of functions (or energies) mapping $\mathcal{X}$ to $\mathbb{R}$, for any $f \in \mathcal{F}$ we can define the probability measure ν_f as a Gibbs measure with density:

$$\frac{d\nu_{\beta f}}{d\tau_{\mathcal{X}}}(x) := Z_{\beta f}^{-1} e^{-\beta f(x)} \qquad \text{with} \qquad Z_{\beta f} := \int_{\mathcal{X}} e^{-\beta f(y)} d\tau_{\mathcal{X}}(y) \,,$$

where $d\nu_{\beta f}/d\tau_{\mathcal{X}}$ is the Radon-Nikodym derivative of $\nu_{\beta f}$ and $Z_{\beta f}$ is the partition function. The parameter $\beta > 0$ is the inverse temperature. We could merge β into $\mathcal{F}$ by considering the function class $\{\beta f | f \in \mathcal{F}\}$, but we have decided to keep them separate to showcase the dependency on β. Gibbs measures are the cornerstone of statistical physics since the seminal works of Boltzmann and Gibbs. Beyond their widespread use across computational sciences, they have also found their application in Machine Learning, by the name of *energy-based models* (EBMs), where the energy function is parametrised using e.g. a neural network.

We denote by $\mathcal{F}_1$-EBMs the energy-based models for which the energy class $\mathcal{F}$ is the unit ball $\mathcal{B}_{\mathcal{F}_1}(1)$ of $\mathcal{F}_1$. Notice that the class $\{\beta f | f \in \mathcal{F}\}$ is equal to the ball $\mathcal{B}_{\mathcal{F}_1}(\beta)$. Such models may be regarded as abstractions of more complex deep EBMs, in that they incorporate feature learning, and they were first studied by Domingo-Enrich et al. (2021), which provide statistical guarantees. They are to be contrasted with $\mathcal{F}_2$-EBMs, for which $\mathcal{F}$ is the unit ball $\mathcal{B}_{\mathcal{F}_2}(1)$. $\mathcal{F}_2$-EBMs, which we study in App. C, have fixed features and showed worse statistical performance in experiments (Domingo-Enrich et al., 2021).

Given samples $\{x_i\}_{i=1}^n$ from a target measure ν_p, training an EBM consists in selecting the best $\nu_{\beta f}$ with energy $f \in \mathcal{F}$ according to a given criterion. Two such criteria are relevant in this paper.

Maximum likelihood. The maximum likelihood estimator (MLE) is defined as $\hat{f} = \arg\max_{f \in \mathcal{F}} \prod_{i=1}^n \frac{d\nu_{\beta f}}{d\tau_{\mathcal{X}}}(x_i)$, or, equivalently, as the minimizer of the cross-entropy $H(\nu_n, \nu_{\beta f})$ with the empirical measure $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} -\frac{1}{n}\sum_{i=1}^n \log\left(\frac{d\nu_{\beta f}}{d\tau_{\mathcal{X}}}(x_i)\right) = \arg\min_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(x_i) + \beta^{-1}\log Z_{\beta f}. \tag{2}$$

The estimated distribution is then simply given by $d\nu_{\beta\hat{f}} = Z_{\beta\hat{f}}^{-1} e^{-\beta\hat{f}} d\tau_{\mathcal{X}}$. By observing that $D_{\mathrm{KL}}(\nu||\nu') = H(\nu,\nu') - H(\nu)$, where $H(\nu) := H(\nu,\nu)$ is the entropy, minimizing the cross-entropy is equivalent to minimizing the KL divergence when the latter is finite. However, such equivalence is only well-defined in the *population* setting where the empirical measure ν_n is replaced by its population counterpart ν_p. An appropriate choice of function class $\mathcal{F}$ induces a regularization that prevents the learned Gibbs measure to overfit to the empirical data measure, and presumably approximate ν_p instead. The MLE enjoys strong statistical properties (Wainwright & Jordan, 2008; Wainwright, 2019), as well as a powerful variational principle as soon as one considers convex function classes (see App. A), but its notorious computational challenges (see Related works section) have motivated alternative approximation metrics to be used to learn EBMs.

Since an arbitrary element f of $\mathcal{F}_1$ can be expressed as $f(x) = \int_\Theta \varphi(x,\theta)\, d\gamma(\theta)$, with $\|f\|_{\mathcal{F}_1}$ equal to the infimum of $\|\gamma\|_{\mathrm{TV}}$ for all such γ, the maximum likelihood function for $\mathcal{F} = \mathcal{B}_{\mathcal{F}_1}(1)$ the problem (2) can be restated as $f_{\mathrm{MLE}} = \int_\Omega \varphi(\cdot,\theta) d\gamma_{\mathrm{MLE}}(\theta)$ where

$$\gamma_{\mathrm{MLE}} = \arg\min_{\substack{\gamma \in \mathcal{M}(\Theta) \\ \|\gamma\|_{\mathrm{TV}} \leq 1}} \frac{1}{n}\sum_{i=1}^n \int_\Theta \varphi(x_i,\theta)\, d\gamma(\theta) + \frac{1}{\beta}\log\left(\int_{\mathcal{X}} \exp\left(-\beta \int_\Theta \varphi(x,\theta)\, d\gamma(\theta)\right) d\tau_{\mathcal{X}}(x)\right) \tag{3}$$

Score matching. An important instance of such weaker metrics is given by Score Matching (SM). The SM metric between two absolutely continuous probability measures ν, ν' is defined as $\mathrm{SM}(\nu, \nu') = \int_{\mathcal{X}} |\nabla \log \frac{d\nu}{d\tau_{\mathcal{X}}}(x) - \nabla \log \frac{d\nu'}{d\tau_{\mathcal{X}}}(x)|^2 \, d\nu(x)$. The SM metric is known in information theory as the relative Fisher information. Note that this metric cannot be trivially extended to the case $\nu = \nu_n$, because empirical measures do not have a density with respect to $\tau_{\mathcal{X}}$. To get around this difficulty, note that, if the target measure ν_p is absolutely continuous with respect to $\tau_{\mathcal{X}}$ and we denote by $f_p(x) = -\beta^{-1} \log \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x)$ its energy, learning an EBM with function class $\mathcal{F}$ under the population loss corresponds to solving $\hat{f} = \arg\min_{f \in \mathcal{F}} \int_{\mathcal{X}} |\nabla f(x) - \nabla f_p(x)|^2 \, d\nu_p(x)$. The insight from Hyvärinen (2005) is that under regularity conditions on f_p, via integration by parts we then have $\int_{\mathcal{X}} |\nabla f - \nabla f_p|^2 \, d\nu_p = \mathbb{E}_{\{x_i\}_{i=1}^n} L(f, \nu_n) + C$, where $\mathbb{E}_{\{x_i\}_{i=1}^n}$ denotes expectation over the data set, C is a constant in f which is therefore irrelevant, and

$$L(f, \nu_n) = \frac{1}{n} \sum_{i=1}^{n} \beta^{-1} \Delta f(x_i) + \frac{1}{2} |\nabla f(x_i)|^2.$$

In practice, we train an EBM via score matching by solving $\hat{f} = \arg\min_{f \in \mathcal{F}} L(f, \nu_n)$. Score matching is computationally more tractable than maximum likelihood and performs well in practice (see Related works section). Statistically, its main drawback is that the SM metric is weaker than the KL divergence, and may fail to distinguish distributions in some instances.

The following proposition, proved in App. G, provides the expression for the loss L and the resulting score matching problem for $\mathcal{F}_1$-EBMs.

Proposition 1. *Suppose that $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ is a manifold without boundaries. Assume that $\int_{\mathcal{X}} |\nabla_x \varphi(x, \theta) \cdot \nabla \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x)| \, d\tau_{\mathcal{X}}(x)$ is upper-bounded by some constant K for all $\theta \in \Theta$. Assume also that $\sup_{\theta \in \Theta} \|\nabla_x \varphi(x, \theta)\| < \eta(x)$ and that $\int_{\mathcal{X}} |\eta(x)|^2 \, d\nu_p(x) < \infty$. The optimization problem to train EBMs under the score matching loss over the ball $\mathcal{B}_{\mathcal{F}_1}(1)$ gives $f_{\mathrm{SM}} = \int_{\Omega} \varphi(\cdot, \theta) d\gamma_{\mathrm{SM}}(\theta)$ where*

$$\gamma_{\mathrm{SM}} = \underset{\substack{\gamma \in \mathcal{M}(\Theta) \\ \|\gamma\|_{TV} \leq 1}}{\arg\min} \int_{\Theta} \int_{\mathcal{X}} \left(\frac{1}{2} \nabla_x \varphi(x, \theta) \cdot \nabla_x \int_{\Theta} \varphi(x, \theta') \, d\gamma(\theta') - \beta^{-1} \Delta_x \varphi(x, \theta) \right) d\nu_n(x) d\gamma(\theta) \quad (4)$$

3 Dual $\mathcal{F}_1$-EBM training via maximum likelihood

As a corollary of the duality result from Subsec. B.2, we derive an alternative objective for $\mathcal{F}_1$-EBMs trained via maximum likelihood, the original objective being (3) and we develop an algorithm to solve this alternative problem. To this end, we make:

Assumption 1. *Let $\varphi : \mathcal{X} \times \Theta \to \mathbb{R}$ be a continuous function such that either $\mathcal{X}$ is compact or (i) for any fixed $\theta \in \Theta$, $\varphi(x, \theta) \leq \xi(x)$ for some strictly positive $\xi : \mathcal{X} \to \mathbb{R}$, and (ii) $\xi(x) + \log(\xi(x)) = o\left(-\log\left(\frac{d\tau_{\mathcal{X}}}{d\lambda}(x) \right) - (d_1 + \epsilon) \log \|x\|_2 \right)$ as $\|x\|_2 \to +\infty$ for some $\epsilon > 0$.*

In particular, this assumption holds for ReLU network energies when setting $\mathcal{X} = \mathbb{R}^{d_1}$, $\Theta = \mathbb{R}^{d_1 + 1}$, $\varphi(x, \theta) = \sigma(\langle (x, 1), \theta \rangle) / \|\theta\|$ and $\tau_{\mathcal{X}}$ Gaussian (and in many other settings).

Theorem 1. *Under Assumption 1, the problem* (3) *is the Fenchel dual of*

$$\min_{\nu \in \mathcal{P}(\mathcal{X})} \max_{\substack{\gamma \in \mathcal{M}(\Theta), \\ \|\gamma\|_{TV} \leq 1}} \beta^{-1} D_{KL}(\nu || \tau_{\mathcal{X}}) + \int_{\Theta} \int_{\mathcal{X}} \varphi(x, \theta) \, d(\nu - \nu_n)(x) \, d\gamma(\theta). \quad (5)$$

Moreover, the solution $\nu^\star$ of (70) *is precisely the Gibbs measure for the optimal $\gamma^\star$ in* (3)*, that is, $\frac{d\nu^\star}{d\tau_{\mathcal{X}}}(x) = \frac{1}{Z_\beta} \exp\left(-\beta \int_{\Theta} \varphi(x, \theta) \, d\gamma^\star(x) \right)$.*

If we replace the $\mathcal{F}_1$ ball by the $\mathcal{F}_2$ ball, the analogous duality result links the maximum likelihood problem with the entropy regularized MMD flow from Arbel et al. (2019) (see App. C).

Training dynamics on nonnegative measures. Let us write the dynamics to solve (5) that can be discretized in terms of parameters and particles (cf Proposition 2 below). To this end we consider the triple $(\gamma^+, \gamma^-, \nu)$ where the nonegative measures $\gamma^\pm$ are defined through the Hahn decomposition of $\gamma = \gamma_+ - \gamma_-$. Then we introduce coupled gradient flows for this triple, in which γ_t^+ and γ_t^- evolve via a Wasserstein-Fisher-Rao gradient flow (Chizat et al., 2018) and ν_t evolves via a Wasserstein gradient flow (Santambrogio, 2017):

$$\begin{aligned} \partial_t \gamma_t^\sigma &= -\alpha\sigma \nabla_\theta \cdot \left(\gamma_t^\sigma \nabla_\theta F_t(\theta)\right) + \alpha \gamma_t^\sigma \left(\sigma F_t(\theta) - K_t\right), \qquad \sigma = \pm 1, \;\; \gamma_t^\sigma = \gamma_t^\pm \\ \partial_t \nu_t &= \nabla_x \cdot \left(\nu_t \left(\nabla_x f_t(x) - \beta^{-1} \nabla \log \frac{d\tau_{\mathcal{X}}}{d\lambda}\right)\right) + \beta^{-1} \Delta_x \nu_t, \end{aligned} \tag{6}$$

where α is a tunable parameter and we defined

$$\begin{aligned} F_t(\theta) = \int_{\mathcal{X}} \varphi(x,\theta)\, d(\nu_t - \nu_n)(x), \quad f_t(x) = \int_\Theta \varphi(x,\theta)\, (d\gamma_t^+ - d\gamma_t^-)(\theta), \\ K_t = \mathbb{1}_{\|\gamma_t^+\|_{\mathrm{TV}} + \|\gamma_t^-\|_{\mathrm{TV}} \geq 1} \int_\Theta F_t(\theta)(d\gamma_t^+ - d\gamma_t^-)(\theta). \end{aligned} \tag{7}$$

The initialization of (6) is $\nu_0 = \nu_n$ and $\gamma_0^\pm = 0$ (such that the initial energy is null). The term K_t keeps the total variation of γ_t below one. The parameter α acts as a relative timescale. Notice that different values of α can potentially lead to different behaviors of the dynamics; setting $\alpha \ll 1$ would correspond to the primal formulation of maximum likelihood with persistent MCMC samples (as in PCD). In contrast if $\alpha \gg 1$, $\gamma_t^\pm$ evolves faster than ν_t and if the optimization is well behaved at all times $\gamma_t = \gamma_t^+ - \gamma_t^-$ remains close to minimizing the inner maximization problem of (5) with $\gamma = \gamma_t$. Initializing $\nu_0 = \nu_n$ is crucial to avoid the kind of metastabilities that curse the behavior of classical (primal) maximum likelihood EBM training.

Proposition 2 below states that the solution (μ_t, ν_t) may be approximated using coupled particle systems (see proof in App. F) and is the basis for Alg. 1. The link between particle systems and measure PDEs is through a classical technique known as propagation of chaos (Sznitman, 1991) and it has been used previously for similar coupled systems in the machine learning literature (Domingo-Enrich et al., 2020).

Proposition 2. *Let $\{\theta_0^{(j)}\}_{j=1}^m$ be initial features sampled uniformly over Θ, let $\{\sigma_j\}_{j=1}^m$ be uniform samples over $\{\pm 1\}$ and let $\{w_0^{(j)} = 1\}_{j=1}^m$ be the initial weight values, which are set to 1. Let $\{X_0^{(i)}\}_{i=1}^N$ be the initial "generated" samples, which are chosen i.i.d. uniformly from the target sample set $\{x_i\}_{i=1}^n$. Consider the system of ODEs/SDEs:*

$$\begin{aligned} \frac{d\theta_t^{(j)}}{dt} &= \alpha \sigma_j \nabla \tilde{F}_t(\theta_t^{(j)}), \qquad \frac{dw_t^{(j)}}{dt} = \alpha w_t^{(j)} (\sigma_j \tilde{F}_t(\theta_t^{(j)}) - \tilde{K}_t) \\ dX_t^{(i)} &= \left(-\nabla \tilde{f}_t(X_t^{(i)}) + \beta^{-1} \nabla \log \frac{d\tau_{\mathcal{X}}}{d\lambda}(X_t^{(i)})\right) dt + \sqrt{2\beta^{-1}}\, dW_t^{(i)} \end{aligned} \tag{8}$$

where

$$\begin{aligned} \tilde{F}_t(\theta) = \frac{1}{N}\sum_{i=1}^N \varphi(X_t^{(i)}, \theta) - \frac{1}{n}\sum_{i=1}^n \varphi(x_i, \theta), \qquad \tilde{f}_t(x) = \frac{1}{m}\sum_{j=1}^m \sigma_j w_t^{(j)} \varphi(x, \theta_t^{(j)}), \\ \tilde{K}_t = \mathbb{1}_{\sum_{j=1}^m w_t^{(j)} \geq m} \frac{1}{m}\sum_{j=1}^m \sigma_j w_t^{(j)} \tilde{F}_t(\theta_t^{(j)}). \end{aligned} \tag{9}$$

are the empirical counterparts of the functions in (7)*. Then the system* (8) *approximates the measure dynamics. Namely, as $m, N \to \infty$:*

- *the empirical measure $\hat{\gamma}_t = \frac{1}{m}\sum_{j=1}^m \sigma_j w_t^{(j)} \delta_{\theta_t^{(j)}}$ converges weakly to the solution $\gamma_t = \gamma_t^+ - \gamma_t^-$ of* (6) *with uniform initialization for any finite time interval $[0, T]$, and*
- *the empirical measure $\hat{\nu}_t = \frac{1}{N}\sum_{i=1}^N \delta_{X_t^{(i)}}$ converges weakly to the solution ν_t of* (6) *for any finite time interval $[0, T]$.*

Importantly, the system of ODEs/SDEs in (8) may be solved via forward Euler steps on $\{\theta_j\}_{j=1}^m$ and $\{w_j\}_{j=1}^m$ (or rather, $\{\log w_j\}_{j=1}^m$), and Euler-Maruyama updates on $\{X_0^{(i)}\}_{i=1}^N$. Such a discretization yields Algorithm 1. Algorithm 1 makes use of a tunable parameter p_R, which stands for the restart probability and will be discussed in Sec. 4 as a natural way to connect maximum likelihood with score matching. To discretize (8) we set $p_R = 0$, i.e., there are no particle restarts.

Algorithm 1 Dual $\mathcal{F}_1$-EBM training ($p_R = 0$: maximum likelihood, $p_R = (s\alpha \wedge 1)$: score matching)

Input: n samples $\{x_i\}_{i=1}^n$ of the target distribution, stepsize s, stepsize ratio α.
Initialize unif. features $(\theta_0^{(j)})_{j=1}^m$ over Θ, weights $(w_0^{(j)})_{j=1}^m$ in $[0, 1)$, signs $(\sigma_j)_{j=1}^m$ over $\{\pm 1\}$.
Initialize generated samples $\{X_0^{(i)}\}_{i=1}^N$ uniformly i.i.d. from $\{x_i\}_{i=1}^n$.
for $t = 0, \dots, T-1$ **do**
  for $i = 1, \dots, N$ **do**
    With probability p_R, replace $X_t^{(i)}$ by some uniformly chosen sample in $\{x_i\}_{i=1}^n$ (see Sec. 4).
    Sample $\zeta_t^{(i)}$ from the d_1-variate standard Gaussian.
    Perform Euler-Maruyama update: $X_{t+1}^{(i)} = X_t^{(i)} - s(\nabla \tilde{f}_t(X_t^{(i)}) + \beta^{-1} \nabla \log \frac{d\tau_{\mathcal{X}}}{d\lambda}(X_t^{(i)})) + \sqrt{2\beta^{-1}s}\, \zeta_t^{(i)}$, where $\tilde{f}_t$ is defined in (9).
  end for
  for $j = 1, \dots, m$ **do**
    Update $\theta_{t+1}^{(j)} = \theta_t^{(j)} + s\alpha\sigma_j \nabla \tilde{F}_t(\theta_t^{(j)})$, where $\tilde{F}_t$ is defined in (9).
    Update $\tilde{w}_{t+1}^{(j)} = w_{t+1}^{(j)} \exp(s\alpha\sigma_j \tilde{F}_t(\theta_t^{(j)}))$.
    Normalize if needed $w_{t+1}^{(j)} = \tilde{w}_{t+1}^{(j)} / \max\left(m^{-1} \sum_{j'=1}^m \tilde{w}_{t+1}^{(j')}, 1\right)$.
  end for
end for
Output: samples $\{X_T^{(i)}\}_{i=1}^N$, energy $f_T(x) := \frac{\beta}{m} \sum_{j=1}^m \sigma_j w_j \varphi(x, \theta_j)$.

4 Links between maximum likelihood and score matching $\mathcal{F}_1$-EBMs

In this section we uncover how the score matching loss fits seamlessly as a variant of Alg. 1, in the form of particle restarts. Interestingly, we can modify the PDE (6) in a way that allows us to make a connection with score matching. To this end, let us introduce the following coupled measure PDE:

$$\begin{aligned} \partial_t \gamma_t^\sigma &= -\alpha\sigma\nabla_\theta \cdot \left(\gamma_t^\sigma \nabla_\theta F_t(\theta)\right) + \alpha\gamma_t^\sigma \left(\sigma F_t(\theta) - K_t\right), \qquad \sigma = \pm 1,\ \ \gamma_t^\sigma = \gamma_t^\pm\,, \\ \partial_t \nu_t &= \nabla_x \cdot \left(\nu_t \left(\nabla_x f_t(x) - \beta^{-1} \nabla \log \frac{d\tau_{\mathcal{X}}}{d\lambda}\right)\right) + \beta^{-1} \Delta_x \nu_t - \alpha\left(\nu_t - \nu_n\right). \end{aligned} \tag{10}$$

Remark that the only difference between this equation and the PDE (6) for dual maximum likelihood training is the term $-\alpha(\nu_t - \nu_n)$, which draws ν_t closer to the empirical target measure ν_n. We have:

Proposition 3. *In the limit $\alpha \to \infty$, the equations for γ_t^σ in* (10) *reduce to*

$$\partial_t \gamma_t^\sigma = \sigma\nabla_\theta \cdot \left(\gamma_t^\sigma \nabla_\theta V(\gamma_t)(\theta)\right) - \gamma_t^\sigma \left(\sigma V(\gamma_t)(\theta) - \bar{V}(\gamma_t)\right), \qquad \sigma = \pm 1,\ \ \gamma_t^\sigma = \gamma_t^\pm \tag{11}$$

where $\gamma_t = \gamma_t^+ - \gamma_t^-$, $\bar{V}(\gamma) = \int_\Theta V(\gamma) d\gamma$, and $V(\gamma)(\theta)$ is the Frechet derivative of the score matching loss $L : \mathcal{M}(\Theta) \to \mathbb{R}$ defined in (4)*.*

That is, in the large α limit, equation (10) is equivalent to the Wasserstein-Fisher-Rao gradient flow of a loss L which, remarkably, is the score matching loss for $\mathcal{F}_1$-EBMs. This means that adding the term $-\alpha(\nu_t - \nu_n)$ to the dual maximum likelihood measure dynamics and letting $\alpha \to \infty$ we recover the score matching dynamics. This additional term can be easily implemented at particle level by replacing each training sample $X_t^{(i)}$ by some random target sample in $\{x_i\}_{i=1}^n$ with probability $p_R = 1 - e^{-\alpha t} = \alpha t + o(t)$ for every time interval of length t (proof in App. G).

Similar birth-death processes were used in (Rotskoff et al., 2019) in the context of neural network regression. Hence, the score matching scheme corresponds to setting the restart probability $p_R = s\alpha$ in Algorithm 1. The restart probability acts as a knob that allows us to interpolate between score matching and maximum likelihood.

In summary, score matching differs from dual maximum likelihood in that the trained measure is being "pulled" towards the target measure at all times via particle restarting. Such constant pulling should be useful to alleviate sampling problems due to metastability issues which may arise with dual maximum likelihood. However, dual maximum likelihood has the upside of providing samples of the learned EBM as a byproduct of training, which score matching does not. A good balance between both algorithms may be to use a restart probability p_R between $s\alpha \wedge 1$ and 0, or even a p_R that decreases with time from one value to the other, in such a way that at the beginning of training we avoid metastability issues by restarting the particles frequently, and at an advanced phase we perform little to no restarting to obtain faithful samples. It is also interesting to contrast our approach to score matching with the works Sutherland et al. (2018); Arbel & Gretton (2018), which using different techniques propose algorithms to train EBMs with RKHS energies via score matching. Finally, notice that a particle discretization of the flow (11) yields an alternative straightforward algorithm to train $\mathcal{F}_1$-EBMs via score matching; see Subsec. G.1. In Subsec. G.1 we show that this algorithm can be linked directly to Alg. 1 with particle restarts, without recurring to measure arguments.

5 Experiments

Setup. To illustrate Alg. 1 we perform numerical experiments on simple synthetic datasets generated by teacher models with energy $f^*(x) = \frac{1}{J}\sum_{j=1}^{J} w_j^* \sigma(\langle \theta_j^*, x\rangle)$, with $\theta_j^* \in \mathbb{S}^d$ for all j. The training is performed using Alg. 1 with the added detail that both the features $\theta_t^{(j)}$ and the particles $X_t^{(i)}$ are constrained to remain on the sphere by adding a projection step in the update of their positions. The code, figures, and videos on the dynamics can be found in the supplementary material. In the main text we consider two planted teacher neurons ($J = 2$) with negative output weights $w_1^* = w_2^* = -10$ in dimension $d = 14$ and $m = 64$ neurons for the student model, but we include additional experiments and videos in App. I and supplementary material. We study setups with two different choices of angles between the teacher neurons, which showcase different behaviors:

- Teacher neurons θ_1^*, θ_2^* forming an angle of 2.87 rad ($\approx$ 164 degrees), and output weights $w_1^* = w_2^* = -10$. The teacher neurons are almost in opposite directions, and the resulting target distribution is bimodal, as the energy has two local minimizers around θ_1^* and θ_2^* (see Figure 5).
- Teacher neurons θ_1^*, θ_2^* forming an angle of 1.37 rad ($\approx$ 78 degrees). The teacher neurons are almost orthogonal and the resulting target distribution is monomodal; indeed, when the angle is less than $\pi/2$, the target energy has a unique minimizer at the geodesic average between θ_1^* and θ_2^* (see Figure 5 in App. I).

Monitoring convergence. In all our experiments, to monitor convergence we use a testing set of n_* data points sampled from the teacher distribution: denoting these samples by $\{x_i^*\}_{i=1}^{n_*}$, we estimate the KL divergence from the student to the teacher via $\log(\frac{1}{n^*}\sum_{i=1}^{n^*} \exp(-\beta f_t(x_i^*) + \beta f^*(x_i^*))) + \frac{1}{n^*}\sum_{i=1}^{n^*}(f_t(x_i^*) + \beta f^*(x_i^*))$ where $f_t(x) = \frac{1}{m}\sum_{j=1}^{m} w_t^{(j)} \sigma(\langle \theta_t^{(j)}, x\rangle)$. Similarly, for the score matching objective we use the estimate $\frac{1}{n^*}\sum_{i=1}^{n^*} \left|\nabla_x f_t(x_i^*) - \nabla_x f^*(x_i^*)\right|^2$.

Comparison of the primal algorithm and the dual algorithms. We defer the empirical study of tuning the restart probability to App. I, and in this section focus on comparing the dual algorithm for maximum likelihood $\mathcal{F}_1$-EBMs (i.e. with $p_R = 0$) to the classical (primal) algorithm, which was the algorithm used in the experiments of Domingo-Enrich et al. (2021). The primal algorithm corresponds to Alg. 1 with $\alpha \ll 1$, while the dual algorithm uses $\alpha \gg 1$. To obtain a principled comparison of the two settings where numerical errors do not blow up, we set s to be the stepsize for the fastest process (particle evolution for the primal, neuron evolution for the dual), and $\min(\alpha, 1/\alpha)s$ the stepsize for the slow process.

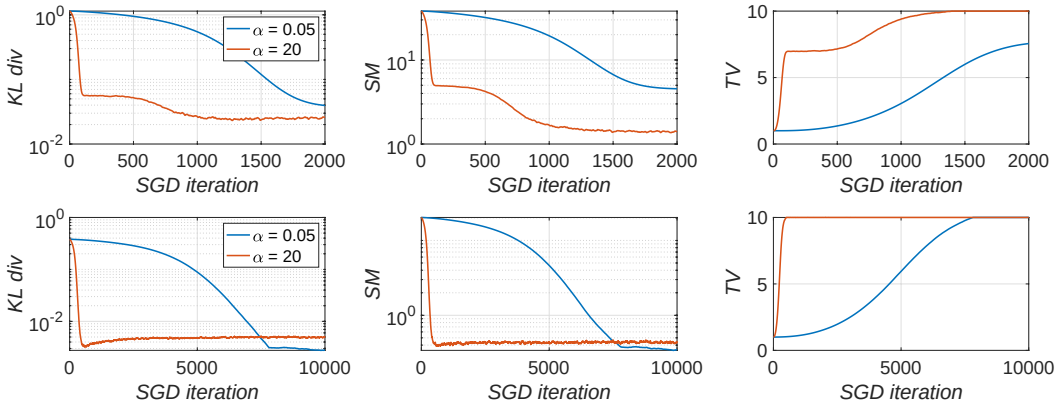

Figure 1: (Top) The evolution of the KL divergence, the score matching metric and the TV norm of the trained measure (i.e., the $\mathcal{F}_1$ norm) during training for Algorithm 1 with $\mathcal{X} = \mathbb{S}^{14}$, $m = 64$, $p_R = 0$, $s = 0.02$, $n = 10^5$, $N = 2 \cdot 10^5$ and $\alpha = 0.05$ (primal training) or $\alpha = 20$ (dual training), showing a speedup by a factor about 10-20 of the latter over the former. The angle between the two teacher neurons is 1.37 rad (monomodal distribution). (Bottom) Same experiments with an angle of 2.87 rad between the two teacher neurons (bimodal distribution).

The results are shown in Figure 6 for the two angle configurations between teacher neurons. We observe that the dual algorithm is several orders of magnitude faster at reaching KL and SM values close to the final ones, which showcases the main advantage of the dual approach. For the monomodal distribution, the final values obtained by the dual algorithm are slightly better than for the primal algorithm; for the bimodal distribution, the converse happens and the convergence is slower for both algorithms, most likely due to metastability. Interestingly, the decrease of the performance metrics seems to stall as soon as the hard $\mathcal{F}_1$-norm threshold is reached.

6 Discussion and outlook

In this work we leverage a Fenchel duality result to recast the maximum likelihood loss for $\mathcal{F}_1$-EBMs into a min-max problem on probability measures over the sample space. We provide mean-field dynamics at the measure level to solve this problem, which lead to a dual algorithm (Alg. 1) after discretization. We observe that if we restart particles at random target samples throughout training, we get an algorithm which is equivalent to training under the score matching loss. We perform experiments in which we learn planted distributions with two-layer ReLU networks, and we observe empirically that our dual algorithm is much faster than the classical one.

At theoretical level, one direction for future work is to obtain convergence results for the dynamics (6) and (10). Domingo-Enrich et al. (2020) study similar coupled Wasserstein-Fisher-Rao gradient flows, but their results only work in the case of weight learning rates much larger than position learning rates. We hypothesize that the additional term $-\alpha(\nu_t - \nu_n)$, which keeps ν_t close to ν_n, might help in the analysis.

At the numerical level, it would be interesting to further test the variant of Algorithm 1 with annealed p_R decreasing in time, to understand under which parameter setup it captures the best features of maximum likelihood and score matching. One could also test Algorithm 1 using deeper neural architectures: while the analysis is more complicated in this case, the scheme itself can be straightforwardly generalized to deep networks.

References

Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *Learning Theory*, pp. 139–153. Springer Berlin Heidelberg, 2006.

Michael Arbel and Arthur Gretton. Kernel conditional exponential family. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1337–1346. PMLR, 2018.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

Andrew Barron. Universal approximation bounds for superpositions of a sigmoidal function. *Information Theory, IEEE Transactions on*, 39:930 – 945, 1993.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Space in Probability and Statistics*. Springer, 2004.

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

Zhengdao Chen, Grant M. Rotskoff, Joan Bruna, and Eric Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks, 2020.

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

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018.

Bo Dai, Hanjun Dai, Arthur Gretton, Le Song, Dale Schuurmans, and Niao He. Kernel exponential family estimation via doubly dual embedding. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2321–2330. PMLR, 2019a.

Bo Dai, Zhen Liu, Hanjun Dai, Niao He, Arthur Gretton, Le Song, and Dale Schuurmans. Exponential family estimation via adversarial dynamics embedding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

Carles Domingo-Enrich, Samy Jelassi, Arthur Mensch, Grant Rotskoff, and Joan Bruna. A mean-field analysis of two-player zero-sum games. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20215–20226. Curran Associates, Inc., 2020.

Carles Domingo-Enrich, Alberto Bietti, Eric Vanden-Eijnden, and Joan Bruna. On energy-based models with overparametrized shallow neural networks, 2021.

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

Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *J. Mach. Learn. Res.*, 8:1217–1260, 2007.

Nelson Dunford and Jacob T Schwartz. *Linear operators. Part I, General theory*. Pure and applied mathematics (Interscience series). Interscience, 1958.

Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *UAI*, 2015.

Weinan E and Stephan Wojtowytsch. On the banach spaces associated with multi-layer relu networks: Function representation, approximation theory and gradient descent dynamics, 2020.

Weinan E, Chao Ma, and Lei Wu. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17:1407–1425, 01 2019.

Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. Philadelphia, Pa: Society for Industrial and Applied Mathematics, 1999.

Josiah Willard Gibbs. *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics*. Cambridge Library Collection - Mathematics. Cambridge University Press, 2010.

Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pp. 513–520, 2007.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

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

E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Rémi Tachet des Combes, and Ioannis Mitliagkas. Adversarial score matching and improved sampling for image generation. *arXiv preprint arXiv:2009.05475*, 2020.

Zahra Kadkhodaie and Eero P Simoncelli. Solving linear inverse problems using the prior implicit in a denoiser. *arXiv preprint arXiv:2007.13640*, 2020.

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

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. 2006.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabas Poczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yujia Li, Kevin Swersky, and Richard Zemel. Generative moment matching networks. In *ICML*, 2015.

Henry McKean. A class of markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences of the United States of America*, 56:1907–11, 01 1967.

Song Mei, Andrea Montanari, and Phan-Minh 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.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

J v Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

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

Marc Ranzato, Christopher Poultney, Sumit Chopra, et al. Efficient learning of sparse representations with an energy-based model. 2007.

Grant M Rotskoff and Eric 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.

Grant M Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Global convergence of neuron birth-death dynamics. In *Proceedings of the 36th International Conference on International Conference on Machine Learning*, Long Beach, CA, USA, 2019.

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

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

Maurice Sion. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958.

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

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

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

Yang Song and Diederik P. Kingma. How to train your energy-based models, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR 2021)*, 2021.

Danica J. Sutherland, Heiko Strathmann, Michael Arbel, and Arthur Gretton. Efficient and principled score estimation with nyström kernel exponential families. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 652–660. PMLR, 2018.

Alain-Sol Sznitman. Topics in propagation of chaos. In Paul-Louis Hennequin (ed.), *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*, pp. 165–251, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg.

Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08. Association for Computing Machinery, 2008.

Tijmen Tieleman and Geoffrey Hinton. Using fast weights to improve persistent contrastive divergence. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09. Association for Computing Machinery, 2009.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7), 2011.

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

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*. PMLR, 2016.

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

CONTENTS

A Preliminaries on Fenchel duality and maxent models **14**

B General duality results **15**

B.1 KL-regularized L^2 regression 16

B.2 KL-regularized L^∞ regression 17

C Dual $\mathcal{F}_2$-EBM training as KL-regularized MMD optimization **17**

C.1 How to train $\mathcal{F}_2$-EBMs implicitly 18

C.2 Comparison with Arbel et al. (2019) 19

C.3 How to recover an explicit form of the energy 20

D Training overparametrized two-layer neural networks via sampling **20**

E Proofs of App. B **21**

F Proofs of Sec. 3 and additional results **37**

F.1 Link of dual $\mathcal{F}_1$-EBMs training with learned MMD training 39

G Links with Score Matching **40**

G.1 Direct optimization of the score matching loss 42

G.2 Comparison with Score-based Generative Models (SGMs) 44

H Proofs of App. C **44**

I Additional experiments **46**

A PRELIMINARIES ON FENCHEL DUALITY AND MAXENT MODELS

The basic theoretical tool of the present paper is Fenchel duality, whose main applications in machine learning are maximum entropy or *maxent* models. We provide a brief description of such models to put in context the results in App. B, which are related.

Fenchel duality. If X is a Banach space and X^* is its dual space, the convex or Fenchel conjugate of $f : X \to \mathbb{R}$ is the function $f^* : X^* \to \mathbb{R}$ defined as $f^*(x^*) = \sup_{x \in X} \langle x^*, x \rangle - f(x)$. The Fenchel strong duality theorem (see Theorem 4) states that under certain conditions, if X, Y are Banach spaces, $f : X \to \mathbb{R} \cup \{+\infty\},\ g : Y \to \mathbb{R} \cup \{+\infty\}$ are convex functions, and $A : X \to Y$ is a bounded linear map, then

$$\inf_{x \in X} \{f(x) + g(Ax)\} = \sup_{y^* \in Y^*} \{-f^*(A^* y^*) - g^*(-y^*)\}. \tag{12}$$

Entropy and log-partition as convex conjugates. For simplicity, let $\mathcal{Y}$ be a finite set and let $\tau_{\mathcal{Y}} \in \mathcal{P}(\mathcal{Y}) \subseteq \mathbb{R}^{|\mathcal{Y}|}$ be a base distribution on $\mathcal{Y}$. Crucially, the KL divergence or relative entropy $D_{\mathrm{KL}}(\nu\|\tau_{\mathcal{Y}})$ is a convex function of ν and its convex conjugate is the log-partition function $v \in \mathbb{R}^{|\mathcal{Y}|} \mapsto \log(\sum_{y\in\mathcal{Y}} \tau_{\mathcal{Y}}(y)\exp(v(y)))$. The functional equivalent of this convex conjugate pair is key both for maximum entropy models, introduced below, as well as for the results of App. B.

Maximum entropy (maxent) models. Let $\Phi : \mathcal{Y} \to \mathbb{R}^q$ be a feature mapping, and $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i} \in \mathcal{P}(\mathcal{Y})$ an empirical measure. One is interested in a statistical model that is 'maximally non committal', i.e. as close as possible to the base measure in KL divergence, given that its feature moments are not too far from those of ν_n. This rationale leads to the l^∞ maxent problem (Ch. 12, Mohri et al. (2012)), which is

$$\min_{\nu\in\mathcal{P}(\mathcal{Y})} D_{\mathrm{KL}}(\nu\|\tau_{\mathcal{Y}}) \quad \text{such that} \quad \|\mathbb{E}_\nu[\Phi(y)] - \mathbb{E}_{\nu_n}[\Phi(y)]\|_\infty \leq \lambda. \tag{13}$$

Let $\nu_w \in \mathcal{P}(\mathcal{Y})$ be the distribution with density $\nu_w/\tau_{\mathcal{Y}} \propto \exp(-\langle w, \Phi(y)\rangle)$. One can apply Fenchel strong duality (equation (12)) on the problem (13), by taking the KL divergence as the function f and the indicator function of the constraint set as $g \circ A$. Using that the log-partition is the convex conjugate, the dual of (13) is

$$\max_{w\in\mathbb{R}^q} -\frac{1}{n}\sum_{i=1}^n \langle w, \Phi(y_i)\rangle - \log\left(\sum_{y\in\mathcal{Y}} \exp(-\langle w, \Phi(y)\rangle)\right) - \lambda\|w\|_1 = \frac{1}{n}\sum_{i=1}^n \log\left(\frac{d\nu_w}{d\tau_{\mathcal{Y}}}(y_i)\right) - \lambda\|w\|_1, \tag{14}$$

Strong duality holds and $\nu_{w^\star}$ is a solution of (13) when $w^\star$ is a solution of (14). That is, solving an entropy maximization problem with an ℓ^∞ constraint on some generalized moments is equivalent to solving a maximum likelihood for the exponential family problem under ℓ^1 regularization. If in (14) we replace the norm in the constraint by the ℓ^2 norm, the corresponding dual problem involves ℓ^2 norm of w instead. The ℓ^∞-ℓ^1 maxent problems (13)-(14) are to be compared with problems (18)-(19) in the next section, while the ℓ^2 maxent problems should be contrasted with problems (15)-(16).

B General duality results

In this section we state Fenchel duality results between KL regularized regression problems over probability measures with problems that are formally equivalent maximum likelihood estimation. On the one hand, in Theorem 2 the metric used for regression is the L^2 distance (not squared, unlike in least squares regression) and the corresponding dual problem is over a properly defined L^2 space. Theorem 2 is the basis for the dual formulation of $\mathcal{F}_2$-EBMs (App. C) and also for a formulation of neural network regression via sampling (App. D). These two topics are deferred to the appendices. On the other hand, in Theorem 3 the regression metric is the L^∞ distance, and the corresponding dual problem is over a space of Radon measures. Theorem 3 is the theoretical foundation for the dual formulation of $\mathcal{F}_1$-EBMs in Sec. 3. The proofs are in App. E.

Let $\mathcal{Y} \subseteq \mathbb{R}^{d_1}$ and let $\tau_{\mathcal{Y}} \in \mathcal{P}(\mathcal{Y})$ be a fixed base probability measure over $\mathcal{Y}$ with full support. Let $\mathcal{Z} \subseteq \mathbb{R}^{d_2}$ and let $\tau_{\mathcal{Z}} \in \mathcal{P}(\mathcal{Z})$ be a fixed base probability measure over $\mathcal{Z}$ with full support. Denote $L^2(\mathcal{Z}) = \{f : \mathcal{Z} \to \mathbb{R} \mid \int_{\mathcal{Z}} f(z)^2\, d\tau_{\mathcal{Z}}(z) < +\infty\}$. Let $g \in L^2(\mathcal{Z})$ be a fixed function.

Assumption 2. *Let $\varphi : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ be a continuous function such that either $\mathcal{Y}$ is compact or (i) for any fixed $z \in \mathcal{Z}$, $\varphi(y, z) = O(\xi_1(y))$ for some strictly positive $\xi_1 : \mathcal{Y} \to \mathbb{R}$, and (ii) $\xi_1(y) + \log(\xi_1(y)) = o\left(-\log\left(\frac{d\tau_{\mathcal{Y}}}{d\lambda}(y)\right) - (d_1+\epsilon)\log\|y\|_2\right)$ as $\|y\|_2 \to +\infty$ for some $\epsilon > 0$, and (iii) the function $\xi_2(y) := (\int_{\mathcal{Z}} \varphi(y, z)^2\, d\tau_{\mathcal{Z}}(z))^{1/2}$ fulfills $\sup_{y\in\mathcal{Y}} |\xi_2(y)|/\xi_1(y) < \infty$.*

Assumption 2 imposes that either $\mathcal{Y}$ is compact, or the map φ has a well-behaved growth in a certain sense, not very stringently. Note that (ii) is merely to ensure that $\xi_1(y)$ has finite expectation under the base measure $\tau_{\mathcal{Y}}$.

B.1 KL-REGULARIZED L^2 REGRESSION

Consider the two problems

$$\min_{\nu \in \mathcal{P}(\mathcal{Y})} \beta^{-1} D_{\mathrm{KL}}(\nu || \tau_{\mathcal{Y}}) + \left(\int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \varphi(y,z)\, d\nu(y) - g(z) \right)^2 d\tau_{\mathcal{Z}}(z) \right)^{1/2}, \tag{15}$$

and

$$\max_{\substack{h \in L^2(\mathcal{Z}) \\ \|h\|_{L^2} \leq 1}} - \int_{\mathcal{Z}} g(z) h(z)\, d\tau_{\mathcal{Z}}(z) - \frac{1}{\beta} \log \left(\int_{\mathcal{Y}} \exp \left(-\beta \int_{\mathcal{Z}} \varphi(y,z) h(z)\, d\tau_{\mathcal{Z}}(z) \right) d\tau_{\mathcal{Y}}(y) \right) \tag{16}$$

Theorem 2. *The problems* (15) *and* (16) *are convex. Suppose that Assumption 2 holds. Then problem* (16) *is the Fenchel dual of problem* (15)*, and strong duality holds. Moreover, the solution* $\nu^\star$ *of* (15) *is unique and its density satisfies*

$$\frac{d\nu^\star}{d\tau_{\mathcal{Y}}}(y) = \frac{1}{Z_\beta} \exp \left(-\beta \int_{\mathcal{Z}} \varphi(y,z)\, h^\star(z) d\tau_{\mathcal{Z}}(z) \right),$$

where $h^\star$ *is a solution of* (16) *and* Z_β *is a normalization constant.*

The proof of this result is in App. E. Note that in the problem (16) we are implicitly optimizing over an RKHS ball, which makes Theorem 2 close to the results from Dai et al. (2019a).

A relevant problem that is very similar to (15) is:

$$\min_{\nu \in \mathcal{P}(\mathcal{Y})} \tilde{\beta}^{-1} D_{KL}(\nu || \tau_{\mathcal{Y}}) + \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \varphi(y,z)\, d\nu(y) - g(z) \right)^2 d\tau_{\mathcal{Z}}(z). \tag{17}$$

The following result links this problem with problem (15).

Proposition 4. *Problems* (15) *and* (17) *are equivalent in the following sense: if* $\nu_1^\star$ *is a solution of* (15) *for* β*, then it is also a solution of* (17) *for*

$$\tilde{\beta} = \beta \left(4 \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \varphi(y,z)\, d\nu_1^\star(y) - g(z) \right)^2 d\tau_{\mathcal{Z}}(z) \right)^{-1/2}$$

provided that the left-most factor is non-zero. Conversely, if $\nu_2^\star$ *is a solution of* (17) *for* $\tilde{\beta}$*, then it is also a solution of* (15) *for*

$$\beta = \tilde{\beta} \left(4 \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \varphi(y,z)\, d\nu_2^\star(y) - g(z) \right)^2 d\tau_{\mathcal{Z}}(z) \right)^{1/2}.$$

And the next lemma provides additional insights into how the problems (15) and (17) differ in the planted case.

Proposition 5. *Suppose* $g : \mathcal{Z} \to \mathbb{R}$ *is of the form* $g(z) = \int_{\mathcal{Y}} \varphi(y,z)\, d\nu_p(y)$ *for some* $\nu_p \in \mathcal{P}(\mathcal{Z})$*, and assume that the (negated) log-density* $E(y) = -\log(\frac{d\nu_p}{d\tau_{\mathcal{Y}}}(y))$ *belongs to the RKHS ball* $B_{\mathcal{F}_2}(\beta_0)$*.*

(a) On the one hand, when $\beta \geq \beta_0$ *the solution* $\nu_1^\star$ *of* (15) *is equal to* ν_p*. That is, there is recovery of the planted target measure and consequently* $\int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \varphi(y,z)^2\, d\nu_1^\star(y) - g(z) \right)^2 d\tau_{\mathcal{Z}}(z) = 0$*.*

(b) On the other hand, for all choices of $\tilde{\beta}$ *finite if* $\nu_2^\star$ *is the solution of* (17)*, the unregularized regression loss at* $\nu_2^\star$ *is not zero:* $\int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \varphi(y,z)^2\, d\nu_2^\star(y) - g(z) \right)^2 d\tau_{\mathcal{Z}}(z) > 0$*. Hence,* $\nu_2^\star \neq \nu_p$ *and there is no recovery.*

B.2 KL-REGULARIZED L^∞ REGRESSION

Consider the two problems

$$\min_{\nu\in\mathcal{P}(\mathcal{Y})} \beta^{-1} D_{\mathrm{KL}}(\nu||\tau_{\mathcal{Y}}) + \left\| \int_{\mathcal{Y}} \varphi(y,\cdot)\, d\nu(y) - g(\cdot) \right\|_{L^\infty}, \tag{18}$$

and

$$\max_{\substack{\gamma\in\mathcal{M}(\mathcal{Z}) \\ \|\gamma\|_{\mathrm{TV}}\leq 1}} -\int_{\mathcal{Z}} g(z)\, d\gamma(z) - \frac{1}{\beta}\log\left(\int_{\mathcal{Y}} \exp\left(-\beta\int_{\mathcal{Z}} \varphi(y,z)\, d\gamma(z)\right) d\tau_{\mathcal{Y}}(y)\right). \tag{19}$$

Theorem 3. *The problems* (18) *and* (19) *are convex. Suppose that Assumption 2 holds and also that (i) there exists $K > 0$ such that $\sup_{z\in\mathcal{Z}} \sup_{y\in\mathcal{Y}} \varphi(y,z)/\xi_1(y) < K$, and (ii) $g(\cdot) \in C_b(\mathcal{Z})$. Then problem* (19) *is the Fenchel dual of problem* (18)*, and strong duality holds. Moreover, the solution $\nu^\star$ of* (18) *is unique and its density satisfies*

$$\frac{d\nu^\star}{d\tau_{\mathcal{Y}}}(y) = \frac{1}{Z_\beta}\exp\left(-\beta\int_{\mathcal{Z}} \varphi(y,z)\, d\gamma^\star(z)\right),$$

where $\gamma^\star$ is a solution of (19) *and Z_β is a normalization constant.*

At this point, we remark the similarity between the maxent problems (13)-(14) and problems (18)-(19). The former are stated in finite dimension and involve a constraint in the minimization problem and a penalization term in the maximization problem; the latter hold in infinite-dimensional settings and involve a a penalization term in the minimization problem and a constraint in the maximization problem.

In Theorem 2 and Theorem 3, the improvement over Domingo-Enrich et al. (2021) is that $\mathcal{Y}$, $\mathcal{Z}$ may be taken unbounded, which makes the results more general and closer to practice. This adds certain technical difficulties in constructing the Banach spaces needed to apply Fenchel duality (proofs in App. E).

C DUAL $\mathcal{F}_2$-EBM TRAINING AS KL-REGULARIZED MMD OPTIMIZATION

Let $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ and let $\tau_{\mathcal{X}} \in \mathcal{P}(\mathcal{X})$ be a fixed base measure. Assume that we have access to i.i.d. samples $\{x_i\}_{i=1}^n$ from an arbitrary target $\nu_p \in \mathcal{P}(\mathcal{X})$, and let $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be the empirical distribution. For any $f : \mathcal{X} \to \mathbb{R}$, denote by ν_f the Gibbs measure of energy f and base measure $\tau_{\mathcal{X}}$, i.e. $\frac{d\nu_f}{d\tau_{\mathcal{X}}}(x) = \exp(-f(x))/Z$. Let $\mathcal{F}_2$ be a random feature RKHS over $\mathcal{X}$ as defined in the second paragraph of Subsec. 2.1, let k be the corresponding kernel defined in (1), and analogously denote $L^2(\Theta) = \{f : \Theta \to \mathbb{R} \mid \int_\Theta f(\theta)^2\, d\tau_\Theta(\theta) < +\infty\}$. We consider the problem of training an energy-based model with energies in the RKHS ball $\mathcal{B}_{\mathcal{F}_2}(\beta)$ of radius β via maximum likelihood, i.e.

$$\begin{aligned} f^\star &= \operatorname*{arg\,min}_{f\in\mathcal{B}_{\mathcal{F}_2}(\beta)} H(\nu_n,\nu_f) = \operatorname*{arg\,min}_{f\in\mathcal{B}_{\mathcal{F}_2}(\beta)} -\frac{1}{n}\sum_{i=1}^n \log\left(\frac{d\nu_f}{d\tau_{\mathcal{X}}}(x_i)\right) \\ &= \operatorname*{arg\,min}_{f\in\mathcal{B}_{\mathcal{F}_2}(\beta)} \frac{1}{n}\sum_{i=1}^n f(x_i) + \log\left(\int_{\mathcal{X}} e^{-f(x)} d\tau_{\mathcal{X}}(x)\right), \end{aligned} \tag{20}$$

where $H(\nu,\nu') = -\int \log(\frac{d\nu'}{d\tau_{\mathcal{X}}})d\nu$ denotes the cross-entropy between two measures. Remark that an arbitrary element f of the RKHS $\mathcal{F}_2$ admits a representation as (Bach, 2017)

$$f(x) = \int_\Theta \varphi(x,\theta)h(\theta)\, d\tau_\Theta(\theta), \quad \text{where } h \in L^2(\Theta). \tag{21}$$

Thus, the problem (20) can be restated as

$$\underset{\substack{h\in L^2(\Theta)\\ \|h\|_{L^2}\leq 1}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\int_\Theta \varphi(x_i,\theta)h(\theta)\ d\tau_\Theta(\theta)+\frac{1}{\beta}\log\left(\int_{\mathcal{X}}\exp\left(-\beta\int_\Theta\varphi(x,\theta)h(\theta)\ d\tau_\Theta(\theta)\right)d\tau_{\mathcal{X}}(x)\right) \tag{22}$$

Problem (22) can be identified with problem (16) up to a sign flip by setting $\mathcal{Y}=\mathcal{X}$, $\mathcal{Z}=\Theta$ and $g(\theta)=\frac{1}{n}\sum_{i=1}^n\int_\Theta\varphi(x_i,\theta)=\int_\Theta\varphi(x,\theta)\ d\nu_n(x)$. Hence, if Assumption 2 holds, by Theorem 2 the problem (22) is the Fenchel dual of

$$\min_{\nu\in\mathcal{P}(\mathcal{X})}\beta^{-1}D_{KL}(\nu||\tau_{\mathcal{X}})+\text{MMD}_k(\nu,\nu_n), \tag{23}$$

where $\text{MMD}_k(\nu,\nu_n)=\left(\int_{\mathcal{X}\times\mathcal{X}}k(x,x')\ d(\nu-\nu_n)(x)\ d(\nu-\nu_n)(x')\right)^{1/2}$ is known as the maximum mean discrepancy (MMD) for the kernel k (Gretton et al., 2012). See Lemma 13 in App. H for the derivation. And the analog of problem (17) is

$$\min_{\nu\in\mathcal{P}(\mathcal{X})}\tilde{\beta}^{-1}D_{KL}(\nu||\tau_{\mathcal{X}})+\text{MMD}_k^2(\nu,\nu_n) \tag{24}$$

The following corollary of Theorem 2 and Proposition 4 describes precisely the link between the solutions of problems (23) and (24) and the solution of the maximum likelihood problem (22).

Corollary 1. *Suppose that Assumption 2 holds for $\mathcal{Y}=\mathcal{X}$ and $\mathcal{Z}=\Theta$. The solution $\nu_1^\star$ of* (23) *is unique and of the form*

$$\frac{d\nu_1^\star}{d\tau_{\mathcal{X}}}(x)=\frac{1}{Z_\beta}\exp\left(-\beta\int_\Theta\varphi(x,\theta)\ h^\star(\theta)d\tau_\Theta(\theta)\right),$$

where $h^\star$ is a solution of (22) *and Z_β is a normalization constant. Additionally, the unique solution $\nu_2^\star$ of* (24) *is equal to the solution $\nu_1^\star$ of* (23) *when $\beta=2\text{MMD}_k(\nu_2^\star,\nu_n)\tilde{\beta}$.*

Undoing the change of variables (21), we see that $f^\star(x)=\beta\int\varphi(x,\theta)\ h^\star(\theta)d\tau_\Theta(\theta)$ is the energy in $\mathcal{B}_{\mathcal{H}}(\beta)$ that maximizes the likelihood. Hence, although problems (23)-(24) are implicit in the sense that they do not involve energy functions, the solutions $\nu_1^\star$ and $\nu_2^\star$ coincide with the Gibbs measure $\nu_{f^\star}$ that is obtained through maximum likelihood EBM training.

Consequently, solving (23) or (24) provides an implicit way to train maximum likelihood $\mathcal{F}_2$-EBMs. Maximum likelihood $\mathcal{F}_2$-EBMs are classically trained via gradient descent on a parametrized form of the energy, via either a feature discretization of (22) or a representer theorem applied on (20). Their computational bottleneck is the gradient estimation procedure, which relies on sampling from the trained model at every step; a task that is exponentially costly in β for non-convex energies.

C.1 How to train $\mathcal{F}_2$-EBMs implicitly

Suppose from now on that $\mathcal{X}$ is either a domain (connected open subset) of $\mathbb{R}^{d_1}$ or a Riemannian manifold embedded in $\mathbb{R}^{d_1}$, case in which differential operators are understood in the Riemannian sense. Since the objective functionals in (23) and (24) are convex in ν (Theorem 2), a natural approach to solve these problems is to approximate their Wasserstein gradient flows. Namely, Lemma 14 in App. H shows that for (23) the Wasserstein gradient flow takes the form of a McKean-Vlasov equation (McKean, 1967):

$$\partial_t\nu_t=\nabla\cdot\left(\nu_t\left(-\beta^{-1}\nabla\log\frac{d\tau_{\mathcal{X}}}{d\lambda}(x)+\frac{\int_{\mathcal{X}}\nabla_x k(x,x')\ d(\nu_t-\nu_n)(x')}{\text{MMD}_k(\nu_t,\nu_n)}\right)\right)+\beta^{-1}\Delta\nu_t, \tag{25}$$

where λ is the Lebesgue or Hausdorff measure over $\mathcal{X}$, and for (24) it is:

$$\partial_t\nu_t=\nabla\cdot\left(\nu_t\left(-\tilde{\beta}^{-1}\nabla\log\frac{d\tau_{\mathcal{X}}}{d\lambda}(x)+2\int_{\mathcal{X}}\nabla_x k(x,x')\ d(\nu_t-\nu_n)(x')\right)\right)+\tilde{\beta}^{-1}\Delta\nu_t, \tag{26}$$

Remark the striking similarity of this equation with the ones found in Rotskoff & Vanden-Eijnden (2018); Mei et al. (2018), which study McKean-Vlasov equations for overparametrized two-layer neural network training. As is customary, we approximate McKean-Vlasov equations via coupled particle systems: in the case of (25),

$$dX_t^{(i)} = \left(\beta^{-1} \nabla \log \frac{d\tau_{\mathcal{X}}}{d\lambda}(X_t^{(i)}) - \frac{\int_{\mathcal{X}} \nabla_x k(X_t^{(i)}, x') \, d(\nu_{t,N} - \nu_n)(x')}{\mathrm{MMD}_k(\nu_{t,N}, \nu_n)} \right) dt + \sqrt{2\beta^{-1}} \, dW_t^{(i)} \tag{27}$$

for $i = 1, \dots, N$, where $\nu_{t,N} = \frac{1}{N} \sum_{i=1}^N \delta_{X_t^{(i)}}$, and in the case of (26),

$$dX_t^{(i)} = \left(\tilde{\beta}^{-1} \nabla \log \frac{d\tau_{\mathcal{X}}}{d\lambda}(X_t^{(i)}) - 2 \int_{\mathcal{X}} \nabla_x k(X_t^{(i)}, x') \, d(\nu_{t,N} - \nu_n)(x') \right) dt + \sqrt{2\tilde{\beta}^{-1}} \, dW_t^{(i)} \tag{28}$$

A classical argument known as propagation of chaos (Sznitman, 1991) shows that when the number of particles N goes to infinity, $(\nu_{t,N})_{t \in [0,T]}$ converges weakly to the solution $(\nu_t)_{t \in [0,T]}$ of (25) for any fixed $T > 0$. Although this is only a qualitative guarantee, Rotskoff & Vanden-Eijnden (2018); Chen et al. (2020) provide quantitative central limit theorems for McKean-Vlasov equations similar to (26). Loosely speaking, they find that the variance is no larger than the Monte-Carlo variance one would obtain by sampling i.i.d. from the solution measure The Euler-Maruyama discretizations of the SDEs (27) and (28) yield two alternative implementable algorithms for implicit EBM training.

Algorithm 2 Implicit $\mathcal{F}_2$-EBM training (discretization of equations (27)/(28))

Input: n samples $\{x_i\}_{i=1}^n$ of the target distribution, N initialization samples $\{X_0^{(i)}\}_{i=1}^N$, inverse temperature β (if (27)), reparametrized inverse temperature $\tilde{\beta}$ (if (27)).
for $t = 0, \dots, T-1$ **do**
  If (27)**:** Compute the $\mathrm{MMD}_k^2(\nu_{t,N}, \nu_n) = \frac{1}{n^2} \sum_{i,j=1}^N k(X_t^{(i)}, X_t^{(j)}) + \frac{1}{m^2} \sum_{i,j=1}^n k(x_i, x_j) - \frac{2}{Nn} \sum_{i=1}^N \sum_{j=1}^n k(X_t^{(i)}, x_j)$
  for $i = 1, \dots, N$ **do**
    Sample $\zeta_t^{(i)}$ from the d_1-variate standard Gaussian.
    Perform Euler-Maruyama update:
    If (27), $X_{t+1}^{(i)} = X_t^{(i)} - s \int_{\mathcal{X}} \nabla_x k(X_t^{(i)}, x') \; d(\nu_{t,N} - \nu_n)(x') / \mathrm{MMD}_k(\nu_{t,N}, \nu_n) + s\beta^{-1} \nabla \log \frac{d\tau_{\mathcal{X}}}{d\lambda}(X_t^{(i)}) + \sqrt{2\beta^{-1} s} \zeta_t^{(i)}$. **If** (28), $X_{t+1}^{(i)} = X_t^{(i)} - 2s \int_{\mathcal{X}} \nabla_x k(X_t^{(i)}, x') \, d(\nu_{t,N} - \nu_n)(x') + s\tilde{\beta}^{-1} \nabla \log \frac{d\tau_{\mathcal{X}}}{d\lambda}(X_t^{(i)}) + \sqrt{2\tilde{\beta}^{-1} s} \zeta_t^{(i)}$.
  end for
end for
Output: samples $\{X_T^{(i)}\}_{i=1}^N$, if (27), energy $E_T(x) := \beta \frac{\int_{\mathcal{X}} \nabla_x k(x, x') \; d(\nu_{T,N} - \nu_n)(x')}{\mathrm{MMD}_k(\nu_{T,N}, \nu_n)}$, if (28), energy $E_T(x) := 2\tilde{\beta} \int_{\mathcal{X}} k(x, x') \, d(\nu_{T,N} - \nu_n)(x')$.

C.2 Comparison with Arbel et al. (2019)

Crucially, Algorithm 2 when discretizing (28) is exactly the algorithm studied by Arbel et al. (2019). They start from pure MMD Wasserstein gradient flows, and they study convergence for those. They introduce noise injection/entropy regularization as a way to obtain certain convergence guarantees, and experimentally in their Figure 1 they observe a dramatic improvement in the training and test error against the pure MMD flow. Our theory justifies this behavior; their algorithm is implicitly training an $\mathcal{F}_2$-EBM and the noise level controls the RKHS radius over which the energy is optimized.

They propose using a schedule in which the noise decreases to zero (in our notation, $\beta \to +\infty$). This corresponds to optimizing over growing RKHS balls. Leveraging statistical learning results from Domingo-Enrich et al. (2021), the generalization error can be written as a statistical (Rademacher complexity) term which increases with the radius β, plus an approximation term decreasing with β. Thus, there exists an optimal non-zero noise level which should be maintained.

C.3 How to recover an explicit form of the energy

Let $\nu^\star$ be the unique stationary solution of (25), which is the unique minimizer of (23) (see Lemma 15). Also by Lemma 15, this solution must fulfill

$$\frac{d\nu^\star}{d\tau_{\mathcal{X}}} = \frac{1}{Z_\beta} \exp\left(-\beta \frac{\int_{\mathcal{X}} k(x,x')\, d(\nu^\star - \nu_n)(x')}{\mathrm{MMD}_k(\nu^\star, \nu_n)} \right) \tag{29}$$

This equality leads us to believe that when we run Algorithm 2, $E_t(x) := \beta \int_{\mathcal{X}} k(x,x')\, d(\nu_{t,N} - \nu_n)(x')/\mathrm{MMD}_k(\nu_{t,N}, \nu_n)$ can be used as an rough estimate of the energy of the trained implicit EBM at time t, although of course this intuition is only accurate when $\nu_{t,N}$ is close enough to the equilibrium measure $\nu^\star$. For consistency with (20), it is also interesting to note that the estimate E_t has constant RKHS norm $\|E_t\|_{\mathcal{H}} = \beta$, since $\| \int_{\mathcal{X}} k(x,x')\, d(\nu_{t,N} - \nu_n)(x') \|_{\mathcal{H}} = \mathrm{MMD}_k(\nu_{t,N}, \nu_n)$.

Similar equations can be derived for the dynamics (26), which lead to an energy estimate of the form $E_t(x) := 2\tilde{\beta} \int_{\mathcal{X}} k(x,x')\, d(\nu_{t,N} - \nu_n)(x')$.

D Training overparametrized two-layer neural networks via sampling

In the previous section we described how the general duality result from App. B can be leveraged to train EBMs implicitly via the Wasserstein gradient flow of a functional formally similar to the two-layer neural network regression loss. In this section we take the reverse approach: we use the results from App. B to describe how overparametrized two-layer neural networks can be trained via techniques developed for maximum likelihood EBMs.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\tau_{\mathcal{X}}$ be a fixed base probability measure over $\mathcal{X}$. Let $\mathcal{F}_1$ be the feature-learning space defined in the first paragraph of Subsec. 2.1. Overparametrized two-layer neural network regression for some target $g : \mathcal{X} \to \mathbb{R}$ corresponds to solving

$$\min_{f \in \mathcal{B}_{\mathcal{F}_1}(\beta_0)} \int_{\mathcal{X}} \big(f(x) - g(x)\big)^2 \, d\tau_{\mathcal{X}}(x) \tag{30}$$

for an arbitrary ball radius β_0. This problem has been tackled via Wasserstein gradient flows and propagation of chaos by several works (Rotskoff & Vanden-Eijnden, 2018; Chizat & Bach, 2018; Mei et al., 2018; Sirignano & Spiliopoulos, 2019). We briefly summarize their construction up to slight differences. Functions in $\mathcal{B}_{\mathcal{F}_1}(\beta_0)$ can be written as $f(x) = \int_\Theta \varphi(x,\theta)\, d\gamma(\theta)$ for some signed Radon measure γ with bounded total variation norm $\|\gamma\|_{\mathrm{TV}} := \int_\Theta d|\gamma|(\theta) \leq \beta_0$. Furthermore, if we set $\Omega = \Theta \times \mathbb{R}$ and take a surjective $\chi : \mathbb{R} \to \mathbb{R}$, we obtain the parametrization $f(x) = \int_\Omega \chi(r)\varphi(x,\theta)\, d\mu(\theta, r)$ for some $\mu \in \mathcal{P}(\Omega)$ such that $\int |\chi(r)|\, d\mu(\theta,r) \leq \beta$. With this characterization, and writing compactly $\omega := (\theta, r)$ and $\tilde{\varphi}(x,\omega) = \chi(r)\varphi(x,\theta)$, we can rewrite (30) as

$$\min_{\mu \in \mathcal{P}(\Omega)} \int_{\mathcal{X}} \left(\int_\Omega \tilde{\varphi}(x,\omega)\, d\mu(\omega) - g(x) \right)^2 d\tau_{\mathcal{X}}(x) + \delta \int_\Omega |\chi|\, d\mu + \tilde{\beta}^{-1} \int_\Omega \log\left(\frac{d\mu}{d\lambda}\right)\, d\mu, \tag{31}$$

where λ denotes the Lebesgue measure over Θ. To go from (30) to (31), we have switched from a constraint on the $\mathcal{F}_1$ norm to a penalization term $\delta \int_\Omega |\chi|\, d\mu$, and we have also added a differential entropy regularizer $\tilde{\beta}^{-1} \int_\Omega \log\left(\frac{d\mu}{d\lambda}\right)\, d\mu$, which Rotskoff & Vanden-Eijnden (2018); Mei et al. (2018) introduce to simplify their analysis.

At this point, remark that if we define the probability measure τ_Ω to have density $\frac{d\tau_\Omega}{d\lambda}(\theta, r) = \exp(-\beta\delta|\chi|)/Z$ w.r.t the Lebesgue measure, then we have

$$\begin{aligned}
&\beta^{-1} \int_\Omega \log\left(\frac{d\mu}{d\lambda}\right)\, d\mu + \delta \int_\Omega |\chi|\, d\mu = \beta^{-1} \int_\Omega \log\left(\frac{d\mu}{d\lambda}\right)\, d\mu - \beta^{-1} \int_\Omega \log(\exp(-\beta\delta|\chi|))\, d\mu \\
&= \beta^{-1} \int_\Omega \log\left(\frac{d\mu}{d\lambda} \frac{1}{\exp(-\beta\delta|\chi|)}\right)\, d\mu = \beta^{-1} \int_\Omega \log\left(\frac{d\mu}{d\tau_\Omega}\right)\, d\mu + K = \tilde{\beta}^{-1} D_{KL}(\mu||\tau_\Omega) + K
\end{aligned}$$

for some constant K arising from the normalization factor of τ_Ω. That is, up to a constant term equation (31) can be rewritten as

$$\min_{\mu\in\mathcal{P}(\Omega)} \int_{\mathcal{X}} \left(\int_\Omega \tilde{\varphi}(x,\omega)\, d\mu(\omega) - g(x) \right)^2 d\tau_{\mathcal{X}}(x) + \tilde{\beta}^{-1} D_{KL}(\mu||\tau_\Omega) \tag{32}$$

The key observation is that is equation is formally equal to (17) when we set $\mathcal{Z} = \mathcal{X}$, $\mathcal{Y} = \Omega$ and $\varphi = \tilde{\varphi}$. Most importantly, as shown by Corollary 2 we can apply Proposition 4 and the Fenchel duality result Theorem 2 to obtain links with the following problem, which is the analog of (16):

$$\max_{\substack{h\in L^2(\mathcal{X},\tau_{\mathcal{X}}) \\ \|h\|_{L^2}\le 1}} -\int_{\mathcal{X}} g(x)h(x)\, d\tau_{\mathcal{X}}(x) - \frac{1}{\beta}\log\left(\int_\Omega \exp\left(-\beta \int_{\mathcal{X}} \tilde{\varphi}(x,\omega)h(x)\, d\tau_{\mathcal{X}}(x) \right) d\tau_\Omega(\omega) \right) \tag{33}$$

Corollary 2. *Let $h^\star$ be a solution of* (33)*. Then, $\mu^\star \in \mathcal{P}(\Omega)$ with density*

$$\frac{d\mu^\star}{d\tau_\Omega}(\omega) = \frac{1}{Z_\beta} \exp\left(-\beta \int_{\mathcal{X}} \tilde{\varphi}(x,\omega)\, h^\star(\omega) d\tau_\Omega(\omega) \right),$$

is a solution of (32) *with* $\tilde{\beta} = \beta \left(4 \int_{\mathcal{X}} \left(\int_\Omega \tilde{\varphi}(x,\omega)\, d\mu^\star(\omega) - g(x) \right)^2 d\tau_{\mathcal{X}}(x) \right)^{-1/2}$.

E PROOFS OF APP. B

The proofs of Theorem 2 and Theorem 3 are based on the proofs found in Appendix E of Domingo-Enrich et al. (2021). We make use of Fenchel strong duality, which is stated in Theorem 4.

Theorem 4. *[Fenchel strong duality; Borwein & 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, $p^* \ge d^*$. Moreover if f, g, and A satisfy either

1. *f and g are lower semi-continuous and $0 \in \mathrm{core}(\mathrm{dom}\, g - A\,\mathrm{dom}\, f)$ where* core *is the algebraic interior and* dom *h, where h is some function, is the set $\{z : h(z) < +\infty\}$,*

2. *or $A\,\mathrm{dom}\, f \cap \mathrm{cont}\, g \neq \emptyset$ 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.

We also rely on a generalization of von Neumann's minimax theorem (Neumann, 1928). For our purposes, the theorem stated below by Kneser (1952) suffices, but a further generalization by Sion (1958) to quasi-convex and quasi-concave functions is more widely known in the literature. Note however that the compactness assumption on one of the sets cannot be relaxed.

Theorem 5. *Kneser (1952) Let A be a non-empty compact convex subset of a locally convex topological vector space space E and B 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:*

- *For each $y \in B$, the function $x \mapsto f(x, y)$ is upper semicontinuous and concave,*
- *For each $x \in A$, the function $y \mapsto f(x, y)$ is convex.*

Then we have

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

We also make use of the Riesz-Markov-Kakutani theorem, which we reproduce in Theorem 6.

Theorem 6. *[Riesz-Markov-Kakutani representation theorem] Let X be a locally compact Hausdorff space and let $C_0(X)$ be the space of continuous functions from X to $\mathbb{R}$ vanishing at infinity, i.e. such that $f(x) \to 0$ when $\|x\|_2 \to \infty$. For any continuous linear functional ψ on $C_0(X)$, there is a unique (countably additive) finite signed regular Borel measure μ on X such that*

$$\forall f \in C_0(X): \quad \psi(f) = \int_X f(x)\, d\mu(x).$$

The norm of ψ as a linear functional is the total variation of μ, that is $\|\psi\| = \|\mu\|_{TV} = |\mu|(X) = \mu_+(X) + \mu_-(X)$, where the decomposition $\mu = \mu_+ - \mu_-$ into positive measures is given by the Hahn decomposition theorem. Finally, ψ is positive if and only if the measure μ is non-negative.

By definition, the space of finite signed Radon measures $\mathcal{M}(X)$ is the same as the space of finite signed regular Borel measures (Radon measures are Borel measures that are finite on compact sets, which is holding directly because we restrict to finite measures). In other words, Theorem 6 states that we have an isometry between the topological dual $C_0^*(X)$ and $\mathcal{M}(X)$. The following theorem is an analogous result for the dual of the Banach space $C_b(X)$ of bounded continuous functions.

Theorem 7. *[Riesz representation theorem for $C_b^*(X)$, Dunford & Schwartz (1958)] Let X be a normal topological space. Let $rba(X)$ be the space of finitely additive finite signed regular Borel measures μ on X. It holds that*

$$C_b^*(X) = rba(X).$$

Finally, we recall the Banach-Alaoglu theorem from functional analysis, which we will use to show compactness and apply Theorem 5.

Theorem 8. *[Banach-Alaoglu theorem] For any topological vector space X with continuous dual space X^*, the closed unit ball of X^* in the dual norm (i.e. $\mathcal{B}_{X^*} = \{x^* \in X^* | \sup_{x\in X}\langle x^*, x\rangle \leq 1\rangle\}$) is compact in the weak-* topology, which the weakest topology on X^* making all maps $\langle x, \cdot\rangle : X^* \to \mathbb{R}$ continuous, as x ranges over X. In particular, for Hilbert spaces H we have that $\mathcal{B}_H$ is compact in the weak-* topology, which coincides with the weak topology in this case.*

Theorem 2. *The problems* (15) *and* (16) *are convex. Suppose that Assumption 2 holds. Then problem* (16) *is the Fenchel dual of problem* (15)*, and strong duality holds. Moreover, the solution $\nu^\star$ of* (15) *is unique and its density satisfies*

$$\frac{d\nu^\star}{d\tau_{\mathcal{Y}}}(y) = \frac{1}{Z_\beta}\exp\left(-\beta\int_{\mathcal{Z}}\varphi(y,z)\, h^\star(z) d\tau_{\mathcal{Z}}(z)\right),$$

where $h^\star$ is a solution of (16) *and Z_β is a normalization constant.*

Proof. We use Theorem 4.

On the one hand, we set $X = \mathcal{M}_{\xi_1}(\mathcal{Y})$, which we define to be the space of Radon measures over $\mathcal{Y}$ such that the weighted total variation

$$\|\nu\|_{\mathrm{TV},\xi_1} := \int_{\mathcal{Y}} \xi_1(y)\, d|\nu|(y)$$

is finite, where $\xi_1 : \mathcal{Y} \to \mathbb{R}$ is the strictly positive function given by Assumption 2(i). By Lemma 1, $\mathcal{M}_{\xi_1}(\mathcal{Y})$ is a Banach space with norm $\|\cdot\|_{\mathrm{TV},\xi_1}$ and its continuous dual contains the set $C_{b,\xi_1}(\mathcal{Y})$ of continuous functions f such that f/ξ_1 is bounded.

On the other hand, we set $Y = L^2(\mathcal{Z}) = \{f : \mathcal{Z} \to \mathbb{R} \mid \int_{\mathcal{Z}} f(z)^2\, d\tau_{\mathcal{Z}}(z) < +\infty\}$, the Hilbert space of square-integrable functions on $\mathcal{Z}$ under the base measure $\tau_{\mathcal{Z}}$, which is of course self-dual.

Define $F : \mathcal{M}_{\xi_1}(\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ as

$$F(\nu) = \begin{cases} \beta^{-1} D_{KL}(\nu||\tau_{\mathcal{Y}}) & \text{if } \nu \in \mathcal{P}(\mathcal{Y}), \\ +\infty & \text{otherwise} \end{cases}. \tag{34}$$

Lemma 2 states that F is a convex functional and that its convex conjugate $F^* : \mathcal{M}^*_{\xi_1}(\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ restricted to $C_{b,\xi_1}(\mathcal{Y})$ satisfies

$$F^*(q) = \beta^{-1} \log \left(\int_{\mathcal{Y}} \exp(\beta q(y'))\, d\tau_{\mathcal{Y}}(y') \right). \tag{35}$$

Define $G : L^2(\mathcal{Z}) \to \mathbb{R} \cup \{+\infty\}$ as

$$G(\psi) = \left(\int_{\mathcal{Z}} \big(\psi(z) - g(z)\big)^2 \, d\tau_{\mathcal{Z}}(z) \right)^{1/2}, \tag{36}$$

Lemma 3 states that G is a convex functional and that its convex conjugate $G^* : L^2(\mathcal{Z}) \to \mathbb{R} \cup \{+\infty\}$ is of the form

$$G^*(\psi) = \begin{cases} \int_K g(z)\psi(z)\, d\tau_{\mathcal{Z}}(z) & \text{if } \|\psi\|_{L^2(\mathcal{Z})} \leq 1, \\ +\infty & \text{otherwise} \end{cases}. \tag{37}$$

Define $A : \mathcal{M}_{\xi_1}(\mathcal{Y}) \to L^2(\mathcal{Z})$ as

$$(A\nu)(z) = \int_{\mathcal{Y}} \varphi(y, z)\, d\nu(y).$$

The linear operator A is well defined and continuous by Lemma 4. Lemma 4 also states that $A^* : L^2(\mathcal{Z}) \to \mathcal{M}^*_{\xi_1}(\mathcal{Y})$ is of the form

$$(A^*h)(y) = \int_{\mathcal{Z}} \varphi(y, z)h(z)\, d\tau_{\mathcal{Z}}(z) \tag{38}$$

Hence, we have that $\min_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y})} \beta^{-1} D_{KL}(\nu||\tau_{\mathcal{Y}}) + (\int_{\mathcal{Z}} (\int_{\mathcal{Y}} \varphi(y, z)\, d\nu(y) - g(z))^2\, d\tau_{\mathcal{Z}}(z))^{1/2}$ can be written as

$$p^* = \inf_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y})} \{F(\nu) + G(A\nu)\}.$$

And problem (16) can be written as

$$d^* = \sup_{\substack{h \in L^2(\mathcal{Z}), \\ \|h\|_{L^2} \leq 1}} \{-F^*(-A^*h) - G^*(h)\}. \tag{39}$$

To apply Theorem 4, it only remains to show that condition 2 holds. That is, we have to check that A dom $F \cap$ cont $G \neq \emptyset$. Consider $\psi = A\nu$ for some $\nu \in \mathcal{P}(\mathcal{Y}) \cap \mathcal{M}_{\xi_1}(\mathcal{Y})$ absolutely continuous w.r.t. $\tau_{\mathcal{Y}}$. Then $\psi \in A$ dom F. Moreover, since G is a continuous functional, we have that cont $G = L^2(\mathcal{Z})$. Thus, ψ also belongs to cont G and we conclude that A dom $F \cap$ cont $G \neq \emptyset$.

By Theorem 4, $p^* = d^*$, and since p^* is finite, we have that the supremum in (39) is attained; let $h^\star$ be one maximizer. We show that $p^* = \inf_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y})} \{F(\nu) + G(A\nu)\} = \inf_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y}) \cap \mathcal{P}(\mathcal{Y})} \{F(\nu) + G(A\nu)\}$ admits a minimizer by the direct method of the calculus of variations. First, notice that F and $G \circ A$ are lower semicontinuous in the topology of weak convergence:

- F by the lower semicontinuity of the KL divergence (Posner, 1975),
- and $G \circ A$ because can be written as a supremum of continuous functions as shown in (44), and thus its sublevel sets are closed because they are the intersection of closed sublevel sets. Closed sublevel sets is equivalent to lower semicontinuity.

Second, $\mathcal{P}(\mathcal{Y}) \cap \mathcal{M}_{\xi_1}(\mathcal{Y})$ is compact, because $\mathcal{P}(\mathcal{Y})$ is compact and $\mathcal{M}_{\xi_1}(\mathcal{Y})$ is closed as it is a Banach space. Hence, the direct method of the calculus of variations applies. Let $\nu^\star$ be one minimizer of p^*.

It remains to show that

$$\frac{d\nu^\star}{d\tau_{\mathcal{Y}}}(y) = \frac{1}{Z_\beta} \exp\left(-\beta \int_{\mathcal{Z}} \varphi(y,z)\, h^\star(z) d\tau_{\mathcal{Z}}(z)\right). \tag{40}$$

We make use of the argument to prove Fenchel weak duality, which is:

$$\begin{aligned}
&\sup_{\substack{h\in L^2(\mathcal{Z}),\\ \|h\|_{L^2}\leq 1}} \{-F^*(-A^*h) - G^*(h)\} = -F^*(-A^*h^\star) - G^*(h^\star) \\
&= -\sup_{\nu\in\mathcal{M}_{\xi_1}(\mathcal{Y})} \left\{\langle -A^*h^\star, \nu\rangle - F(\nu)\right\} - \sup_{\psi\in L^2(\mathcal{Z})} \left\{\langle h^\star, \psi\rangle - G(\psi)\right\} \\
&\leq -\sup_{\nu\in\mathcal{M}_{\xi_1}(\mathcal{Y})} \left\{\langle -A^*h^\star, \nu\rangle - F(\nu) + \langle h^\star, A\nu\rangle - G(A\nu)\right\} \\
&= -\sup_{\nu\in\mathcal{M}_{\xi_1}(\mathcal{Y})} \left\{-F(\nu) - G(A\nu)\right\} = \inf_{\nu\in\mathcal{M}_{\xi_1}(\mathcal{Y})} \{F(\nu) + G(A\nu)\} \\
&= F(\nu^\star) + G(A\nu^\star)
\end{aligned} \tag{41}$$

Thus, for strong duality to hold we must have that

$$\nu^\star = \operatorname*{arg\,min}_{\nu\in\mathcal{M}_{\xi_1}(\mathcal{Y})} \left\{\langle -A^*h^\star, \nu\rangle - F(\nu)\right\} \tag{42}$$

By Lemma 5(i), this implies that equation (40) holds, and by Lemma 5(ii) we have that $\nu^\star = \arg\min_{\nu\in\mathcal{P}(\mathcal{Y})}\{F(\nu) + G(A\nu)\}$. □

Lemma 1. *Let $\mathcal{M}_{\xi_1}(\mathcal{Y})$ be the vector space of Radon measures over $\mathcal{Y}$ such that the weighted total variation $\|\nu\|_{TV,\xi_1} := \int_{\mathcal{Y}} \xi_1(y)\, d|\nu|(y)$ is finite, where $\xi_1 : \mathcal{Y} \to \mathbb{R}$ is the strictly positive function given by Assumption 2(i). $\mathcal{M}_{\xi_1}(\mathcal{Y})$ is a Banach space with norm $\|\cdot\|_{TV,\xi_1}$.*

Let $C_{b,\xi_1}(\mathcal{Y})$ be the set of continuous functions f such that $f/\xi_1 \in C_b(\mathcal{Y})$, i.e. is a bounded continuous function. The continuous dual $\mathcal{M}^*_{\xi_1}(\mathcal{Y})$ contains the set $C_{b,\xi_1}(\mathcal{Y})$.

Proof. If we define the linear map $\tilde{\xi_1} : \mathcal{M}_{\xi_1}(\mathcal{Y}) \to \mathcal{M}(\mathcal{Y})$ as $\nu \mapsto \tilde{\nu}$, where $\tilde{\nu}$ is absolutely continuous w.r.t ν and has density $\frac{d\tilde{\nu}}{d\nu}(y) = \xi_1(y)$, we have that $\|\nu\|_{\mathrm{TV},\xi_1} = \|\tilde{\xi_1}(\nu)\|_{\mathrm{TV}}$. Notice that $\tilde{\xi_1}$ is surjective, because for all $\tilde{\nu} \in \mathcal{M}(\mathcal{Y})$, the measure ν with density $\frac{d\nu}{d\tilde{\nu}}(y) = \xi_1(y)^{-1}$ is a Radon measure (possibly not signed, because we cannot guarantee that ν_+ nor ν_- is finite) such that $\tilde{\xi_1}\nu = \tilde{\nu}$. $\tilde{\xi_1}$ is a surjective isometry between $\mathcal{M}_{\xi_1}(\mathcal{Y})$ and $\mathcal{M}(\mathcal{Y})$, which shows that $\mathcal{M}_{\xi_1}(\mathcal{Y})$ is a Banach space.

Let $\mathcal{M}^*(\mathcal{Y})$ be the dual space of $\mathcal{M}(\mathcal{Y})$. By the Riesz-Markov-Kakutani representation theorem (Theorem 6) and the fact that the double dual space contains the primal space, $\mathcal{M}^*(\mathcal{Y})$ immediately contains the set of continuous functions $C_0(\mathcal{Y})$ on $\mathcal{Y}$ vanishing at infinity. Furthermore, $\mathcal{M}^*(\mathcal{Y})$ contains the larger set of bounded continuous functions $C_b(\mathcal{Y})$, because if $f \in C_b(\mathcal{Y})$, for any $\tilde{\nu} \in \mathcal{M}(\mathcal{Y})$,

$$\langle f, \tilde{\nu}\rangle = \int_{\mathcal{Y}} f(y)\, d\tilde{\nu}(y) \leq \sup_{y\in\mathcal{Y}} f(y) \|\tilde{\nu}\|_{\mathrm{TV}}.$$

In an analogous way, $\mathcal{M}^*_{b,\xi_1}(\mathcal{Y})$ contains the set $C_{b,\xi_1}(\mathcal{Y})$, because if $f \in C_{b,\xi_1}(\mathcal{Y})$, for any $\nu \in \mathcal{M}_{b,\xi_1}(\mathcal{Y})$,

$$\langle f, \nu\rangle = \int_{\mathcal{Y}} \frac{f(y)}{\xi_1(y)} \xi_1(y)\, d\nu(y) = \int_{\mathcal{Y}} \frac{f(y)}{\xi_1(y)}\, d\tilde{\nu}(y) \leq \|\tilde{\nu}\|_{\mathrm{TV}} \sup_{y\in\mathcal{Y}} \frac{f(y)}{\xi_1(y)} = \|\nu\|_{\mathrm{TV},\xi_1} \sup_{y\in\mathcal{Y}} \frac{f(y)}{\xi_1(y)}$$

□

Lemma 2. *$F : \mathcal{M}_{\xi_1}(\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ defined in equation (34) is convex. The restriction of its convex conjugate $F^* : \mathcal{M}^*_{\xi_1}(\mathcal{Y}) \to \mathbb{R}$ to the set $C_{b,\xi_1}(\mathcal{Y}) := \{f \in C(\mathcal{Y}) \mid f(\cdot)/\xi_1(\cdot) \in C_b(\mathcal{Y})\} \subseteq \mathcal{M}^*_{\xi_1}(\mathcal{Y})$ is given by equation (35).*

Proof. It is well known that the KL divergence is convex. We compute the (restriction of the) convex conjugate via a classical argument (c.f. Lemma B.37 of Mohri et al. (2012)): for any $q : \mathcal{Z} \to \mathbb{R}$ belonging to $C_{b,\xi_1}(\mathcal{Y})$, define $\tilde{q} \in \mathcal{P}(\mathcal{Y})$ with density $\frac{d\tilde{q}}{d\tau_{\mathcal{Y}}}(y) = \exp(\beta q(y))/\int_{\mathcal{Y}} \exp(\beta q(y'))d\tau_{\mathcal{Y}}(y')$. Then,

$$\begin{aligned}
F^*(q) &= \sup_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y}) \cap \mathcal{P}(\mathcal{Y})} \int_{\mathcal{Y}} q(y)\, d\nu(y) - \beta^{-1} D_{KL}(\nu||\tau_{\mathcal{Y}}) \\
&= \sup_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y}) \cap \mathcal{P}(\mathcal{Y})} \int_{\mathcal{Y}} \log(\exp(q(y)))\, d\nu(y) - \beta^{-1} \int_{\mathcal{Y}} \log\left(\frac{d\nu}{d\tau_{\mathcal{Y}}}(y)\right)\, d\nu(y) \\
&= \sup_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y}) \cap \mathcal{P}(\mathcal{Y})} \beta^{-1} \int_{\mathcal{Y}} \log\left(\frac{d\tau_{\mathcal{Y}}}{d\nu}(y) \exp(\beta q(y))\right)\, d\nu(y) \\
&= \sup_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y}) \cap \mathcal{P}(\mathcal{Y})} \beta^{-1} \int_{\mathcal{Y}} \log\left(\frac{d\tilde{q}}{d\nu}(y)\right)\, d\nu(y) \\
&= \sup_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y}) \cap \mathcal{P}(\mathcal{Y})} -\beta^{-1} D_{KL}(\nu||\tilde{q}) + \beta^{-1} \log\left(\int_{\mathcal{Y}} \exp(\beta q(y'))\, d\tau_{\mathcal{Y}}(y')\right) \\
&= \beta^{-1} \log\left(\int_{\mathcal{Y}} \exp(\beta q(y'))\, d\tau_{\mathcal{Y}}(y')\right)
\end{aligned}$$

It remains to justify the last equality, which follows from checking that $\tilde{q} \in \mathcal{M}_{\xi_1}(\mathcal{Y}) \cap \mathcal{P}(\mathcal{Y})$. For this, we need to see that $\|\tilde{q}\|_{\mathrm{TV},\xi_1}$ is finite making use of Assumption 2(ii):

$$\begin{aligned}
&\xi_1(y) + \log(\xi_1(y)) = o\left(-\log\left(\frac{d\tau_{\mathcal{Y}}}{d\lambda}(y)\right) - (d+\epsilon)\log\|y\|_2\right) \text{ as } \|y\|_2 \to +\infty \\
&\implies \lim_{\|y\|_2 \to +\infty} \log(\xi_1(y)) + q(y) - \log Z_q + \log\left(\frac{d\tau_{\mathcal{Y}}}{d\lambda}(y)\right) + (d+\epsilon)\log\|y\|_2 = -\infty \\
&\implies \exp\left(\log(\xi_1(y)) + q(y) - \log Z_q + \log\left(\frac{d\tau_{\mathcal{Y}}}{d\lambda}(y)\right) + (d+\epsilon)\log\|y\|_2\right) = o(1) \\
&\implies \|\tilde{q}\|_{\mathrm{TV},\xi_1} = \int_{\mathcal{Y}} \xi_1(y)\, d|\tilde{q}|(y) = \int_{\mathcal{Y}} \exp\left(\log(\xi_1(y)) + q(y) - \log Z_q + \log\left(\frac{d\tau_{\mathcal{Y}}}{d\lambda}(y)\right)\right) d\lambda(y) \\
&= \int_{\mathbb{R}_+} \int_{\mathbb{S}^{d-1}} \mathbb{1}_{r\theta \in \mathcal{Y}} \exp\left(\log(\xi_1(r\theta)) + q(r\theta) - \log Z_q + \log\left(\frac{d\tau_{\mathcal{Y}}}{d\lambda}(r\theta)\right) + (d+\epsilon)\log r\right) K_1 r^{-1-\epsilon}\, dr d\lambda(\theta) \\
&\leq \int_{\mathbb{R}_+} \int_{\mathbb{S}^{d-1}} K_2 K_1 r^{-1-\epsilon}\, dr d\lambda(\theta) = \frac{K_2 K_1 \mathrm{vol}(\mathbb{S}^{d-1})}{\epsilon}.
\end{aligned} \tag{43}$$

In this equation, λ denotes the Lebesgue measure over $\mathcal{Y}$. In the last inequality, we use that $\exp\left(\log(\xi_1(y)) + q(y) - \log Z_q + \log\left(\frac{d\tau_{\mathcal{Y}}}{d\lambda}(y)\right) + (d+\epsilon)\log\|y\|_2\right) = o(1)$ to show the existence of some constant bound K_2 of this expression over all $y \in \mathcal{Y}$. □

Lemma 3. *$G : L^2(\mathcal{Z}) \to \mathbb{R} \cup \{+\infty\}$ given by equation (36) is convex. Its convex conjugate $G^* : L^2(\mathcal{Z}) \to \mathbb{R} \cup \{+\infty\}$ is given by equation (37).*

Proof. We can easily check that G is convex by writing

$$\left(\int_{\mathcal{Z}} \big(\psi(z) - g(z)\big)^2 \, d\tau_{\mathcal{Z}}(z)\right)^{1/2} = \sup_{\substack{h \in L^2(\mathcal{Z}), \\ \|h\|_{L^2} \leq 1}} \int_{\mathcal{Z}} \big(\psi(z) - g(z)\big) \, h(z) \, d\tau_{\mathcal{Z}}(x), \tag{44}$$

(in more compact notation $\|\chi - g\|_{L^2(\mathcal{Z})} = \sup_{h \in L^2(\mathcal{Z}),\, \|h\|_{L^2} \leq 1} \langle \chi - g, h \rangle$), and recalling that a supremum of convex functions is convex.

By definition, for any $\psi \in L^2(\mathcal{Z}, \tau_{\mathcal{Z}})$, we have that $G^*(\psi)$ is equal to

$$\begin{aligned}
&\sup_{\chi \in L^2(\mathcal{Z})} \Big\{ \langle \chi, \psi \rangle_{L^2(\mathcal{Z})} - G(\chi) \Big\} = \sup_{\chi \in L^2(\mathcal{Z})} \Big\{ \langle \chi, \psi \rangle_{L^2(\mathcal{Z})} - \|\chi - g\|_{L^2(\mathcal{Z})} \Big\} \\
&= \sup_{\chi \in L^2(\mathcal{Z})} \left\{ \langle \chi, \psi \rangle_{L^2(\mathcal{Z})} - \sup_{\substack{\hat{\psi} \in L^2(\mathcal{Z}), \\ \|\hat{\psi}\|_{L^2} \leq 1}} \langle \chi - g, \hat{\psi} \rangle_{L^2(\mathcal{Z})} \right\} \\
&= \sup_{\chi \in L^2(\mathcal{Z})} \inf_{\substack{\hat{\psi} \in L^2(\mathcal{Z}), \\ \|\hat{\psi}\|_{L^2} \leq 1}} \Big\{ \langle \chi, \psi - \hat{\psi} \rangle_{L^2(\mathcal{Z})} + \langle g, \hat{\psi} \rangle_{L^2(\mathcal{Z})} \Big\}.
\end{aligned} \tag{45}$$

At this point, we want to apply Theorem 5. For that, we set $A = \mathcal{B}_{L^2(\mathcal{Z})} \subseteq E = L^2(\mathcal{Z})$ and $B = F = L^2(\mathcal{Z})$. We can endow B with the strong (or norm) topology, but A requires a weaker topology that makes it compact. We endow A with the weak-* topology, which by the Banach-Alaoglu theorem (Theorem 8) for Hilbert spaces makes it compact. We have that $(\hat{\psi}, \chi) \mapsto H(\hat{\psi}, \chi) = -\langle \chi, \psi - \hat{\psi} \rangle_{L^2(\mathcal{Z})} - \langle g, \hat{\psi} \rangle_{L^2(\mathcal{Z})}$ is concave in $\hat{\psi} \in A$ and convex in $\chi \in B$, because it is affine in both variables. H is continuous in χ (via Cauchy-Schwarz) and it is continuous in $\hat{\psi}$ in the weak-* (or weak) topology, because it is precisely the weakest one that makes maps of the form $\hat{\psi} \mapsto \langle \hat{\psi}, g - \chi \rangle_{L^2(\mathcal{Z})}$ continuous. Thus, we obtain that $\sup_{\hat{\psi} \in A} \inf_{\chi \in B} H(\chi, \hat{\psi}) = \inf_{\chi \in B} \sup_{\hat{\psi} \in A} H(\chi, \hat{\psi})$. Alternatively, if we flip the signs, we get that the right-hand side of (45) is equal to:

$$\begin{aligned}
&\inf_{\substack{\hat{\psi} \in L^2(\mathcal{Z}), \\ \|\hat{\psi}\|_{L^2} \leq 1}} \sup_{\chi \in L^2(\mathcal{Z})} \Big\{ \langle \chi, \psi - \hat{\psi} \rangle_{L^2(\mathcal{Z})} + \langle g, \hat{\psi} \rangle_{L^2(\mathcal{Z})} \Big\} \\
&= \begin{cases} \int_K g(z) \psi(z) \, d\tau_{\mathcal{Z}}(z) & \text{if } \|\psi\|_{L^2} \leq 1, \\ +\infty & \text{otherwise} \end{cases}
\end{aligned}$$

The equality holds because unless $\hat{\psi} = \psi$, the value of the supremum is $+\infty$. □

Lemma 4. *The linear operator $A : \mathcal{M}_{\xi_1}(\mathcal{Y}) \to L^2(\mathcal{Z})$ defined as $(A\nu)(z) = \int_{\mathcal{Y}} \varphi(y, z) \, d\nu(y)$ is well defined and continuous. Its operator norm is upper bounded by $\sup_{y \in \mathcal{Y}} |\xi_2(y)|/\xi_1(y)$, where ξ_1 and ξ_2 are defined in Assumption 2. Moreover, the adjoint operator $A^* : L^2(\mathcal{Z}) \to \mathcal{M}^*_{\xi_1}(\mathcal{Y})$ is defined as $(A^* h)(y) = \int_{\mathcal{Z}} \varphi(y, z) h(z) \, d\tau_{\mathcal{Z}}(z)$.*

Proof. Remark that $\int_{\mathcal{Y}} \varphi(y, \cdot) \, d\nu(y)$ does belong to $L^2(\mathcal{Z})$ because

$$\begin{aligned}
&\int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \varphi(y, z) \, d\nu(y) \right)^2 d\tau_{\mathcal{Z}}(z) = \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \frac{\varphi(y, z)}{\xi_1(y)} \, d\tilde{\nu}(y) \right)^2 d\tau_{\mathcal{Z}}(z) \\
&= \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \frac{\varphi(y, z)}{\xi_1(y)} \|\tilde{\nu}\|_{\mathrm{TV}} \, d\frac{\tilde{\nu}}{\|\tilde{\nu}\|_{\mathrm{TV}}}(y) \right)^2 d\tau_{\mathcal{Z}}(z) \leq \|\tilde{\nu}\|_{\mathrm{TV}} \int_{\mathcal{Z}} \int_{\mathcal{Y}} \left(\frac{\varphi(y, z)}{\xi_1(y)} \right)^2 d|\tilde{\nu}|(y) \, d\tau_{\mathcal{Z}}(z) \\
&= \|\tilde{\nu}\|_{\mathrm{TV}} \int_{\mathcal{Y}} \frac{1}{\xi_1(y)^2} \int_{\mathcal{Z}} \varphi(y, z)^2 \, d\tau_{\mathcal{Z}}(z) \, d|\tilde{\nu}|(y) = \|\tilde{\nu}\|_{\mathrm{TV}} \int_{\mathcal{Y}} \frac{\xi_2(y)^2}{\xi_1(y)^2} \, d|\tilde{\nu}|(y) \\
&\leq \|\tilde{\nu}\|_{\mathrm{TV}}^2 \left(\sup_{y \in \mathcal{Y}} \frac{|\xi_2(y)|}{\xi_1(y)} \right)^2.
\end{aligned}$$

In the first equality we have used the change of variable $\tilde{\nu} = \tilde{\xi_1}(\nu)$. In the first inequality we have used the Cauchy-Schwarz inequality, and in the following equality we used Fubini's theorem, which holds because the integrand is positive. In the last equality we have used the definition of ξ_2 given by Assumption 2(iii). Also by Assumption 2(iii), the right-most expression is finite, implying that $A\nu \in L^2(\mathcal{Z})$. Furthermore, since $\|\tilde{\nu}\|_{\text{TV}} = \|\nu\|_{\text{TV},\xi_1}$, we also conclude that A is a continuous operator with norm bounded by $\sup_{y\in\mathcal{Y}} |\xi_2(y)|/\xi_1(y)$.

We have that $A^* : L^2(\mathcal{Z}) \to \mathcal{M}^*_{\xi_1}(\mathcal{Y})$ is defined as $(A^*h)(y) = \int_{\mathcal{Z}} \varphi(y,z)h(z)\, d\tau_{\mathcal{Z}}(z)$, because

$$\begin{aligned}\langle A\nu, h\rangle &= \int_{\mathcal{Z}} (A\nu)(z)h(z)\, d\tau_{\mathcal{Z}}(z) = \int_{\mathcal{Z}}\int_{\mathcal{Y}} \varphi(y,z)\, d\nu(y)\, h(z)\, d\tau_{\mathcal{Z}}(z) \\ &= \int_{\mathcal{Y}}\int_{\mathcal{Z}} \varphi(y,z)h(z)\, d\tau_{\mathcal{Z}}(z)\, d\nu(y).\end{aligned} \tag{46}$$

In the last equality we have applied Fubini's theorem, which holds because by the Cauchy-Schwarz inequality,

$$\begin{aligned}&\int_{\mathcal{Y}}\int_{\mathcal{Z}} |\varphi(y,z)h(z)|\, d\tau_{\mathcal{Z}}(z)\, d\nu(y) \le \|\tilde{\nu}\|_{\text{TV}} \int_{\mathcal{Y}}\int_{\mathcal{Z}} \frac{|\varphi(y,z)|}{\xi_1(y)}|h(z)|\, d\tau_{\mathcal{Z}}(z)\, d\frac{|\tilde{\nu}|}{\|\tilde{\nu}\|_{\text{TV}}}(y) \\ &\le \left(\int_{\mathcal{Y}}\int_{\mathcal{Z}} \left(\frac{|\varphi(y,z)|}{\xi_1(y)}\right)^2 d\tau_{\mathcal{Z}}(z)\, d|\tilde{\nu}|(y)\right)^{1/2} \left(\int_{\mathcal{Y}}\int_{\mathcal{Z}} |h(z)|^2\, d\tau_{\mathcal{Z}}(z)\, d|\tilde{\nu}|(y)\right)^{1/2} \\ &= \left(\int_{\mathcal{Y}} \left(\frac{\xi_2(y)}{\xi_1(y)}\right)^2 d|\tilde{\nu}|(y)\right)^{1/2} \|h\|_{L^2(\mathcal{Z},\tau_{\mathcal{Z}})}\|\tilde{\nu}\|_{\text{TV}}^{1/2} \le \|\tilde{\nu}\|_{\text{TV}}\|h\|_{L^2(\mathcal{Z},\tau_{\mathcal{Z}})} \sup_{y\in\mathcal{Y}} \frac{|\xi_2(y)|}{\xi_1(y)} < +\infty\end{aligned} \tag{47}$$

As a safety check, notice that when $h \in L^2(\mathcal{Z})$, we have that $\int_{\mathcal{Z}} \varphi(\cdot,z)h(z)\, d\tau_{\mathcal{Z}}(z)$ indeed belongs to $\mathcal{M}^*_{\xi_1}(\mathcal{Y})$, because

$$\int_{\mathcal{Z}} \left|\frac{\varphi(y,z)h(z)}{\xi_1(y)}\right| d\tau_{\mathcal{Z}}(z) \le \left(\int_{\mathcal{Z}} \left|\frac{\varphi(y,z)}{\xi_1(y)}\right|^2 d\tau_{\mathcal{Z}}(z)\right)^{1/2} \|h\|_{L^2(\mathcal{Z},\tau_{\mathcal{Z}})} \le \sup_{y\in\mathcal{Y}} \frac{|\xi_2(y)|}{\xi_1(y)}\|h\|_{L^2(\mathcal{Z},\tau_{\mathcal{Z}})}$$

is uniformly bounded over $y \in \mathcal{Y}$ and thus $\int_{\mathcal{Z}} \varphi(\cdot,z)h(z)\, d\tau_{\mathcal{Z}}(z) \in C_{b,\xi_1}(\mathcal{Y}) \subseteq \mathcal{M}^*_{\xi_1}(\mathcal{Y})$. □

Lemma 5. *(i) Let F as defined in equation* (34)*, A^* as defined in* (38) *and $h^\star$ as in* (41)*. Then, the unique $\nu^\star = \arg\max_{\nu\in\mathcal{M}_{\xi_1}(\mathcal{Y})} \{\langle -A^*h^\star, \nu\rangle - F(\nu)\}$ satisfies*

$$\frac{d\nu^\star}{d\tau_{\mathcal{Y}}}(y) = \frac{1}{Z_{\nu^\star}} \exp\left(-\beta \int_{\mathcal{Z}} \varphi(y,z)\, h^\star(z) d\tau_{\mathcal{Z}}(z)\right).$$

and we also have that $\nu^\star = \arg\max_{\nu\in\mathcal{P}(\mathcal{Y})} \{\langle -A^*h^\star, \nu\rangle - F(\nu)\}$.

(ii) We also have that $\nu^\star = \arg\min_{\nu\in\mathcal{P}(\mathcal{Y})}\{F(\nu) + G(A\nu)\}$.

Proof. First, notice that

$$\arg\max_{\nu\in\mathcal{M}_{\xi_1}(\mathcal{Y})} \{\langle -A^*h^\star, \nu\rangle - F(\nu)\} = \arg\max_{\nu\in\mathcal{M}_{\xi_1}(\mathcal{Y})\cap\mathcal{P}(\mathcal{Y})} \{\langle -A^*h^\star, \nu\rangle - F(\nu)\}$$

because $F(\nu) = +\infty$ when $\nu \notin \mathcal{P}(\mathcal{Y})$. Since the KL-divergence is strictly convex, this problem (42) has a unique solution $\nu^\star$.

Now, define $\nu_1 \in \mathcal{P}(\mathcal{Y})$ with density $\frac{d\nu_1^\star}{d\tau_{\mathcal{Y}}}(y) = \frac{1}{Z_{\nu_1^\star}} \exp\left(-\beta \int_{\mathcal{Z}} \varphi(y,z)\, h^\star(z) d\tau_{\mathcal{Z}}(z)\right)$ (the following arguments show that indeed this measure is normalizable).

Consider the relaxation

$$\arg\max_{\nu\in\mathcal{P}(\mathcal{Y})} \{\langle -A^*h^\star, \nu\rangle - F(\nu)\}. \tag{48}$$

This problem is strictly convex (because the KL divergence is) and it has at most one solution, which is the unique solution of an Euler-Lagrange equation. This Euler-Lagrange equation is satisfied by $\nu_1^\star$, hence $\nu_1^\star = \arg\max_{\nu\in\mathcal{P}(\mathcal{Y})} \left\{\langle -A^* h^\star, \nu\rangle - F(\nu)\right\}$.

We will see that $\nu_1^\star$ belongs to $\mathcal{M}_{\xi_1}(\mathcal{Y})$, which implies that $\nu_1^\star = \nu^\star$. Remark that problem (48) has Euler-Lagrange condition $\frac{d\nu_1^\star}{d\tau_\mathcal{Y}}(y) = \frac{1}{Z_{\nu_1^\star}} \exp\left(-\beta \int_\mathcal{Z} \varphi(y,z)\, h^\star(z) d\tau_\mathcal{Z}(z)\right)$. Next, notice that by the Cauchy-Schwarz inequality and the definition of ξ_2 in Assumption 2(iii),

$$\left|-\beta \int \varphi(y,z)\, h^\star(z) d\tau_\mathcal{Z}(z)\right| \leq \beta \left(\int_\mathcal{Z} \varphi(y,z)^2 d\tau_\mathcal{Z}(z)\right)^{1/2} \left(\int h^\star(z)^2 d\tau_\mathcal{Z}(z)\right)^{1/2} = \beta\xi_2(y)\|h^\star\|_{L^2}. \tag{49}$$

Thus, $-\beta \int \varphi(\cdot,z)\, h^\star(z) d\tau_\mathcal{Z}(z) \in C_{b,\xi_1}(\mathcal{Y})$, and in analogy with (43),

$$\begin{aligned}
\|\nu_1^\star\|_{\mathrm{TV},\xi_1} &= \int_\mathcal{Y} \xi_1(y)\, d\nu_1^\star(y) \\
&= \int_\mathcal{Y} \exp\left(\log(\xi_1(y)) - \beta \int \varphi(y,z)\, h^\star(z) d\tau_\mathcal{Z}(z) - \log Z_{\nu_1^\star} + \log\left(\frac{d\tau_\mathcal{Y}}{d\lambda}(y)\right)\right) d\lambda(y) \\
&\leq \int_\mathcal{Y} \exp\left(\log(\xi_1(y)) + \beta\xi_2(y)\|h^\star\|_{L^2} - \log Z_{\nu_1^\star} + \log\left(\frac{d\tau_\mathcal{Y}}{d\lambda}(y)\right)\right) d\lambda(y) \\
&\leq \int_{\mathbb{R}_+} \int_{\mathbb{S}^{d-1}} K_2 K_1 r^{-1-\epsilon}\, dr d\lambda(\theta) = \frac{K_2 K_1 \mathrm{vol}(\mathbb{S}^{d-1})}{\epsilon}.
\end{aligned}$$

In the second equality we used the Euler-Lagrange condition, in the first inequality we used equation (49) and in the second inequality we skipped a step which proceeds as in (43); the key point is that $\beta\xi_2(\cdot)\|h^\star\|_{L^2}$ is $O(\xi_1)$ by Assumption 2(iii) and thus $\exp\Bigg(\log(\xi_1(y)) + \beta\xi_2(\cdot)\|h^\star\|_{L^2} - \log Z_{\nu_1^\star} + \log\left(\frac{d\tau_\mathcal{Y}}{d\lambda}(y)\right) + (d+\epsilon)\log\|y\|_2\Bigg) = o(1)$.

(ii) Consider the problem

$$\arg\min_{\nu\in\mathcal{P}(\mathcal{Y})} \beta^{-1} D_{KL}(\nu||\tau_\mathcal{Y}) + \left(\int_\mathcal{Z} \left(\int_\mathcal{Y} \varphi(y,z)\, d\nu(y) - g(z)\right)^2 d\tau_\mathcal{Z}(z)\right)^{1/2}.$$

If it exists, the unique solution $\nu_2^\star \in \mathcal{P}(\mathcal{Y})$ of this problem is the unique solution of the following Euler-Lagrange condition:

$$\begin{cases} \frac{d\nu_2^\star}{d\tau_\mathcal{Y}}(y) = \exp\left(-\frac{\beta \int_\mathcal{Z} \varphi(y,z)(\int_\mathcal{Y} \varphi(y',z)\, d\nu_2^\star(y') - g(z))\, d\tau_\mathcal{Z}(z)}{(\int_\mathcal{Z} (\int_\mathcal{Y} \varphi(y',z)\, d\nu_2^\star(y') - g(z))^2\, d\tau_\mathcal{Z}(z))^{1/2}}\right) & \text{if } \int_\mathcal{Y} \varphi(y',\cdot)\, d\nu_2^\star(y') \neq g(\cdot) \\ g(\cdot) = \int_\mathcal{Y} \varphi(y',\cdot)\, d\nu_2^\star(y') & \text{otherwise.} \end{cases} \tag{50}$$

Going back to (41), we observe that for strong duality hold we must have

$$A\nu^\star = \arg\max_{\psi\in L^2(\mathcal{Z})} \left\{\langle h^\star, \psi\rangle - G(\psi)\right\} = \arg\max_{\psi\in L^2(\mathcal{Z})} \left\{\langle h^\star, \psi\rangle - \|\psi - g\|_{L^2(\mathcal{Z})}\right\} \tag{51}$$

The Euler-Lagrange condition for $\arg\max_{\psi\in L^2(\mathcal{Z})} \left\{\langle h^\star, \psi\rangle - \|\psi - g\|_{L^2(\mathcal{Z})}\right\}$ is:

$$h^\star - \frac{\psi - g}{\|\psi - g\|_{L^2(\mathcal{Z})}} = 0,$$

which in the case $\|h^\star\|_{L^2(\mathcal{Z})} \neq 1$ implies that $\psi = g$. Thus, for (51) to hold we must have either $h^\star = \frac{A\nu^\star - g}{\|A\nu^\star - g\|_{L^2(\mathcal{Z})}}$ or $A\nu = g$. In either of the two cases, using part (i) we see that $\nu^\star$ satisfies (50), which means that $\nu^\star = \nu_2^\star$. □

Theorem 9. *Let $\eta_1 : \mathcal{Z} \to \mathbb{R}$ be a strictly positive function such that $\sup_{y\in\mathcal{Y}} \varphi(y,z)/\xi_1(y) = o(\eta_1(z))$ and $g(z) = o(\eta_1(z))$ as $z \to \infty$. Let $\mathcal{M}_{\eta_1}(\mathcal{Z})$ be the space of (countably additive) signed Radon measures γ over $\mathcal{Z}$ such that $\|\gamma\|_{TV,\eta_1} := \int_{\mathcal{Z}} \eta_1(z)\, d|\gamma|(z)$ is finite. Consider the problem*

$$\min_{\nu\in\mathcal{P}(\mathcal{Y})} \beta^{-1} D_{KL}(\nu||\tau_{\mathcal{Y}}) + \left\| \frac{1}{\eta_1(\cdot)} \left(\int_{\mathcal{Y}} \varphi(y,\cdot)\, d\nu(y) - g(\cdot) \right) \right\|_{L^\infty}. \tag{52}$$

and the problem

$$\max_{\substack{\gamma\in\mathcal{M}_{\eta_1}(\mathcal{Z}) \\ \|\gamma\|_{TV,\eta_1}\le 1}} -\int_{\mathcal{Z}} g(z)\, d\gamma(z) - \frac{1}{\beta} \log \left(\int_{\mathcal{Y}} \exp\left(-\beta \int_{\mathcal{Z}} \varphi(y,z)\, d\gamma(z) \right) d\tau_{\mathcal{Y}}(y) \right) \tag{53}$$

The two problems (52) *and* (53) *are convex. The problem* (53) *is the dual problem of* (52)*. Moreover, the solution $\nu^\star$ of* (52) *is unique and its density satisfies*

$$\frac{d\nu^\star}{d\tau_{\mathcal{Y}}}(y) = \frac{1}{Z_\beta} \exp\left(-\beta \int \varphi(y,z)\, d\gamma^\star(z) \right),$$

where $\gamma^\star$ is a solution of (53) *and Z_β is a normalization constant.*

Proof. We apply Theorem 4.

As in the proof of Theorem 2, we set $X = \mathcal{M}_{\xi_1}(\mathcal{Y})$, which is the Banach space of Radon measures over $\mathcal{Y}$ such that the weighted total variation $\|\nu\|_{\text{TV},\xi_1} := \int_{\mathcal{Y}} \xi_1(y)\, d|\nu|(y)$ is finite, and whose continuous dual $\mathcal{M}^*_{\xi_1}(\mathcal{Y})$ contains the set $C_{b,\xi_1}(\mathcal{Y})$ of continuous functions f such that $f/\xi_1 \in C_b(\mathcal{Y})$.

Unlike in the proof of Theorem 2, we set $Y = C_{0,\eta_1}(\mathcal{Z})$, which we define to be the space of continuous functions $f : \mathcal{Z} \to \mathbb{R}$ such that $\lim_{\|z\|\to\infty} f(z)/\eta_1(z) = 0$. By Lemma 6, $C_{0,\eta_1}(\mathcal{Z})$ is a Banach space endowed with the norm $\|f\|_{C_{0,\eta_1}} = \sup_{z\in\mathcal{Z}} f(z)/\eta_1(z)$, and we have that the continuous dual space $C^*_{0,\eta_1}(\mathcal{Z})$ is equal to the set $\mathcal{M}_{\eta_1}(\mathcal{Z})$ of Radon measures γ over $\mathcal{Z}$ such that $\|\gamma\|_{\text{TV},\eta_1} := \int_{\mathcal{Z}} \eta_1(z)\, d|\gamma|(z)$ is finite.

$F : \mathcal{M}_{\xi_1}(\mathcal{Y}) \to \mathbb{R}\cup\{+\infty\}$ and its convex conjugate $F^* : \mathcal{M}^*_{\xi_1}(\mathcal{Y}) \to \mathbb{R}\cup\{+\infty\}$ are as specified in equations (34)-(35) in the proof of Theorem 2.

Define $G : C_{0,\eta_1}(\mathcal{Z}) \to \mathbb{R}\cup\{+\infty\}$ as

$$G(\psi) = \sup_{\substack{\gamma\in\mathcal{M}_{\eta_1}(\mathcal{Z}) \\ \|\gamma\|_{\text{TV},\eta_1}\le 1}} \int_{\mathcal{Z}} (\psi(z) - g(z))\, d\gamma(z),$$

which by Lemma 7 can also be written as

$$G(\psi) = \sup_{z\in\mathcal{Z}} \frac{|\psi(z) - g(z)|}{\eta_1(z)}.$$

Also by Lemma 7, the convex conjugate $G^* : \mathcal{M}_{\eta_1}(\mathcal{Z}) \to \mathbb{R}$ is of the form

$$G^*(\gamma) = \begin{cases} \int_{\mathcal{Z}} g(z)\, d\gamma(z) & \text{if } \|\gamma\|_{\text{TV},\eta_1} \le 1 \\ +\infty & \text{otherwise} \end{cases}$$

The linear operator $A : \mathcal{M}_{\xi_1}(\mathcal{Y}) \to C_{0,\eta_1}(\mathcal{Z})$ is defined as $(A\nu)(z) = \int_{\mathcal{Y}} \varphi(y,z)\, d\nu(y)$. It is well defined and continuous by Lemma 8. Lemma 8 also states that the adjoint operator $A^* : \mathcal{M}_{\eta_1}(\mathcal{Z}) \to \mathcal{M}^*_{\xi_1}(\mathcal{Y})$ is $(A^*\gamma)(y) = \int_{\mathcal{Z}} \varphi(y,z)\, d\gamma(z)$. Hence, we have that problem (52) can be written as

$$p^* = \inf_{\nu\in\mathcal{M}_{\xi_1}(\mathcal{Y})} \{F(\nu) + G(A\nu)\}.$$

And problem (53) can be written as

$$d^* = \sup_{\substack{\gamma \in \mathcal{M}_{\eta_1}(\mathcal{Z}) \\ \|\gamma\|_{\mathrm{TV},\eta_1} \leq 1}} \{-F^*(-A^*\gamma) - G^*(\gamma)\}. \tag{54}$$

To apply Theorem 4, it only remains to show that condition 2 holds. That is, we have to check that $A \text{ dom } F \cap \text{cont } G \neq \emptyset$. Consider $\psi = A\nu$ for some $\nu \in \mathcal{P}(\mathcal{Y})$ absolutely continuous w.r.t. $\tau_{\mathcal{Y}}$. Then $\psi \in A \text{ dom } F$. Moreover, since G is a continuous functional, we have that $\text{cont } G = C_{0,\eta_1}(\mathcal{Z})$. Thus, ψ also belongs to $\text{cont } G$ and we conclude that $A \text{ dom } F \cap \text{cont } G \neq \emptyset$.

By Theorem 4, $p^* = d^*$, and since p^* is finite, we have that the supremum in (54) is attained; let $\gamma^\star$ be one maximizer. As in the proof of Theorem 2, we prove the existence of a minimizer $\nu^\star$ of p^* by the direct method of the calculus of variations, and the link between $\gamma^\star$ and $\nu^\star$ is analogous. □

Lemma 6. *Let $C_{0,\eta_1}(\mathcal{Z})$ be the vector space of functions $f : \mathcal{Z} \to \mathbb{R}$ such that $\lim_{\|z\|\to\infty} f(z)/\eta_1(z) = 0$. $C_{0,\eta_1}(\mathcal{Z})$ is a Banach space with norm $\|f\|_{C_{0,\eta_1}} = \sup_{z\in\mathcal{Z}} f(z)/\eta_1(z)$. The continuous dual space $C^*_{0,\eta_1}(\mathcal{Z})$ is equal to the set $\mathcal{M}_{\eta_1}(\mathcal{Z})$ of Radon measures γ over $\mathcal{Z}$ such that $\|\gamma\|_{TV,\eta_1} := \int_{\mathcal{Z}} \eta_1(z)\, d|\gamma|(z)$ is finite.*

Proof. Define the linear map $\hat{\eta_1} : C_{0,\eta_1}(\mathcal{Z}) \to C_0(\mathcal{Z})$ as $f \mapsto \tilde{\eta_1}(f) := f/\eta_1$, where $C_0(\mathcal{Z})$ is the Banach space of continuous functions on $\mathcal{Z}$ vanishing at infinity, endowed with the supremum norm. Notice that for all $f \in C_{0,\eta_1}(\mathcal{Z})$, we have that $\|\tilde{\eta_1}(f)\|_{C_0} = \|f\|_{C_{0,\eta_1}}$. Remark also that $\tilde{\eta_1}$ is surjective, because if $\tilde{f} \in C_0(\mathcal{Z})$, there exists $f := \eta_1\tilde{f}$ such that $\tilde{\eta_1}(f) = \tilde{f}$. Thus, $\tilde{\eta_1}$ is a surjective isometry between $C_{0,\eta_1}(\mathcal{Z})$ and $C_0(\mathcal{Z})$, which shows that $C_{0,\eta_1}(\mathcal{Z})$ is a Banach space.

And in analogy with Lemma 1, the linear mapping $\tilde{\eta_1} : \mathcal{M}_{\eta_1}(\mathcal{Z}) \to \mathcal{M}(\mathcal{Z})$ defined as $\gamma \mapsto \tilde{\gamma}$ such that $\frac{d\tilde{\gamma}}{d\gamma}(z) = \eta_1(z)$ is a surjective isometry. To show that $C^*_{0,\eta_1}(\mathcal{Z})$ is $\mathcal{M}_{\eta_1}(\mathcal{Z})$, we will show both inclusions. Given $\gamma \in \mathcal{M}_{\eta_1}(\mathcal{Z})$, for any $f \in C_{0,\eta_1}(\mathcal{Z})$ we have that

$$\begin{aligned}\int_{\mathcal{Z}} f(z)\, d\gamma(z) &= \int_{\mathcal{Z}} \frac{f(z)}{\eta_1(z)}\eta_1(z)\, d\gamma(z) = \int_{\mathcal{Z}} \hat{\eta_1}(f)(z)\, d\tilde{\eta_1}(\gamma)(z) \\ &\leq \|\hat{\eta_1}(f)\|_{C_0}\|\tilde{\eta_1}(\gamma)\|_{\mathrm{TV}} = \|f\|_{C_{0,\eta_1}}\|\gamma\|_{\mathrm{TV},\eta_1}.\end{aligned}$$

Thus, $\mathcal{M}_{\eta_1}(\mathcal{Z}) \subseteq C^*_{0,\eta_1}(\mathcal{Z})$. Conversely, let $\gamma \in C^*_{0,\eta_1}(\mathcal{Z})$. Since $\hat{\eta_1}^{-1} : C_0(\mathcal{Z}) \to C_{0,\eta_1}(\mathcal{Z})$ is a surjective isometry, we have that $\gamma \circ \hat{\eta_1}^{-1} \in C^*_0(\mathcal{Z})$. By the Riesz-Markov-Kakutani theorem (Theorem 6) the continuous dual space $C^*_0(\mathcal{Z})$ is the Banach space $\mathcal{M}(\mathcal{Z})$ of finite signed Radon measures with norm $\|\cdot\|_{\mathrm{TV}}$. Thus, there exists $\tilde{\gamma} \in \mathcal{M}(\mathcal{Z})$ such that for any $\hat{f} \in C_0(\mathcal{Z})$, $\int_{\mathcal{Z}} \hat{f}(z)\, d\tilde{\gamma}(z) = \langle \gamma \circ \hat{\eta_1}^{-1}, \hat{f}\rangle$. Since $\langle \gamma \circ \hat{\eta_1}^{-1}, \hat{f}\rangle = \langle \gamma, \hat{\eta_1}^{-1}(\hat{f})\rangle$ and $\int_{\mathcal{Z}} f(z)\, d\tilde{\gamma}(z) = \int_{\mathcal{Z}} \hat{f}(z)\eta_1(z)\frac{1}{\eta_1(z)}\, d\tilde{\gamma}(z) = \int_{\mathcal{Z}} \hat{\eta_1}^{-1}(\hat{f})(z)\, d\tilde{\eta_1}^{-1}(\tilde{\gamma})(z)$, we have that

$$\forall f \in C_{0,\eta_1}(\mathcal{Z}), \quad \int_{\mathcal{Z}} f(z)\, d\tilde{\eta_1}^{-1}(\tilde{\gamma})(z) = \langle \gamma, f\rangle,$$

proving that $C^*_{0,\eta_1}(\mathcal{Z}) \subseteq \mathcal{M}_{\eta_1}(\mathcal{Z})$. □

Lemma 7. *$G : C_{0,\eta_1}(\mathcal{Z}) \to \mathbb{R} \cup \{+\infty\}$ given by equation (36) is convex. Its convex conjugate $G^* : \mathcal{M}_{\eta_1}(\mathcal{Z}) \to \mathbb{R} \cup \{+\infty\}$ is given by equation (37).*

Proof. G is convex because it is the supremum of linear functions. $G^* : \mathcal{M}_{\eta_1}(\mathcal{Z}) \to \mathbb{R}$ is defined as

$$\begin{aligned} G^*(\gamma) &= \sup_{\psi \in C_{0,\eta_1}(\mathcal{Z})} \left\{ \int_{\mathcal{Z}} \psi(z)\, d\gamma(z) - \sup_{\substack{\gamma' \in \mathcal{M}_{\eta_1}(\mathcal{Z}) \\ \|\gamma'\|_{\mathrm{TV},\eta_1} \leq 1}} \int_{\mathcal{Z}} (\psi(z) - g(z))\, d\gamma'(z) \right\} \\ &= \sup_{\psi \in C_{0,\eta_1}(\mathcal{Z})} \inf_{\substack{\gamma' \in \mathcal{M}_{\eta_1}(\mathcal{Z}) \\ \|\gamma'\|_{\mathrm{TV},\eta_1} \leq 1}} \left\{ \int_{\mathcal{Z}} \psi(z)\, d(\gamma - \gamma')(z) + \int_{\mathcal{Z}} g(z)\, d\gamma'(z) \right\} \end{aligned} \tag{55}$$

At this point, we want to apply Theorem 5 in a similar fashion to the proof of Lemma 3. In this case, we set $A = \mathcal{B}_{\mathcal{M}_{\eta_1}(\mathcal{Z})} \subseteq E = \mathcal{M}_{\eta_1}(\mathcal{Z})$ and $B = F = \mathcal{M}_{\eta_1}(\mathcal{Z})$. We can endow B with the strong (or norm) topology, but A requires a weaker topology that makes it compact. Since $\mathcal{M}_{\eta_1}(\mathcal{Z})$ is the continuous dual of $C_{0,\eta_1}(\mathcal{Z})$, we endow A with the weak-* topology, which by the Banach-Alaoglu theorem (Theorem 8) makes it compact. We have that $(\gamma', \psi) \mapsto H(\gamma', \psi) = -\int_{\mathcal{Z}} \psi(z)\, d(\gamma - \gamma')(z) - \int_{\mathcal{Z}} g(z)\, d\gamma'(z)$ is concave in $\gamma' \in A$ and convex in $\psi \in B$ because it is affine in both variables. H is continuous in ψ because $\int_{\mathcal{Z}} \psi(z)\, d(\gamma - \gamma')(z) = \int_{\mathcal{Z}} \frac{\psi(z)}{\eta_1(z)} \eta_1(z)\, d(\gamma - \gamma')(z) \leq \|\psi\|_{C_{0,\eta_1}} \|\gamma - \gamma'\|_{\text{TV},\eta_1}$. H is continuous in γ' in the weak-* topology, because it is precisely the weakest one that makes maps of the form $\gamma' \mapsto \int_{\mathcal{Z}} (g(z) - \psi(z))\, d\gamma'(z)$ continuous. Thus, $\sup_{\gamma' \in A} \inf_{\psi \in B} H(\gamma', \psi) = \sup_{\gamma' \in A} \inf_{\psi \in B} H(\gamma', \psi)$, and flipping the signs, the right-hand side of (55) is equal to:

$$\inf_{\substack{\gamma' \in \mathcal{M}_{\eta_1}(\mathcal{Z}) \\ \|\gamma'\|_{\text{TV},\eta_1} \leq 1}} \sup_{\psi \in C_{0,\eta_1}(\mathcal{Z})} \left\{ \int_{\mathcal{Z}} \psi(z)\, d(\gamma - \gamma')(z) + \int_{\mathcal{Z}} g(z)\, d\gamma'(z) \right\}$$
$$= \begin{cases} \int_{\mathcal{Z}} g(z)\, d\gamma(z) & \text{if } \|\gamma\|_{\text{TV},\eta_1} \leq 1 \\ +\infty & \text{otherwise} \end{cases}.$$

□

Lemma 8. *The linear operator $A : \mathcal{M}_{\xi_1}(\mathcal{Y}) \to C_{0,\eta_1}(\mathcal{Z})$ defined as $(A\nu)(z) = \int_{\mathcal{Y}} \varphi(y, z)\, d\nu(y)$ is well defined and continuous. Its operator norm is upper bounded by $\sup_{y \in \mathcal{Y}} |\xi_2(y)|/\xi_1(y)$, where ξ_1 and ξ_2 are defined in Assumption 2. Moreover, the adjoint operator $A^* : \mathcal{M}_{\eta_1}(\mathcal{Z}) \to \mathcal{M}^*_{\xi_1}(\mathcal{Y})$ is $(A^*\gamma)(y) = \int_{\mathcal{Z}} \varphi(y, z)\, d\gamma(z)$.*

Proof. Remark that $\int_{\mathcal{Y}} \varphi(y, \cdot)\, d\nu(y)$ does belong to $C_{0,\eta_1}(\mathcal{Z})$ because

$$\lim_{\|z\| \to \infty} \frac{\int_{\mathcal{Y}} \varphi(y, z)\, d\nu(y)}{\eta_1(z)} = \int_{\mathcal{Y}} \lim_{\|z\| \to \infty} \frac{\varphi(y, z)}{\eta_1(z)}\, d\nu(y) = 0$$

The second equality follows from the assumption that $\varphi(y, z) = o(\eta_1(z))$ for all $y \in \mathcal{Y}$. The first equality holds by the dominated convergence theorem, which can be applied because the integral of the absolute value can be uniformly dominated for all $z \in \mathcal{Z}$:

$$\begin{aligned} \int_{\mathcal{Y}} \left| \frac{\varphi(y, z)}{\eta_1(z)} \right| d|\nu|(y) = \int_{\mathcal{Y}} \left| \frac{\varphi(y, z)}{\eta_1(z)\xi_1(y)} \right| \xi_1(y)\, d|\nu|(y) &\leq \|\nu\|_{\text{TV},\xi_1} \sup_{y \in \mathcal{Y}} \frac{\varphi(y, z)}{\eta_1(z)\xi_1(y)} \\ &\leq \|\nu\|_{\text{TV},\xi_1} K, \end{aligned} \tag{56}$$

for some constant K. In the first inequality we used the definition of $\|\cdot\|_{\text{TV},\xi_1}$. In the last inequality we used that $\sup_{y \in \mathcal{Y}} \varphi(y, z)/\xi_1(y) = o(\eta_1(z))$ as $\|z\| \to \infty$ by the definition of η_1. Equation (56) also proves that A is continuous, because $\|A\nu\|_{C_{0,\eta_1}} = \sup_{z \in \mathcal{Z}} |\int_{\mathcal{Y}} \varphi(y, z)\, d\nu(y)|/\eta_1(z)$.

We have that $A^* : \mathcal{M}_{\eta_1}(\mathcal{Z}) \to \mathcal{M}^*_{\xi_1}(\mathcal{Y})$ is defined as $(A^*\gamma)(y) = \int_{\mathcal{Z}} \varphi(y, z)\, d\gamma(z)$, because

$$\int_{\mathcal{Z}} (A\nu)(z)\, d\gamma(z) = \int_{\mathcal{Z}} \int_{\mathcal{Y}} \varphi(y, z)\, d\nu(y)\, d\gamma(z) = \int_{\mathcal{Y}} \int_{\mathcal{Z}} \varphi(y, z)\, d\gamma(z)\, d\nu(y).$$

In the last equality we applied Fubini's theorem, which holds because

$$\begin{aligned} &\int_{\mathcal{Z}} \int_{\mathcal{Y}} |\varphi(y, z)|\, d|\nu|(y)\, d|\gamma|(z) = \int_{\mathcal{Z}} \int_{\mathcal{Y}} \frac{|\varphi(y, z)|}{\eta_1(z)\xi_1(y)} \xi_1(y)\, d|\nu|(y)\, \eta_1(z)\, d|\gamma|(z) \\ &\leq \|\nu\|_{\text{TV},\xi_1} \|\gamma\|_{\text{TV},\eta_1} \sup_{y \in \mathcal{Y}} \frac{\varphi(y, z)}{\eta_1(z)\xi_1(y)} = \|\nu\|_{\text{TV},\xi_1} \|\gamma\|_{\text{TV},\eta_1} K, \end{aligned}$$

for some constant K. □

Theorem 3. *The problems* (18) *and* (19) *are convex. Suppose that Assumption 2 holds and also that (i) there exists $K > 0$ such that $\sup_{z\in\mathcal{Z}} \sup_{y\in\mathcal{Y}} \varphi(y,z)/\xi_1(y) < K$, and (ii) $g(\cdot) \in C_b(\mathcal{Z})$. Then problem* (19) *is the Fenchel dual of problem* (18)*, and strong duality holds. Moreover, the solution $\nu^\star$ of* (18) *is unique and its density satisfies*

$$\frac{d\nu^\star}{d\tau_\mathcal{Y}}(y) = \frac{1}{Z_\beta} \exp\left(-\beta \int_\mathcal{Z} \varphi(y,z)\, d\gamma^\star(z)\right),$$

where $\gamma^\star$ is a solution of (19) *and Z_β is a normalization constant.*

Proof. The proof makes use of Theorem 9. We choose η_1 to be in the family $C = \{\eta_r : \mathcal{Z} \to \mathbb{R},\ \eta_r(z) = \max\{\exp(\|z\| - r), 1\} \mid r \in (0, +\infty)\}$. First, we prove that

$$\begin{aligned}
&\sup_{\eta_1\in C} \max_{\substack{\gamma\in\mathcal{M}_{\eta_1}(\mathcal{Z}) \\ \|\gamma\|_{\mathrm{TV},\eta_1}\le 1}} -\int_\mathcal{Z} g(z)\, d\gamma(z) - \frac{1}{\beta}\log\left(\int_\mathcal{Y} \exp\left(-\beta\int_\mathcal{Z}\varphi(y,z)\, d\gamma(z)\right) d\tau_\mathcal{Y}(y)\right) \\
&= \max_{\substack{\gamma\in\mathcal{M}(\mathcal{Z}) \\ \|\gamma\|_{\mathrm{TV}}\le 1}} -\int_\mathcal{Z} g(z)\, d\gamma(z) - \frac{1}{\beta}\log\left(\int_\mathcal{Y} \exp\left(-\beta\int_\mathcal{Z}\varphi(y,z)\, d\gamma(z)\right) d\tau_\mathcal{Y}(y)\right).
\end{aligned} \tag{57}$$

The right-hand side is larger or equal than the left-hand side because for all $r \in (0, +\infty)$, we have that $\mathcal{M}_{\eta_r}(\mathcal{Z}) \subseteq \mathcal{M}(\mathcal{Z})$, as $\|\gamma\|_{\mathrm{TV},\eta_r} \ge \|\gamma\|_{\mathrm{TV}}$.

Lemma 9(i) states that $\{\mathcal{M}_{\eta_r}(\mathcal{Z}) \mid r \in (0, +\infty)\}$ is dense in $\mathcal{M}(\mathcal{Z})$ in the TV norm topology. Lemma 9(ii) states that the objective functional of (57) is continuous in the TV norm topology. These two facts imply the equality in (57). To show that the maximum is attained in the left-hand side, we apply Lemma 9(iii). Let $\gamma^\star$ be a maximizer.

Second, we prove that

$$\begin{aligned}
&\sup_{\eta_1\in C} \min_{\nu\in\mathcal{M}_{\xi_1}(\mathcal{Y})} \beta^{-1} D_{KL}(\nu||\tau_\mathcal{Y}) + \left\| \frac{1}{\eta_1(\cdot)} \left(\int_\mathcal{Y} \varphi(y,\cdot)\, d\nu(y) - g(\cdot)\right)\right\|_{L^\infty} \\
&= \min_{\nu\in\mathcal{M}_{\xi_1}(\mathcal{Y})} \beta^{-1} D_{KL}(\nu||\tau_\mathcal{Y}) + \left\| \int_\mathcal{Y} \varphi(y,\cdot)\, d\nu(y) - g(\cdot)\right\|_{L^\infty}.
\end{aligned} \tag{58}$$

Remark that for all $\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y})$, $\left\|\int_\mathcal{Y} \varphi(y,\cdot)\, d\nu(y) - g(\cdot)\right\|_{L^\infty}$ is finite because $g \in C_b(\mathcal{Z})$ and $\int_\mathcal{Y} \varphi(y,\cdot)\, d\nu(y) \in C_b(\mathcal{Z})$ because

$$\sup_{z\in\mathcal{Z}} \left|\int_\mathcal{Y} \varphi(y,z)\, d\nu(y)\right| = \sup_{z\in\mathcal{Z}} \left|\int_\mathcal{Y} \frac{\varphi(y,z)}{\xi_1(y)} \xi_1(y)\, d\nu(y)\right| \le K\|\nu\|_{\mathrm{TV},\eta_1}. \tag{59}$$

Given $\delta > 0$, let $R > 0$ such that $\|\int_\mathcal{Y} \varphi(y,\cdot)\ d\nu(y) - g(\cdot)\|_{L^\infty} - \sup_{z\in\mathcal{Z}\cap\mathcal{B}_{\mathbb{R}^{d_2}}(R)} |\int_\mathcal{Y} \varphi(y,\cdot)\ d\nu(y) - g(\cdot)| \le \delta$. Hence, for $r > R$, $\|\int_\mathcal{Y} \varphi(y,\cdot)\ d\nu(y) - g(\cdot)\|_{L^\infty} - \|(\int_\mathcal{Y} \varphi(y,\cdot)\, d\nu(y) - g(\cdot))/\eta_r(\cdot)\|_{L^\infty} \le \delta$. Thus, equality holds in (58). By the direct method of the calculus of variations (see the proof of Theorem 2), we have that a minimizer $\nu^\star$ for the right-hand side of (58) exists.

Applying Theorem 9 on the right-hand sides of equations (57) and (58), we see that they are equal. Thus, the left-hand sides are equal. Let us set F and F^* as in the proof of Theorem 9. Let us set $G : C_b(\mathcal{Y}) \to \mathbb{R}$ as

$$G(\psi) = \|\psi(\cdot) - g(\cdot)\|_{L^\infty} = \sup_{\gamma'\in\mathcal{M}(\mathcal{Z}),\|\gamma'\|_{\mathrm{TV}}\le 1} \int_\mathcal{Z} (\psi(z) - g(z))\, d\gamma'(z), \tag{60}$$

and define $\tilde{G} : \mathcal{M}(\mathcal{Z}) \to \mathbb{R}$ as

$$\begin{cases} \int_\mathcal{Z} g(z)\, d\gamma(z) & \text{if } \|\gamma\|_{\mathrm{TV}} \le 1 \\ +\infty & \text{otherwise} \end{cases}. \tag{61}$$

By Lemma 11, for any $\gamma \in \mathcal{M}(\mathcal{Z})$, we have $\tilde{G}(\gamma) = \sup_{\psi \in C_b(\mathcal{Z})} \big\{\langle \gamma, \psi \rangle - G(\psi)\big\}$.

Then, the equality between the left-hand sides of (57) and (58) can be rewritten as $\sup_{\gamma \in \mathcal{M}(\mathcal{Z}), \|\gamma\|_{\text{TV}} \leq 1} \{-F^*(-A^*\gamma) - \tilde{G}(\gamma)\} = \inf_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y})} \{F(\nu) + G(A\nu)\}$. We reproduce the argument of (41) and we conclude that

$$
\begin{aligned}
&\sup_{\substack{\gamma \in \mathcal{M}(\mathcal{Z}) \\ \|\gamma\|_{\text{TV}} \leq 1}} \{-F^*(-A^*\gamma) - G^*(\gamma)\} = -F^*(-A^*\gamma^\star) - G^*(\gamma^\star) \\
&= - \sup_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y})} \big\{\langle -A^*\gamma^\star, \nu \rangle - F(\nu)\big\} - \sup_{\psi \in C_b(\mathcal{Z})} \big\{\langle \gamma^\star, \psi \rangle - G(\psi)\big\} \\
&\leq - \sup_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y})} \big\{\langle -A^*\gamma^\star, \nu \rangle - F(\nu) + \langle \gamma^\star, A\nu \rangle - G(A\nu)\big\} \\
&= - \sup_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y})} \big\{\langle -F(\nu) - G(A\nu)\big\} = \inf_{\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y})} \{F(\nu) + G(A\nu)\} \\
&= F(\nu^\star) + G(A\nu^\star)
\end{aligned}
$$

In the first equality we used that $G^*(\gamma^\star) = \sup_{\psi \in C_b(\mathcal{Z})} \big\{\langle \gamma^\star, \psi \rangle - G(\psi)\big\}$, which holds because for $\gamma \in \mathcal{M}(\mathcal{Z})$,

$$
\begin{aligned}
&\sup_{\psi \in C_b(\mathcal{Z})} \left\{ \int_{\mathcal{Z}} \psi(z)\, d\gamma(z) - \sup_{\substack{\gamma' \in \mathcal{M}(\mathcal{Z}) \\ \|\gamma'\|_{\text{TV}} \leq 1}} \int_{\mathcal{Z}} (\psi(z) - g(z))\, d\gamma'(z) \right\} \\
&= \sup_{\psi \in C_b(\mathcal{Z})} \inf_{\substack{\gamma' \in \mathcal{M}(\mathcal{Z}) \\ \|\gamma'\|_{\text{TV}} \leq 1}} \left\{ \int_{\mathcal{Z}} \psi(z)\, d(\gamma - \gamma')(z) + \int_{\mathcal{Z}} g(z)\, d\gamma'(z) \right\} \\
&= \inf_{\substack{\gamma' \in \mathcal{M}(\mathcal{Z}) \\ \|\gamma'\|_{\text{TV}} \leq 1}} \sup_{\psi \in C_b(\mathcal{Z})} \left\{ \int_{\mathcal{Z}} \psi(z)\, d(\gamma - \gamma')(z) + \int_{\mathcal{Z}} g(z)\, d\gamma'(z) \right\} \\
&= \begin{cases} \int_{\mathcal{Z}} g(z)\, d\gamma(z) & \text{if } \|\gamma\|_{\text{TV}} \leq 1 \\ +\infty & \text{otherwise} \end{cases}.
\end{aligned}
$$

The link between $\gamma^\star$ and $\nu^\star$ is analogous to the proof of Theorem 2 (see Lemma 5(i)). The fact that $\nu^\star = \arg\min_{\nu \in \mathcal{P}(\mathcal{Y})} \{F(\nu) + G(A\nu)\}$ holds by an analogous reasoning. □

Lemma 9. *(i) For any $r > 0$ let $\eta_r : \mathcal{Z} \to \mathbb{R}$, $\eta_r(z) = \max\{\exp(\|z\| - r), 1\}$. The set $\{\mathcal{M}_{\eta_r}(\mathcal{Z}) \mid r \in (0, +\infty)\}$ is dense in $\mathcal{M}(\mathcal{Z})$ in the TV norm topology.*

(ii) The functional $\gamma \mapsto -\int_{\mathcal{Z}} g(z)\, d\gamma(z) - \frac{1}{\beta} \log\left(\int_{\mathcal{Y}} \exp\left(-\beta \int_{\mathcal{Z}} \varphi(y, z)\, d\gamma(z)\right) d\tau_{\mathcal{Y}}(y)\right)$ is continuous in the TV norm topology, and a fortiori, its first variation has bounded supremum norm.

(iii) The functional $\gamma \mapsto -\int_{\mathcal{Z}} g(z)\, d\gamma(z) - \frac{1}{\beta} \log\left(\int_{\mathcal{Y}} \exp\left(-\beta \int_{\mathcal{Z}} \varphi(y, z)\, d\gamma(z)\right) d\tau_{\mathcal{Y}}(y)\right)$ has a maximizer over $\mathcal{B}_{\mathcal{M}(\mathcal{Z})} = \{\gamma \in \mathcal{M}(\mathcal{Z}) \mid \|\gamma\|_{TV} \leq 1\}$.

Proof. To prove (i), let $(r_n)_{n \geq 0}$ be a real sequence converging to $+\infty$. For any $\gamma \in \mathcal{M}(\mathcal{Z})$, we can build a sequence of measures $\gamma_{r_n} \in \mathcal{M}_{\eta_{r_n}}(\mathcal{Z})$ defined with density $\frac{d\gamma_{r_n}}{d\gamma}(z) = \min\{\exp(-\|z\| + r), 1\}$. For any $\delta > 0$, there exists $R > 0$ such that $\|\gamma\|_{\text{TV}} - \int_{\mathcal{Z} \cap \mathcal{B}_{\mathbb{R}^{d_2}}(R)} d|\gamma|(z) \leq \delta$. Notice that for all $r_n > R$,

$$
\|\gamma - \gamma_{r_n}\|_{\text{TV}} \leq \int_{\mathcal{Z} \setminus \mathcal{B}_{\mathbb{R}^{d_2}}(r_n)} d|\gamma|(z) \leq \delta
$$

To prove (ii), notice that the first variation of the log-partition at γ is the function

$$z \mapsto \frac{-\beta \int_{\mathcal{Y}} \exp\left(-\beta \int_{\mathcal{Z}} \varphi(y,z')\, d\gamma(z')\right) \varphi(y,z)\, d\tau_{\mathcal{Y}}(y)}{\int_{\mathcal{Y}} \exp\left(-\beta \int_{\mathcal{Z}} \varphi(y,z')\, d\gamma(z')\right) d\tau_{\mathcal{Y}}(y)},$$

which has supremum norm bounded by

$$\frac{\beta K \int_{\mathcal{Y}} \xi_1(y) \exp\left(-\beta \int_{\mathcal{Z}} \varphi(y,z')\, d\gamma(z')\right) d\tau_{\mathcal{Y}}(y)}{\int_{\mathcal{Y}} \exp\left(-\beta \int_{\mathcal{Z}} \varphi(y,z')\, d\gamma(z')\right) d\tau_{\mathcal{Y}}(y)}.$$

And this bound is finite because we can apply the argument of (43) with $q(y) = \int_{\mathcal{Z}} \varphi(\cdot, z')\, d\gamma(z')$, as

$$\sup_{y\in\mathcal{Y}} \left| \frac{\int_{\mathcal{Z}} \varphi(y,z')\, d\gamma(z')}{\xi_1(y)} \right| \leq \int_{\mathcal{Z}} \left| \frac{\varphi(y,z')}{\xi_1(y)} \right| d\gamma(z') \leq K\|\gamma\|_{\text{TV}}, \implies \int_{\mathcal{Z}} \varphi(\cdot,z')\, d\gamma(z') \in C_{b,\xi_1}(\mathcal{Z}).$$

Moreover, the first variation of the map $\gamma \mapsto -\int_{\mathcal{Z}} g(z)\, d\gamma(z)$ is $-g$, which also has bounded supremum norm by the assumption of Theorem 3.

To prove the existence of a maximizer in (iii), we use the direct method of the calculus of variations. The functional is concave; the first term is linear and the second term is the negated convex conjugate of the KL-divergence composed with a linear map. We cannot use the TV norm topology for $\mathcal{M}(\mathcal{Z})$, because it does not make $\mathcal{B}_{\mathcal{M}(\mathcal{Z})}$ compact. We observe that the weak-* topology of $\text{rba}(\mathcal{Z})$ is the right choice. Here, $\text{rba}(\mathcal{Z})$ is the space of finitely additive finite signed regular Borel measures, which contains the space $\mathcal{M}(\mathcal{Z})$ of countably additive finite signed regular Borel measures, and it is the dual of $C_b(\mathcal{Z})$; see Theorem 7. On the one hand, $\mathcal{B}_{\mathcal{M}(\mathcal{Z})}$ is compact in the weak-* topology of $\text{rba}(\mathcal{Z})$ by Lemma 10.

On the other hand, we check that the functional is upper semicontinuous in this topology. The first term of the functional is continuous (thus, upper semicontinuous) in the weak-* topology of $\text{rba}(\mathcal{Z})$, because $-g \in C_b(\mathcal{Z})$ by assumption and $\text{rba}(\mathcal{Z}) = C_b^*(\mathcal{Z})$. We write the second term as

$$\begin{aligned}
&-\frac{1}{\beta} \log\left(\int_{\mathcal{Y}} \exp\left(-\beta \int_{\mathcal{Z}} \varphi(y,z)\, d\gamma(z)\right) d\tau_{\mathcal{Y}}(y)\right) \\
&= -\sup_{\nu\in\mathcal{P}(\mathcal{Y})} \left\{ -\int_{\mathcal{Y}}\int_{\mathcal{Z}} \varphi(y,z)\, d\gamma(z)\, d\nu(y) - \beta^{-1} D_{KL}(\nu||\tau_{\mathcal{Y}}) \right\} \qquad (62) \\
&= \inf_{\nu\in\mathcal{P}(\mathcal{Y})} \left\{ \int_{\mathcal{Z}}\int_{\mathcal{Y}} \varphi(y,z)\, d\nu(y)\, d\gamma(z) + \beta^{-1} D_{KL}(\nu||\tau_{\mathcal{Y}}) \right\},
\end{aligned}$$

where the first equality follows from the argument in Lemma E and in the second equality we used Fubini's theorem. Remark that for a fixed $\nu \in \mathcal{M}_{\xi_1}(\mathcal{Y})$, $\int_{\mathcal{Y}} \varphi(y,\cdot)\, d\nu(y) \in C_b(\mathcal{Z})$ because of equation (59). Hence, the mapping $\gamma \mapsto \int_{\mathcal{Z}}\int_{\mathcal{Y}} \varphi(y,z)\, d\nu(y)\, d\gamma(z) + \beta^{-1} D_{KL}(\nu||\tau_{\mathcal{Y}})$ is continuous (thus, upper semicontinuous) in the weak-* topology of $\text{rba}(\mathcal{Z})$. The pointwise infimum of upper semicontinuous functions is upper semicontinuous, and thus (62) is upper semicontinuous as well. □

Lemma 10. *The unit TV norm ball of $\mathcal{B}_{\mathcal{M}(\mathcal{Z})}$, seen as a subset of $\text{rba}(\mathcal{Z})$, is compact in the weak-* topology of $\text{rba}(\mathcal{Z})$.*

Proof. If we endow $\mathcal{M}(\mathcal{Z})$ with the weak-* topology given by its predual $C_0(\mathcal{Z})$ (Theorem 6), the Banach-Alaoglu theorem (Theorem 8) states that $\mathcal{B}_{\mathcal{M}(\mathcal{Z})}$ is compact. Since the weak-* topology is Hausdorff, and Hausdorff compact spaces are closed, we have that $\mathcal{B}_{\mathcal{M}(\mathcal{Z})}$ is closed in the weak-* topology of $\mathcal{M}(\mathcal{Z})$. To show that $\mathcal{B}_{\mathcal{M}(\mathcal{Z})}$ is also closed in weak-* topology of $\text{rba}(\mathcal{Z})$, suppose that $\gamma \in \text{rba}(\mathcal{Z})$ is such that $(\gamma_n)_n \to \gamma$ in weak-* topology of $\text{rba}(\mathcal{Z})$ for some sequence $(\gamma_n)_n \subseteq$

$\mathcal{M}(\mathcal{Z})$. Then, since $C_0(\mathcal{Z}) \subseteq C_b(\mathcal{Z})$, $(\gamma_n)_n \to \gamma$ in weak-* topology of $\mathcal{M}(\mathcal{Z})$, and the closedness of $\mathcal{M}(\mathcal{Z})$ implies that $\gamma \in \mathcal{M}(\mathcal{Z})$.

We have that the TV norm closed unit ball $\mathcal{B}_{\text{rba}(\mathcal{Z})}$ of $\text{rba}(\mathcal{Z})$, which includes $\mathcal{B}_{\mathcal{M}(\mathcal{Z})}$, is compact in the weak-* topology again by the Banach-Alaoglu theorem. Since $\mathcal{B}_{\mathcal{M}(\mathcal{Z})}$ is a closed subset of the compact space $\mathcal{B}_{\text{rba}(\mathcal{Z})}$, it is itself compact in the weak-* topology of $\mathcal{B}_{\text{rba}(\mathcal{Z})}$. □

Lemma 11. *The function $\tilde{G}$ defined in* (61) *is such that $\tilde{G}(\gamma) = \sup_{\psi \in C_b(\mathcal{Z})} \{\langle \gamma, \psi \rangle - G(\psi)\}$, where G is defined in* (60)*.*

Proof. By definition $\tilde{G}(\gamma)$ is

$$\begin{aligned} &\sup_{\psi \in C_b(\mathcal{Z})} \left\{ \int_{\mathcal{Z}} \psi(z)\, d\gamma(z) - \sup_{\substack{\gamma' \in \mathcal{M}(\mathcal{Z}) \\ \|\gamma'\|_{\text{TV}} \leq 1}} \int_{\mathcal{Z}} (\psi(z) - g(z))\, d\gamma'(z) \right\} \\ &= \sup_{\psi \in C_b(\mathcal{Z})} \inf_{\substack{\gamma' \in \mathcal{M}(\mathcal{Z}) \\ \|\gamma'\|_{\text{TV}} \leq 1}} \left\{ \int_{\mathcal{Z}} \psi(z)\, d(\gamma - \gamma')(z) + \int_{\mathcal{Z}} g(z)\, d\gamma'(z) \right\} \end{aligned} \tag{63}$$

We want to apply Theorem 5 to flip the supremum and the infimum. We set B as in the proof of Lemma 7. The set A requires a careful construction. $\mathcal{M}(\mathcal{Z})$, which is the space finite countably additive regular Borel measures, is included in the Banach space of finite finitely additive regular Borel measures $\text{rba}(\mathcal{Z})$ endowed with the total variation norm, which by Theorem 7 is the continuous dual of $C_b(\mathcal{Z})$. $\text{rba}(\mathcal{Z})$ can be endowed with the weak-* topology of $\text{rba}(\mathcal{Z})$, which is the weakest one that makes maps of the form $\gamma \mapsto \int_{\mathcal{Z}} f(z)\, d\gamma(z)$ continuous for any $f \in C_b(\mathcal{Z})$. We set $A = \mathcal{B}_{\mathcal{M}(\mathcal{Z})}$ to be the TV norm ball of $\mathcal{M}(\mathcal{Z})$, as a subset of $\text{rba}(\mathcal{Z})$ endowed with its the weak-* topology of $\text{rba}(\mathcal{Z})$. Notice that $\gamma' \mapsto \int_{\mathcal{Z}} (g(z) - \psi(z))\, d\gamma'(z)$ is continuous in the weak-* topology of $\text{rba}(\mathcal{Z})$ because $g - \psi \in C_b(\mathcal{Z})$.

It only remains to show that $\mathcal{B}_{\mathcal{M}(\mathcal{Z})}$ is compact in the weak-* topology of $\text{rba}(\mathcal{Z})$. Thus, Theorem 5 can be applied, which means that the right-hand side of (63) is equal to

$$\begin{aligned} &\inf_{\substack{\gamma' \in \mathcal{M}(\mathcal{Z}) \\ \|\gamma'\|_{\text{TV}} \leq 1}} \sup_{\psi \in C_b(\mathcal{Z})} \left\{ \int_{\mathcal{Z}} \psi(z)\, d(\gamma - \gamma')(z) + \int_{\mathcal{Z}} g(z)\, d\gamma'(z) \right\} \\ &= \begin{cases} \int_{\mathcal{Z}} g(z)\, d\gamma(z) & \text{if } \|\gamma\|_{\text{TV}} \leq 1 \\ +\infty & \text{otherwise} \end{cases}. \end{aligned}$$

□

Proposition 4. *Problems* (15) *and* (17) *are equivalent in the following sense: if $\nu_1^\star$ is a solution of* (15) *for β, then it is also a solution of* (17) *for*

$$\tilde{\beta} = \beta \left(4 \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \varphi(y, z)\, d\nu_1^\star(y) - g(z) \right)^2 d\tau_{\mathcal{Z}}(z) \right)^{-1/2}$$

provided that the left-most factor is non-zero. Conversely, if $\nu_2^\star$ is a solution of (17) *for $\tilde{\beta}$, then it is also a solution of* (15) *for*

$$\beta = \tilde{\beta} \left(4 \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \varphi(y, z)\, d\nu_2^\star(y) - g(z) \right)^2 d\tau_{\mathcal{Z}}(z) \right)^{1/2}.$$

Proof. The Euler-Lagrange condition for (15) is

$$0 = \beta^{-1} \log\left(\frac{d\nu_1^\star}{d\tau_\mathcal{Y}}(y)\right) + \frac{\int_\mathcal{Z} \left(\int_\mathcal{Y} \varphi(y', z)\, d\nu_1^\star(y') - g(z)\right) \varphi(y, z)\, d\tau_\mathcal{Z}(z)}{\left(\int_\mathcal{Z} \left(\int_\mathcal{Y} \varphi(y', z)\, d\nu_1^\star(y') - g(z)\right)^2 d\tau_\mathcal{Z}(z)\right)^{1/2}} + K, \quad \forall y \in \mathcal{Y}$$

for some K. Thus,

$$\frac{d\nu_1^\star}{d\tau_\mathcal{Y}}(y) = \frac{1}{Z} \exp\left(-\beta \frac{\int_\mathcal{Z} \left(\int_\mathcal{Y} \varphi(y', z)\, d\nu_1^\star(y') - g(z)\right) \varphi(y, z)\, d\tau_\mathcal{Z}(z)}{\left(\int_\mathcal{Z} \left(\int_\mathcal{Y} \varphi(y', z)\, d\nu_1^\star(y') - g(z)\right)^2 d\tau_\mathcal{Z}(z)\right)^{1/2}}\right), \quad \forall y \in \mathcal{Y}. \tag{64}$$

The Euler-Lagrange condition for (17) is

$$0 = \tilde{\beta}^{-1} \log\left(\frac{d\nu_2^\star}{d\tau_\mathcal{Y}}(y)\right) + 2 \int_\mathcal{Z} \left(\int_\mathcal{Y} \varphi(y', z)\, d\nu_2^\star(y') - g(z)\right) \varphi(y, z)\, d\tau_\mathcal{Z}(z) + K, \quad \forall y \in \mathcal{Y} \tag{65}$$

for some K. Hence,

$$\frac{d\nu_2^\star}{d\tau_\mathcal{Y}}(y) = \frac{1}{Z} \exp\left(-2\tilde{\beta} \int_\mathcal{Z} \left(\int_\mathcal{Y} \varphi(y', z)\, d\nu_2^\star(y') - g(z)\right) \varphi(y, z)\, d\tau_\mathcal{Z}(z)\right).$$

Comparing (64) with (64), we see that $\nu_1^\star$ is equal to $\nu_2^\star$ when $\tilde{\beta}$ is set such that

$$2\tilde{\beta} = \frac{\beta}{\left(\int_\mathcal{Z} \left(\int_\mathcal{Y} \varphi(y', z)\, d\nu_1^\star(y') - g(z)\right)^2 d\tau_\mathcal{Z}(z)\right)^{1/2}}$$

Conversely, the solution $\nu_2^\star$ for a certain $\tilde{\beta}$ is equal to $\nu_1^\star$ when β is set such that

$$2\tilde{\beta} = \frac{\beta}{\left(\int_\mathcal{Z} \left(\int_\mathcal{Y} \varphi(y', z)\, d\nu_2^\star(y') - g(z)\right)^2 d\tau_\mathcal{Z}(z)\right)^{1/2}}.$$

□

Proposition 5. *Suppose $g : \mathcal{Z} \to \mathbb{R}$ is of the form $g(z) = \int_\mathcal{Y} \varphi(y, z)\, d\nu_p(y)$ for some $\nu_p \in \mathcal{P}(\mathcal{Z})$, and assume that the (negated) log-density $E(y) = -\log(\frac{d\nu_p}{d\tau_\mathcal{Y}}(y))$ belongs to the RKHS ball $B_{\mathcal{F}_2}(\beta_0)$.*

(a) On the one hand, when $\beta \geq \beta_0$ the solution $\nu_1^\star$ of (15) *is equal to ν_p. That is, there is recovery of the planted target measure and consequently $\int_\mathcal{Z} \left(\int_\mathcal{Y} \varphi(y, z)^2\, d\nu_1^\star(y) - g(z)\right)^2 d\tau_\mathcal{Z}(z) = 0$.*

(b) On the other hand, for all choices of $\tilde{\beta}$ finite if $\nu_2^\star$ is the solution of (17)*, the unregularized regression loss at $\nu_2^\star$ is not zero: $\int_\mathcal{Z} \left(\int_\mathcal{Y} \varphi(y, z)^2\, d\nu_2^\star(y) - g(z)\right)^2 d\tau_\mathcal{Z}(z) > 0$. Hence, $\nu_2^\star \neq \nu_p$ and there is no recovery.*

Proof. To prove (a), we use duality. Strong duality holds between (15) and (16) and moreover by Theorem 2 the respective solutions $\nu_1^\star$ and $h^\star$ of the two problems are linked by:

$$\frac{d\nu_1^\star}{d\tau_\mathcal{Y}}(y) = \frac{1}{Z_{\nu_1^\star}} \exp\left(-\beta \int_\mathcal{Z} \varphi(y, z)\, h^\star(z) d\tau_\mathcal{Z}(z)\right). \tag{66}$$

Remark that an arbitrary element f of the RKHS $\mathcal{F}_2$ admits a representation as

$$f(y) = \int_\Theta \varphi(y, z) h(z)\, d\tau_\mathcal{Z}(z), \quad \text{where } h \in L^2(\mathcal{Z}), \quad \text{and } \|f\|_{\mathcal{F}_2} = \|h\|_{L^2(\mathcal{Z})}. \tag{67}$$

For an arbitrary f, denote by ν_f the probability measure with density $\frac{d\nu_f}{d\tau_{\mathcal{Y}}}(y) = \exp(-f(y))/\int_{\mathcal{Y}} \exp(-f(y'))\, d\tau_{\mathcal{Y}}(y')$. Using (67) and $g(z) = \int_{\mathcal{Y}} \varphi(y,z)\, d\nu_p(y)$, we rewrite the problem (16) as

$$\begin{aligned}
&\operatorname*{arg\,min}_{\substack{h\in L^2(\mathcal{Z}) \\ \|h\|_{L^2}\leq 1}} \int_{\mathcal{Z}}\int_{\mathcal{Y}} \varphi(y,z)\, d\nu_p(y)\, h(z)\, d\tau_{\mathcal{Z}}(z) + \frac{1}{\beta}\log\left(\int_{\mathcal{Y}} \exp\left(-\beta\int_{\mathcal{Z}} \varphi(y,z)h(z)\, d\tau_{\mathcal{Z}}(z)\right) d\tau_{\mathcal{Y}}(y)\right) \\
&= \operatorname*{arg\,min}_{f\in\mathcal{B}_{\mathcal{F}_2}(\beta)} \int_{\mathcal{Y}} f(y)\, d\nu_p(y) + \log\left(\int_{\mathcal{Y}} e^{-f(y)} d\tau_{\mathcal{Y}}(y)\right) = \operatorname*{arg\,min}_{f\in\mathcal{B}_{\mathcal{F}_2}(\beta)} -\int_{\mathcal{Y}} \log\left(\frac{d\nu_f}{d\tau_{\mathcal{Y}}}(y)\right) d\nu_p(y) \\
&= \operatorname*{arg\,min}_{f\in\mathcal{B}_{\mathcal{F}_2}(\beta)} H(\nu_p,\nu_f) = \operatorname*{arg\,min}_{f\in\mathcal{B}_{\mathcal{F}_2}(\beta)} H(\nu_p,\nu_f) - H(\nu_p,\nu_p) = \operatorname*{arg\,min}_{f\in\mathcal{B}_{\mathcal{F}_2}(\beta)} D_{KL}(\nu_p||\nu_f).
\end{aligned}$$

In the first equality we have used Fubini's theorem to exchange the integrals in the first term, using the same reasoning as in (46)-(47). In the second equality we use the definition of ν_f. In the third equality, H denotes the cross-entropy, and in the fourth one, we use that $H(\nu_p,\nu_p)$ is finite because ν_p is absolutely continuous w.r.t. $\tau_{\mathcal{Y}}$. The fifth equality is by the definition of the KL divergence.

From this viewpoint, we have that the solution $f^\star = \arg\min_{f\in\mathcal{B}_{\mathcal{F}_2}(\beta)} D_{KL}(\nu_p||\nu_f)$ is linked to the solution $h^\star$ of (16): $f^\star(\cdot) = \beta\int_{\mathcal{Z}} \varphi(\cdot,z)\, h^\star(z) d\tau_{\mathcal{Z}}(z)$. Plugging this into (66), we obtain that

$$\frac{d\nu_1^\star}{d\tau}(y) = \frac{1}{Z_{\nu_1^\star}} \exp\left(-f^*(y)\right). \tag{68}$$

Since we have assumed that $E = -\log(\frac{d\nu_p}{d\tau_{\mathcal{Y}}}) \in \mathcal{B}_{\mathcal{F}_2}(\beta_0)$ with $\beta > \beta_0$, the unique solution of $\arg\min_{f\in\mathcal{B}_{\mathcal{F}_2}(\beta)} D_{KL}(\nu_p||\nu_f)$ is $f^\star = E$, which through (68) implies that $\nu_1^\star = \nu_E = \nu_p$.

To show (b), we use the Euler-Lagrange equation of (17), which is stated in (65). Since $\beta^{-1}\log\left(\frac{d\nu_2^\star}{d\tau_{\mathcal{Y}}}(y)\right) \neq 0$ for all $y\in\mathcal{Y}$, we must have that

$$\int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \varphi(y',z)\, d\nu_2^\star(y') - g(z)\right) \varphi(y,z)\, d\tau_{\mathcal{Z}}(z) \neq K, \tag{69}$$

does not hold uniformly over $y\in\mathcal{Y}$ for any constant K.

If we had $\int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \varphi(y',z)\, d\nu_2^\star(y') - g(z)\right)^2 d\tau_{\mathcal{Z}}(z) = 0$, that would mean that for all $z\in\mathcal{Z}$, $\int_{\mathcal{Y}} \varphi(y',z)\, d\nu_2^\star(y') - g(z) = 0$. This would imply that (69) is equal to zero for all $y\in\mathcal{Y}$, yielding a contradiction. □

F Proofs of Sec. 3 and additional results

Theorem 1. *Under Assumption 1, the problem* (3) *is the Fenchel dual of*

$$\min_{\nu\in\mathcal{P}(\mathcal{X})} \max_{\substack{\gamma\in\mathcal{M}(\Theta), \\ \|\gamma\|_{TV}\leq 1}} \beta^{-1} D_{KL}(\nu||\tau_{\mathcal{X}}) + \int_\Theta\int_{\mathcal{X}} \varphi(x,\theta)\, d(\nu-\nu_n)(x)\, d\gamma(\theta). \tag{5}$$

Moreover, the solution $\nu^\star$ of (70) *is precisely the Gibbs measure for the optimal $\gamma^\star$ in* (3)*, that is,* $\frac{d\nu^\star}{d\tau_{\mathcal{X}}}(x) = \frac{1}{Z_\beta}\exp\left(-\beta\int_\Theta \varphi(x,\theta)\, d\gamma^\star(x)\right)$.

Proof. The proof follows from applying Theorem 3 with $\mathcal{Y} = \mathcal{X}$ and $\mathcal{Z} = \Theta$. Assumption 2 holds because it is implied by Assumption 1 when one sets $\xi_1 = \xi$. Assumption 1 also implies that φ satisfies the assumption (i) in Theorem 3. Assumption (ii) in Theorem 3 is also fulfilled because

$g(\theta) = \frac{1}{n}\sum_{i=1}^n \varphi(x_i,\theta) \le \frac{1}{n}\sum_{i=1}^n \xi(x_i)$, which means that $g \in C_b(\Theta)$. By Theorem 3, we see that problem (3) is the Fenchel dual of

$$\min_{\nu\in\mathcal{P}(\mathcal{X})} \beta^{-1} D_{\mathrm{KL}}(\nu||\tau_{\mathcal{X}}) + \left\| \int_{\mathcal{X}} \varphi(x,\cdot)\, d(\nu-\nu_n)(x) \right\|_{L^\infty}. \tag{70}$$

and we also obtain the characterization for the measure $\nu^\star$. Since $\left\|\int_{\mathcal{X}} \varphi(x,\cdot)\, d(\nu-\nu_n)(x)\right\|_{L^\infty} = \sup_{\gamma\in\mathcal{M}(\Theta),\|\gamma\|_{\mathrm{TV}}\le 1} \int_\Theta \int_{\mathcal{X}} \varphi(x,\theta)\, d(\nu-\nu_n)(x)\, d\gamma(\theta) = \sup_{f\in\mathcal{B}_{\mathcal{F}_1}} \int_{\mathcal{X}} f(x)\, d(\nu-\nu_n)(x)$ using Fubini's theorem, we can rewrite (70) as

$$\min_{\nu\in\mathcal{P}(\mathcal{X})} \max_{\substack{\gamma\in\mathcal{M}(\Theta),\\ \|\gamma\|_{\mathrm{TV}}\le 1}} \beta^{-1} D_{\mathrm{KL}}(\nu||\tau_{\mathcal{X}}) + \int_\Theta \int_{\mathcal{X}} \varphi(x,\theta)\, d(\nu-\nu_n)(x)\, d\gamma(\theta).$$

□

Proposition 2. *Let $\{\theta_0^{(j)}\}_{j=1}^m$ be initial features sampled uniformly over Θ, let $\{\sigma_j\}_{j=1}^m$ be uniform samples over $\{\pm 1\}$ and let $\{w_0^{(j)} = 1\}_{j=1}^m$ be the initial weight values, which are set to 1. Let $\{X_0^{(i)}\}_{i=1}^N$ be the initial "generated" samples, which are chosen i.i.d. uniformly from the target sample set $\{x_i\}_{i=1}^n$. Consider the system of ODEs/SDEs:*

$$\begin{aligned} \frac{d\theta_t^{(j)}}{dt} &= \alpha\sigma_j \nabla \tilde{F}_t(\theta_t^{(j)}), \qquad \frac{dw_t^{(j)}}{dt} = \alpha w_t^{(j)}(\sigma_j \tilde{F}_t(\theta_t^{(j)}) - \tilde{K}_t) \\ dX_t^{(i)} &= \left(-\nabla \tilde{f}_t(X_t^{(i)}) + \beta^{-1}\nabla \log \frac{d\tau_{\mathcal{X}}}{d\lambda}(X_t^{(i)})\right) dt + \sqrt{2\beta^{-1}}\, dW_t^{(i)} \end{aligned} \tag{8}$$

where

$$\begin{aligned} \tilde{F}_t(\theta) = \frac{1}{N}\sum_{i=1}^N \varphi(X_t^{(i)},\theta) - \frac{1}{n}\sum_{i=1}^n \varphi(x_i,\theta), \qquad \tilde{f}_t(x) = \frac{1}{m}\sum_{j=1}^m \sigma_j w_t^{(j)} \varphi(x,\theta_t^{(j)}), \\ \tilde{K}_t = \mathbb{1}_{\sum_{j=1}^m w_t^{(j)} \ge m} \frac{1}{m}\sum_{j=1}^m \sigma_j w_t^{(j)} \tilde{F}_t(\theta_t^{(j)}). \end{aligned} \tag{9}$$

are the empirical counterparts of the functions in (7)*. Then the system* (8) *approximates the measure dynamics. Namely, as $m, N \to \infty$:*

- *the empirical measure $\hat{\gamma}_t = \frac{1}{m}\sum_{j=1}^m \sigma_j w_t^{(j)} \delta_{\theta_t^{(j)}}$ converges weakly to the solution $\gamma_t = \gamma_t^+ - \gamma_t^-$ of* (6) *with uniform initialization for any finite time interval $[0,T]$, and*
- *the empirical measure $\hat{\nu}_t = \frac{1}{N}\sum_{i=1}^N \delta_{X_t^{(i)}}$ converges weakly to the solution ν_t of* (6) *for any finite time interval $[0,T]$.*

Proof. For $\sigma = \pm 1$, define the empirical measures $\hat{\gamma}_t^\sigma = \frac{1}{m}\sum_{j=1}^m \mathbb{1}_{\sigma_j=\sigma} w_t^{(j)} \delta_{\theta_t^{(j)}}$, Given a test function χ on Θ, we have that

$$\begin{aligned} &\frac{d}{dt}\int_\Theta \chi(\theta)\, d\hat{\gamma}_t^\sigma(\theta) = \frac{d}{dt}\left(\frac{1}{m}\sum_{j=1}^m \mathbb{1}_{\sigma_j=\sigma} w_t^{(j)} f(\theta_t^{(j)})\right) \\ &= \frac{1}{m}\sum_{j=1}^m \mathbb{1}_{\sigma_j=\sigma} \frac{dw_t^{(j)}}{dt}\chi(\theta_t^{(j)}) + \mathbb{1}_{\sigma_j=\sigma} w_t^{(j)} \frac{d}{dt}\chi(\theta_t^{(j)}) \\ &= \frac{\alpha}{m}\sum_{j=1}^m \mathbb{1}_{\sigma_j=\sigma} w_t^{(j)}(\sigma_j \tilde{F}_t(\theta_t^{(j)}) - \tilde{K}_t)\chi(\theta_t^{(j)}) + \mathbb{1}_{\sigma_j=\sigma} w_t^{(j)} \nabla\chi(\theta_t^{(j)}) \cdot \sigma_j \nabla \tilde{F}_t(\theta_t^{(j)}) \\ &= \alpha \int_\Theta \left((\sigma\tilde{F}_t(\theta) - \tilde{K}_t)\chi(\theta) + \sigma\nabla\tilde{F}_t(\theta)\cdot\nabla\chi(\theta)\right) d\hat{\gamma}_t^\sigma(\theta) \end{aligned}$$

This is the weak formulation of the first equation in (6). We also observe that the forward Kolmogorov equation of the third equation in (8) is the Fokker-Planck equation in the second line of (6). The propagation of chaos argument that allows us to establish convergence $\hat{\gamma}_t \to \gamma_t$ and $\hat{\nu}_t \to \hat{\nu}$ is classical (Sznitman, 1991) and can be found for a very similar coupled setting in Domingo-Enrich et al. (2020). □

F.1 Link of dual $\mathcal{F}_1$-EBMs training with learned MMD training

We show that training dual $\mathcal{F}_1$-EBMs is equivalent to learning a certain form of MMD with feature learning. This observation provides a clearer link between dual $\mathcal{F}_1$-EBMs and dual $\mathcal{F}_2$-EBMs, in which the kernel is fixed (equation (1)). Feature-learning MMD has been the subject of several works and has been shown to outperform fixed-kernel MMD (Li et al., 2017). In particular, we have the following:

Proposition 6. *The solutions or saddle points of* (5) *are the saddle points of*

$$\min_{\nu \in \mathcal{P}(\mathcal{X})} \max_{\gamma \in \mathcal{P}(\Theta)} \beta^{-1} D_{KL}(\nu || \tau_{\mathcal{X}}) + MMD_{k_\gamma}(\nu, \nu_n),$$

where $MMD_{k_\gamma}(\nu, \nu_n) = (\int_{\mathcal{X}\times\mathcal{X}} k_\gamma(x, x')\, d(\nu - \nu_n)(x)\, d(\nu - \nu_n)(x'))^{1/2}$ *and the kernel* $k_\gamma = \int_\Theta \varphi(x, \theta)\varphi(x', \theta)\, d\gamma(\theta)$ *is well defined for any* $\gamma \in \mathcal{P}(\Theta)$.

Proof. The second term in the objective of (5) is $\int_\Theta \int_{\mathcal{X}} \varphi(x, \theta)\, d(\nu - \nu_n)(x)\, d\gamma(\theta)$. For any $\gamma \in \mathcal{M}(\Theta)$, $\|\gamma\|_{\text{TV}} \leq 1$, we apply the Cauchy-Schwarz inequality and then Fubini's theorem:

$$\int_\Theta \int_{\mathcal{X}} \varphi(x, \theta)\, d(\nu - \nu_n)(x)\, d\gamma(\theta) \leq \left(\int_\Theta \left(\int_{\mathcal{X}} \varphi(x, \theta)\, d(\nu - \nu_n)(x) \right)^2 d|\gamma|(\theta) \right)^{1/2} \tag{71}$$

$$= \left(\int_{\mathcal{X}\times\mathcal{X}} \int_\Theta \varphi(x, \theta)\varphi(x', \theta)\, d|\gamma|(\theta)\, d(\nu - \nu_n)(x)\, d(\nu - \nu_n)(x') \right)^{1/2}$$

$$= \left(\int_{\mathcal{X}\times\mathcal{X}} k_\gamma(x, x')\, d(\nu - \nu_n)(x)\, d(\nu - \nu_n)(x') \right)^{1/2} = MMD_{k_\gamma}(\nu - \nu_n).$$

For any $\nu \in \mathcal{P}(\Theta)$, notice that for all measures

$$\gamma^\star \in \operatorname*{arg\,max}_{\gamma \in \mathcal{M}(\Theta), \|\gamma\|_{\text{TV}} \leq 1} \int_\Theta \int_{\mathcal{X}} \varphi(x, \theta)\, d(\nu - \nu_n)(x)\, d\gamma(\theta) \tag{72}$$

and all measures

$$\gamma^\star \in \operatorname*{arg\,max}_{\gamma \in \mathcal{P}(\mathcal{X})} \left(\int_\Theta \left(\int_{\mathcal{X}} \varphi(x, \theta)\, d(\nu - \nu_n)(x) \right)^2 d|\gamma|(\theta) \right)^{1/2},$$

we must have

$$\text{supp}(\gamma^\star) \subseteq \left\{ \theta' \in \Theta \;\middle|\; \left| \int_{\mathcal{X}} \varphi(x, \theta')\, d(\nu - \nu_n)(x) \right| = \max_{\theta \in \Theta} \left| \int_{\mathcal{X}} \varphi(x, \theta)\, d(\nu - \nu_n)(x) \right| \right\}.$$

Hence, for any measure $\gamma^\star$ fulfilling (72),

$$\int_\Theta \int_{\mathcal{X}} \varphi(x, \theta)\, d(\nu - \nu_n)(x)\, d\gamma^\star(\theta) = \max_{\theta \in \Theta} \left| \int_{\mathcal{X}} \varphi(x, \theta)\, d(\nu - \nu_n)(x) \right|$$

$$= \left(\int_\Theta \left(\int_{\mathcal{X}} \varphi(x, \theta)\, d(\nu - \nu_n)(x) \right)^2 d|\gamma^\star|(\theta) \right)^{1/2},$$

which shows that at maximizers, all the terms of (71) are equal, concluding the proof. □

G LINKS WITH SCORE MATCHING

Proposition 1. *Suppose that $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ is a manifold without boundaries. Assume that $\int_{\mathcal{X}} |\nabla_x \varphi(x,\theta) \cdot \nabla \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x)| \, d\tau_{\mathcal{X}}(x)$ is upper-bounded by some constant K for all $\theta \in \Theta$. Assume also that $\sup_{\theta\in\Theta} \|\nabla_x \varphi(x,\theta)\| < \eta(x)$ and that $\int_{\mathcal{X}} |\eta(x)|^2 \, d\nu_p(x) < \infty$. The optimization problem to train EBMs under the score matching loss over the ball $\mathcal{B}_{\mathcal{F}_1}(1)$ gives $f_{\mathrm{SM}} = \int_\Omega \varphi(\cdot,\theta) d\gamma_{\mathrm{SM}}(\theta)$ where*

$$\gamma_{\mathrm{SM}} = \underset{\substack{\gamma\in\mathcal{M}(\Theta)\\ \|\gamma\|_{TV}\leq 1}}{\arg\min} \int_\Theta \int_{\mathcal{X}} \left(\frac{1}{2} \nabla_x \varphi(x,\theta) \cdot \nabla_x \int_\Theta \varphi(x,\theta') \, d\gamma(\theta') - \beta^{-1} \Delta_x \varphi(x,\theta) \right) d\nu_n(x) d\gamma(\theta) \tag{4}$$

Proof. The score matching metric between two absolutely continuous measures ν and ν_p is

$$\mathrm{SM}(\nu_p,\nu) = \int_{\mathcal{X}} \left\| \nabla \log \frac{d\nu}{d\tau_{\mathcal{X}}}(x) - \nabla \log \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x) \right\|^2 d\nu_p(x)$$

If constrain the density of ν to belong to the $\mathcal{F}_1$ ball of radius β, we can write $\log \frac{d\nu}{d\tau_{\mathcal{X}}}(x) = -\int_\Theta \varphi(x,\theta) \, d\gamma(\theta)$ for some $\gamma \in \mathcal{M}(\Theta)$ such that $\|\gamma\|_{\mathrm{TV}} \leq \beta$. Thus, the minimization problem of $\mathrm{SM}(\nu,\nu_p)$ over this class of energies can be written as

$$\min_{\substack{\gamma\in\mathcal{M}(\Theta)\\ \|\gamma\|_{\mathrm{TV}}\leq\beta}} \int_{\mathcal{X}} \left\| -\int_\Theta \nabla_x \varphi(x,\theta) \, d\gamma(\theta) - \nabla \log \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x) \right\|^2 d\nu_p(x)$$

Following Hyvärinen (2005), the objective functional can be expressed as

$$\int_{\mathcal{X}} \left(\left\| \int_\Theta \nabla_x \varphi(x,\theta) \, d\gamma(\theta) \right\|^2 + \left\| \nabla \log \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x) \right\|^2 + 2 \int_\Theta \nabla_x \varphi(x,\theta) \cdot \nabla \log \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x) \, d\gamma(\theta) \right) d\nu_p(x).$$

The middle term is constant w.r.t. γ, hence it is irrelevant. We use Fubini's theorem in the third term

$$\begin{aligned}
&\int_{\mathcal{X}} \int_\Theta \nabla_x \varphi(x,\theta) \cdot \nabla \log \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x) \, d\gamma(\theta) \, d\nu_p(x) = \int_\Theta \int_{\mathcal{X}} \nabla_x \varphi(x,\theta) \cdot \nabla \log \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x) \, d\nu_p(x) \, d\gamma(\theta) \\
&= \int_\Theta \int_{\mathcal{X}} \nabla_x \varphi(x,\theta) \cdot \nabla \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x) \, d\tau_{\mathcal{X}}(x) \, d\gamma(\theta) = -\int_\Theta \int_{\mathcal{X}} \Delta \varphi(x,\theta) \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x) \, d\tau_{\mathcal{X}}(x) \, d\gamma(\theta) \\
&= -\int_\Theta \int_{\mathcal{X}} \Delta \varphi(x,\theta) \, d\nu_p(x) \, d\gamma(\theta).
\end{aligned} \tag{73}$$

In the fourth equality of (73) we applied integration by parts. Fubini's theorem can be applied in the first equality because

$$\begin{aligned}
\int_\Theta \int_{\mathcal{X}} |\nabla_x \varphi(x,\theta) \cdot \nabla \log \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x)| \, d\nu_p(x) \, d|\gamma|(\theta) &= \int_\Theta \int_{\mathcal{X}} |\nabla_x \varphi(x,\theta) \cdot \nabla \frac{d\nu_p}{d\tau_{\mathcal{X}}}(x)| \, d\tau_{\mathcal{X}}(x) d|\gamma|(\theta) \\
&\leq K \|\gamma\|_{\mathrm{TV}} < +\infty
\end{aligned}$$

We also use similar arguments for the first term:

$$\begin{aligned}
&\int_{\mathcal{X}} \left\| \int_\Theta \nabla_x \varphi(x,\theta) \, d\gamma(\theta) \right\|^2 d\nu_p(x) = \int_{\mathcal{X}} \int_\Theta \int_\Theta \nabla_x \varphi(x,\theta) \cdot \nabla_x \varphi(x,\theta') \, d\gamma(\theta) \, d\gamma(\theta') \, d\nu_p(x) \\
&= \int_\Theta \int_{\mathcal{X}} \nabla_x \varphi(x,\theta) \cdot \int_\Theta \nabla_x \varphi(x,\theta') \, d\gamma(\theta') \, d\nu_p(x) \, d\gamma(\theta)
\end{aligned}$$

In this equation we can apply Fubini's theorem because

$$\begin{aligned}
&\int_\Theta \int_{\mathcal{X}} \left| \int_\Theta \nabla_x \varphi(x,\theta) \cdot \nabla_x \varphi(x,\theta')\, d\gamma(\theta) \right| \, d\nu_p(x)\, d|\gamma|(\theta') \\
&\leq \int_\Theta \int_{\mathcal{X}} \int_\Theta \|\nabla_x \varphi(x,\theta)\| \|\nabla_x \varphi(x,\theta')\|\, d|\gamma|(\theta)\, d\nu_p(x)\, d|\gamma|(\theta') \\
&\leq \|\gamma\|_{\mathrm{TV}} \int_{\mathcal{X}} \int_\Theta \eta(x)^2\, d\nu_p(x)\, d|\gamma|(\theta') < +\infty,
\end{aligned}$$

by the assumption that $\int_{\mathcal{X}} \eta(x)^2\, d\nu_p(x) < +\infty$. The proof is concluded by exchanging ν_p by its empirical version ν_n. □

Proposition 3. *In the limit $\alpha \to \infty$, the equations for γ_t^σ in* (10) *reduce to*

$$\partial_t \gamma_t^\sigma = \sigma \nabla_\theta \cdot \left(\gamma_t^\sigma \nabla_\theta V(\gamma_t)(\theta) \right) - \gamma_t^\sigma \left(\sigma V(\gamma_t)(\theta) - \bar{V}(\gamma_t) \right), \qquad \sigma = \pm 1, \;\; \gamma_t^\sigma = \gamma_t^\pm \quad (11)$$

where $\gamma_t = \gamma_t^+ - \gamma_t^-$, $\bar{V}(\gamma) = \int_\Theta V(\gamma) d\gamma$, and $V(\gamma)(\theta)$ is the Frechet derivative of the score matching loss $L : \mathcal{M}(\Theta) \to \mathbb{R}$ defined in (4)*.*

Proof. Let us start from the dynamics (10). For a domain $\mathcal{X}$ without boundary, Duhamel's principle states that the solution $u(x,t)$ of

$$\begin{cases} \partial_t u(x,t) - Lu(x,t) = f(x,t) \\ u(x,0) = 0 \end{cases}$$

is equal to $u(x,t) = \int_0^t P_s f(x,t)\, ds$, where $P_s f$ is the solution of

$$\begin{cases} \partial_t u(x,t) - Lu(x,t) = 0 \\ u(x,s) = f(x,s) \end{cases}$$

While it is typically stated for classical PDEs, in our case we consider Duhamel's principle in the weak sense, i.e. the equalities hold when integrated with respect to test functions.

We can apply Duhamel's principle for the second equation of (10), with $u(\cdot,t) = \nu_t - \nu_0$, $Lu = -\alpha u$ and $f(x,t) = \nabla_x \cdot \left(\nu_t \nabla_x \int_\Omega \tilde{\varphi}(x,\omega)\, d\mu_t(\omega) \right) + \beta^{-1} \Delta_x \nu_t + \alpha(\nu_n - \nu_0)$. Notice that the solution $P_s f$ of

$$\begin{cases} \partial_t u(x,t) + \alpha u(x,t) = 0 \\ u(x,s) = f(x,s) \end{cases}$$

is $P_s f(x,t) = f(x,s) e^{-\alpha(t-s)}$. By Duhamel's principle we obtain that

$$\begin{aligned}
\nu_t - \nu_0 &= \int_0^t P_s f(x,t)\, ds \\
&= \int_0^t \left(\nabla_x \cdot \left(\nu_s \nabla_x \int_\Theta \varphi(x,\theta)\, d\gamma_s(\theta) \right) + \beta^{-1} \Delta_x \nu_s + \alpha(\nu_n - \nu_0) \right) e^{-\alpha(t-s)}\, ds.
\end{aligned}$$

Since $\alpha \int_0^t e^{-\alpha(t-s)}\, ds = 1 - e^{-\alpha t}$, this is equivalent to

$$\nu_t = \nu_0 e^{-\alpha t} + \nu_n (1 - e^{-\alpha t}) + \int_0^t e^{-\alpha(t-s)} \left(\nabla_x \cdot \left(\nu_s \nabla_x \int_\Omega \varphi(x,\omega)\, d\gamma_s(\omega) \right) + \beta^{-1} \Delta \nu_s \right) ds \quad (74)$$

From (74), we see that as $\alpha \to +\infty$,

$$\alpha(\nu_t - \nu_n) \to \nabla_x \cdot \left(\nu_t \nabla_x \int_\Omega \varphi(x,\omega)\, d\gamma_t(\omega) \right) + \beta^{-1} \Delta \nu_t, \quad (75)$$

or alternatively, for any test function f,

$$\alpha \int_{\mathcal{X}} f(x) d(\nu_t - \nu_n)(x) \to -\int_{\mathcal{X}} \nabla f(x) \cdot \nabla_x \int_{\Theta} \varphi(x, \theta)\, d\gamma_t(\theta)\, d\nu_t(x) + \beta^{-1} \int_{\mathcal{X}} \Delta f(x)\, d\nu_t(x). \tag{76}$$

Moreover, (75) implies that $\alpha \to +\infty$, $\nu_t \to \nu_n$. Applying this into (76), we obtain that

$$\alpha \int_{\mathcal{X}} f(x) d(\nu_t - \nu_n)(x) \to -\int_{\mathcal{X}} \nabla f(x) \cdot \nabla_x \int_{\Theta} \varphi(x, \theta)\, d\mu_t(\theta)\, d\nu_n(x) + \beta^{-1} \int_{\mathcal{X}} \Delta f(x)\, d\nu_n(x).$$

Plugging this into the definition of F_t in (7), we get that

$$\alpha F_t(\theta) \to -\int_{\mathcal{X}} \nabla_x \varphi(x, \theta) \cdot \nabla_x \int_{\Theta} \varphi(x, \theta)\, d\mu_t(\theta)\, d\nu_n(x) + \beta^{-1} \int_{\mathcal{X}} \Delta f(x)\, d\nu_n(x).$$

Using this in the first equation of (10), we get that in the limit $\alpha \to +\infty$,

$$\begin{aligned}
&\partial_t \gamma_t^\sigma \\
&= \sigma \nabla_\theta \cdot \left(\gamma_t^\sigma \left(\nabla_\theta \int_{\mathcal{X}} \nabla_x \varphi(x, \theta) \cdot \int_{\Theta} \nabla_x \varphi(x, \theta')\, d\mu_t(\theta')\, d\nu_n(x) + \beta^{-1} \nabla_\theta \int_{\mathcal{X}} \Delta_x \varphi(x, \theta)\, d\nu_t(x) \right) \right) \\
&+ \mu_t \left(-\sigma \int_{\mathcal{X}} \nabla_x \varphi(x, \theta) \cdot \int_{\Theta} \nabla_x \varphi(x, \theta)\, d\gamma_t(\theta)\, d\nu_n(x) + \sigma \beta^{-1} \int_{\mathcal{X}} \Delta_x \varphi(x, \theta)\, d\nu_n(x) - \tilde{K}_t \right) \\
&= \frac{1}{2\beta^2} \left(\sigma \nabla_\theta \cdot \left(\gamma_t^\sigma \nabla_\theta V(\gamma_t)(\theta) \right) - \gamma_t^\sigma \left(\sigma V(\gamma_t)(\theta) - \bar{V}(\gamma_t) \right) \right)
\end{aligned}$$

which is (11) up to a time reparametrization. □

G.1 Direct optimization of the score matching loss

Let L be defined in Proposition 3. The first variation $\frac{\delta L}{\delta \mu}(\mu)(\omega)$ of L at μ is

$$\frac{\delta L}{\delta \gamma}(\gamma)(\theta) = \int_{\mathcal{X}} \left(2\beta^2 \nabla_x \varphi(x, \theta) \cdot \nabla_x \int_{\Theta} \varphi(x, \theta')\, d\gamma(\theta') - 2\beta \Delta_x \varphi(x, \theta) \right) d\nu_n(x).$$

We optimize (4) via the Wasserstein-Fisher-Rao (WFR) gradient flow (11). This measure PDE can be approximated via a particle system ODE (equation (12)), and the corresponding particle system may be discretized into Algorithm 3.

Lemma 12. *Let $\{x_i\}_{i=1}^n$ be samples from a target distribution ν_p. Let $\{\theta_0^{(j)}\}_{j=1}^m$ be features sampled uniformly over Θ, let $\{\sigma_j\}_{j=1}^m$ be uniform samples over $\{\pm 1\}$ and let $\{w_0^{(j)} = 1\}_{j=1}^m$ be the initial weight values, which are set to 1. Equation* (11) *can be simulated by evolving the features $\{\theta^{(j)}\}_{j=1}^m$ and the weights $\{w^{(j)}\}_{j=1}^m$ via the following ODE:*

$$\frac{d\theta_t^{(j)}}{dt} = -\sigma_j \nabla_\theta \left(\frac{1}{n} \sum_{i=1}^n \nabla_x \varphi(x_i, \theta_t^{(j)}) \frac{1}{m} \sum_{j'=1}^m \sigma_{(j')} w_t^{(j')} \nabla_x \varphi(x_i, \theta_t^{(j')}) - \frac{\beta^{-1}}{n} \sum_{i=1}^n \Delta_x \varphi(x_i, \theta_t^{(j)}) \right),$$

$$\frac{d \log w_t^{(j)}}{dt} = -\left(\frac{\sigma_j}{n} \sum_{i=1}^n \nabla_x \varphi(x_i, \theta_t^{(j)}) \frac{1}{m} \sum_{j'=1}^m \sigma_{j'} w_t^{(j')} \nabla_x \varphi(x_i, \theta_{j'}) - \frac{\sigma_j \beta^{-1}}{n} \sum_{i=1}^n \Delta_x \varphi(x_i, \theta_t^{(j)}) - K(t) \right),$$

where

$$\begin{aligned}
K(t) &= \mathbb{1}_{\|\gamma_t^+\|_{TV} + \|\gamma_t^-\|_{TV} \geq 1} \\
&\times \frac{1}{m} \sum_{j=1}^m \sigma_j w_j \left(\frac{1}{n} \sum_{i=1}^n \nabla_x \varphi(x_i, \theta_j) \cdot \frac{1}{m} \sum_{j'=1}^m \sigma_{j'} w_{j'} \nabla_x \varphi(x_i, \theta_{j'}) - \frac{\beta^{-1}}{n} \sum_{i=1}^n \Delta_x \varphi(x_i, \theta_j) \right).
\end{aligned}$$

Namely, up to a time reparametrization with factor $2\beta^2$, the time-dependent measure $\hat{\gamma}_t = \frac{1}{m}\sum_{j=1}^m \sigma_j w_t^{(j)} \delta_{\theta_t^{(j)}}$ converges weakly to the solution $\gamma_t = \gamma_t^+ - \gamma_t^-$ of (11) with uniform initialization, for any finite time interval $[0, T]$, as $m \to \infty$.

Proof. We check that $\hat{\gamma}_t$ is a weak solution of (11) as in Proposition 2, and use propagation of chaos. □

Algorithm 3 $\mathcal{F}_1$-EBM training via score matching

Input: n samples $\{x_i\}_{i=1}^n$ of the target distribution, stepsize s.
Initialize features $(\theta_0^{(j)})_{j=1}^m$ unif. over Θ, weights $(w_0^{(j)} = 1)_{j=1}^m$, signs $(\sigma_j)_{j=1}^m$ unif. over $\{\pm 1\}$.

Initialize generated samples $\{X_0^{(i)}\}_{i=1}^N$ uniformly i.i.d. from $\{x_i\}_{i=1}^n$.
for $t = 0, \ldots, T-1$ **do**
  for $j = 1, \ldots, m$ **do**
    Set $\theta_{t+1}^{(j)} = \theta_t^{(j)} - s\sigma_j \nabla_\theta \left(\frac{1}{n}\sum_{i=1}^n \nabla_x \varphi(x_i, \theta_t^{(j)}) \cdot \frac{1}{m}\sum_{j'=1}^m \sigma_{j'} w_t^{(j')} \nabla_x \varphi(x_i, \theta_{j'}) \right) + s\beta^{-1}\sigma_j \nabla_\theta \left(\frac{1}{n}\sum_{i=1}^n \Delta_x \varphi(x_i, \theta_j) \right)$.
    Update $\tilde{w}_{t+1}^{(j)} = w_{t+1}^{(j)} \exp(-\frac{s\sigma_j}{n}\sum_{i=1}^n \nabla_x \varphi(x_i, \theta_t^{(j)}) \cdot \frac{1}{m}\sum_{j'=1}^m \sigma_{j'} w_t^{(j')} \nabla_x \varphi(x_i, \theta_t^{(j')}) + \frac{s\beta^{-1}\sigma_j}{n}\sum_{i=1}^n \Delta_x \varphi(x_i, \theta_t^{(j)}))$
    Normalize if needed $w_{t+1}^{(j)} = \tilde{w}_{t+1}^{(j)} / \max(\frac{1}{m}\sum_{j'=1}^m \tilde{w}_{t+1}^{(j')}, 1)$.
  end for
end for
Energy $E_T(x) := \frac{\beta}{m}\sum_{j=1}^m w_j \sigma_j \varphi(x, \theta_j)$.

Proposition 7. *The Algorithm 3 is equivalent to Algorithm 1 with (i) base probability measure proportional to Lebesgue, i.e. $\nabla \log \frac{d\tau_{\mathcal{X}}}{d\lambda} = 0$, (ii) replacement probability $p_r = 1$ and (iii) noisy updates.*

Proof. For any iteration t and particle i, let $k_{t+1,i}$ be a uniform independent integer random variable over $\{1, \ldots, n\}$, i.e. $x_{k_{t+1,i}}$ is a uniform random sample from $\{x_{i'}\}_{i'=1}^n$. We may rewrite the updates on $\{\theta_{t+1}^{(j)}\}$, $\{w_{t+1}^{(j)}\}$ for Algorithm 1 with $p_r = 1$, $\beta^{-1} = 0$ as

$$
\begin{aligned}
X_{t+1}^{(i)} &= x_{k_{t+1,i}} - \frac{s}{m}\sum_{j=1}^m w_t^{(j)} \sigma_j \nabla_x \varphi(x_{k_{t+1,i}}, \theta^{(j)}) + \sqrt{2\beta^{-1}s}\,\zeta_t^{(i)}, \\
\theta_{t+1}^{(j)} &= \theta_t^{(j)} + s\alpha\sigma_j w_t^{(j)} \left(\frac{1}{N}\sum_{i=1}^N \nabla_\theta \varphi(X_{t+1}^{(i)}, \theta_t^{(j)}) - \frac{1}{n}\sum_{i=1}^n \nabla_\theta \varphi(x_i, \theta_t^{(j)}) \right), \\
\tilde{w}_{t+1}^{(j)} &= w_{t+1}^{(j)} \exp\left(\frac{s\alpha}{N}\sum_{i=1}^N \varphi(X_{t+1}^{(i)}, \theta_t^{(j)}) - \frac{s\alpha}{n}\sum_{i=1}^n \varphi(x_i, \theta_t^{(j)}) \right), \\
w_{t+1}^{(j)} &= \tilde{w}_{t+1}^{(j)} / \max\left(\frac{1}{m}\sum_{j'=1}^m \tilde{w}_{t+1}^{(j')}, 1 \right).
\end{aligned}
\tag{77}
$$

Notice that in the regime of small stepsize $s \ll 1$, we can use a second order Taylor approximation:

$$\begin{aligned}\nabla_\theta \varphi(X_{t+1}^{(i)}, \theta_t^{(j)}) &\approx \nabla_\theta \varphi(x_{k_{t+1,i}}, \theta_t^{(j)}) + \langle \nabla_{x,\theta}\varphi(x_{k_{t+1,i}}), X_{t+1}^{(i)} - x_{k_{t+1,i}} \rangle \\ &+ \beta^{-1} s \langle \zeta_t^{(i)}, \nabla_{x,x,\theta}\varphi(x_{k_{t+1,i}}, \theta_t^{(j)})\zeta_t^{(i)} \rangle + o(s) \\ &= \nabla_\theta \varphi(x_{k_{t+1,i}}, \theta_t^{(j)}) - 2s \nabla_{x,\theta}\varphi(x_{k_{t+1,i}}, \theta_t^{(j)}) \cdot \frac{1}{m}\sum_{j'=1}^m w^{(j')}\sigma_{j'}\nabla_x \varphi(x_{k_{t+1,i}}, \theta_t^{(j')}) \\ &+ \sqrt{2\beta^{-1}s}\langle \nabla_{x,\theta}\varphi(x_{k_{t+1,i}}), \zeta_t^{(i)} \rangle + \beta^{-1} s \langle \zeta_t^{(i)}, \nabla_{x,x,\theta}\varphi(x_{k_{t+1,i}}, \theta_t^{(j)})\zeta_t^{(i)} \rangle + o(s).\end{aligned}$$

Notice that that $\mathbb{E}[\langle \zeta_t^{(i)}, \nabla_{x,x,\theta}\varphi(x_{k_{t+1,i}}, \theta_t^{(j)})\zeta_t^{(i)} \rangle | \theta_t^{(j)}] = \frac{1}{n}\sum_{i=1}^n \nabla_\theta \nabla_{x,x}\varphi(x_i, \theta_t^{(j)})$. Moreover,

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \nabla_\theta \varphi(x_{k_{t+1,i}}, \theta_t^{(j)}) - 2s\nabla_{x,\theta}\varphi(x_{k_{t+1,i}}, \theta_t^{(j)}) \cdot \frac{1}{m}\sum_{j'=1}^m w^{(j')}\sigma_{j'}\nabla_x\varphi(x_{k_{t+1,i}}, \theta_t^{(j')}) \Big| \theta_t\right] \\ &= \frac{1}{n}\sum_{i=1}^n \nabla_\theta \varphi(x_i, \theta_t^{(j)}) - 2s\nabla_{x,\theta}\varphi(x_i, \theta_t^{(j)}) \cdot \frac{1}{m}\sum_{j'=1}^m w^{(j')}\sigma_{j'}\nabla_x\varphi(x_i, \theta_t^{(j')})\end{aligned}$$

Making use of these observations and the expression of the update on $\theta_{t+1}^{(j)}$ in (77), we get that

$$\begin{aligned}\mathbb{E}\left[\theta_{t+1}^{(j)} - \theta_t^{(j)} | \theta_t\right] &= -s^2\alpha\sigma_j w_t^{(j)} \left(\frac{1}{n}\sum_{i=1}^n \nabla_{x,\theta}\varphi(x_{k_{t+1,i}}, \theta_t^{(j)}) \cdot \frac{1}{m}\sum_{j=1}^m w_t^{(j')}\sigma_j \nabla_x \varphi(x_{k_{t+1,i}}, \theta_t^{(j)})\right) \\ &+ s^2\alpha\sigma_j w_t^{(j)} \frac{\beta^{-1}}{n}\sum_{i=1}^n \nabla_\theta \nabla_{x,x}\varphi(x_i, \theta_t^{(j)})\end{aligned}$$

And this is equal to the update in Algorithm 3 after renaming the stepsize $s^2\alpha \to s$. The analogous argument holds for the update on $\log \tilde{w}_{t+1}^{(j)}$. □

G.2 Comparison with Score-based Generative Models (SGMs)

A recent series of works Song & Ermon (2019; 2020); Song et al. (2021); Song & Kingma (2021); Kadkhodaie & Simoncelli (2020); Jolicoeur-Martineau et al. (2020); Dhariwal & Nichol (2021) have leveraged the link between score matching and reversing a diffusion process (ie, *denoising*) to propose flexible and powerful generative models (SGMs). While our work shows connections with score matching, our approach is somewhat far from SGMs. Indeed, SGMs proceed by estimating various score functions of noisy versions of the data distribution, rather than the original data distribution, and later use these estimates for obtaining new samples using a Langevin diffusion. In contrast, the score matching loss that we consider is directly given by the score matching metric through the classical trick introduced by Hyvärinen (2005), and our Langevin sampling process is built into the training dynamics. Also, while our work makes use of SDEs to evolve the generated samples, we do not use a forward-backward framework in the style of certain SGMs Song et al. (2021).

H Proofs of App. C

Lemma 13. *If Assumption 2 holds, the Fenchel dual of the problem* $\min_{\nu\in\mathcal{P}(\mathcal{X})} \beta^{-1}D_{KL}(\nu||\tau_{\mathcal{X}}) + MMD_k(\nu, \nu_n)$ *is the problem* $\max_{f\in\mathcal{B}_{\mathcal{F}_2}(\beta)} -\frac{1}{n}\sum_{i=1}^n f(x_i) - \log\left(\int_{\mathcal{X}} e^{-f(x)}d\tau_{\mathcal{X}}(x)\right)$.

Proof. We apply Theorem 2 to show that the problem

$$\min_{\nu \in \mathcal{P}(\mathcal{X})} \beta^{-1} D_{KL}(\nu||\tau_{\mathcal{X}}) + \left(\int_{\Theta} \left(\int_{\mathcal{X}} \varphi(x,\theta)\, d(\nu - \nu_n)(x) \right)^2 d\tau_{\Theta}(\theta) \right)^{1/2} \tag{78}$$

has dual problem (22). It remains only to show that the second term of (78) is equal to $\text{MMD}_k(\nu, \nu_n)$. To obtain this, observe that

$$\begin{aligned}
&\int_{\Theta} \left(\int_{\mathcal{X}} \varphi(x,\theta)\, d(\nu - \nu_n)(x) \right)^2 d\tau_{\Theta}(\theta) \\
&= \int_{\mathcal{X}\times\mathcal{X}} \int_{\Theta} \varphi(x,\theta)\varphi(x',\theta) d\tau_{\Theta}(\theta)\, d(\nu - \nu_n)(x)\, d(\nu - \nu_n)(x') \\
&= \int_{\mathcal{X}\times\mathcal{X}} k(x,x')\, d(\nu - \nu_n)(x)\, d(\nu - \nu_n)(x'),
\end{aligned}$$

The first equality holds by Fubini's theorem following an argument similar to equations (46)-(47). The second equality follows from the characterization (1) of the kernel k. □

Lemma 14. *The Wasserstein gradient flow for the objective functional of* (23) *is given by* (25)*.*

Proof. The proof is standard. If we denote $L(\nu) = \beta^{-1} D_{KL}(\nu||\tau_{\mathcal{X}}) + \text{MMD}_k(\nu, \nu_n)$, the first variation of L at any $\nu \in \mathcal{P}(\mathcal{X})$ is $\frac{\delta L}{\delta \nu}(\nu) : \mathcal{X} \to \mathbb{R}$ such that for all $\nu' \in \mathcal{P}(\mathcal{X})$, $\lim_{\epsilon \to 0} (L(\nu + \epsilon(\nu' - \nu)) - L(\nu))/\epsilon = \int_{\mathcal{X}} d(\nu' - \nu)(x)$. In this case, for any absolutely continuous $\nu \in \mathcal{P}(\mathcal{X})$,

$$\frac{\delta L}{\delta \nu}(\nu)(x) = \beta^{-1} \log \frac{d\nu}{d\lambda}(x) + \beta^{-1} - \beta^{-1} \log \frac{d\tau_{\mathcal{X}}}{d\lambda}(x) + \frac{\int_{\mathcal{X}} k(x,x')\, d(\nu_t - \nu_n)(x')}{\text{MMD}_k(\nu_t, \nu_n)} \tag{79}$$

and its gradient is

$$\nabla \frac{\delta L}{\delta \nu}(\nu)(x) = \beta^{-1} \frac{\nabla \frac{d\nu}{d\lambda}(x)}{\frac{d\nu}{d\lambda}(x)} - \beta^{-1} \log \frac{d\tau_{\mathcal{X}}}{d\lambda}(x) + \frac{\int_{\mathcal{X}} k(x,x')\, d(\nu_t - \nu_n)(x')}{\text{MMD}_k(\nu_t, \nu_n)}$$

It is well known (Santambrogio, 2017) that the Wasserstein gradient flow of a functional L is a solution of the measure PDE

$$\partial_t \nu_t = \nabla \cdot \left(\nu_t \nabla \frac{\delta L}{\delta \nu}(\nu_t)(x) \right).$$

□

Lemma 15. *If $\mathcal{X}$ is arc-connected, the unique stationary solution $\nu^\star$ of* (25) *is the unique minimizer of* (23)*. The stationary solution must satisfy* (29)*.*

Proof. We follow the same reasoning as Rotskoff & Vanden-Eijnden (2018); Mei et al. (2018), skipping some techical details. Denoting $L(\nu) = \beta^{-1} D_{KL}(\nu||\tau_{\mathcal{X}}) + \text{MMD}_k(\nu, \nu_n)$, all stationary solutions $\nu^\star$ of the Wasserstein gradient flow of L must satisfy

$$\nabla \frac{\delta L}{\delta \nu}(\nu^\star)(x) = 0, \quad \forall x \in \text{supp}(\nu^\star) \tag{80}$$

Because of the KL term, $\text{supp}(\nu^\star) = \mathcal{X}$. Since L is strictly convex because MMD is convex and D_{KL} is strictly convex, L has at most one minimizer, which is uniquely specified by the Euler-Lagrange condition

$$\frac{\delta L}{\delta \nu}(\nu^\star)(x) = K, \quad \forall x \in \mathcal{X}, \text{ for some } K. \tag{81}$$

When $\mathcal{X}$ is arc-connected, (80) implies (81).

To show that the solution must satisfy (29), we just develop (81) as in (79) and isolate. □

I ADDITIONAL EXPERIMENTS

In this section we provide further insight into the training of teacher-student models with two planted student neurons of negative weights.

Experiments on teacher-student models in $d = 2$. We analyze the case of a teacher with two neurons, both with the same negative weight $w_j^* = -10$, in $d = 2$ (i.e. on the sphere) and we train the student setting $\beta = 20$ such that the approximation errors. This low-dimensional example allows for a visual representation of the training dynamics (see videos `KLdual_1e3_points.mp4` and `KLdual_1e4_points.mp4` in the supplementary material). Figure 2 shows that the densities of the Gibbs distribution associated to the teacher and the student at the end of training are very concentrated in two separated regions on the sphere. This means that sampling this distribution by Langevin dynamics, which is required in the late stages of training in the primal formulation, would be challenging due to strong metastability. Our aim is to illustrate that our dual formulation avoids this metastability issue in the sampling.

For different values of p_R, and for $n = 10^3, 10^4$ training data points, Figure 4 shows the evolution of the KL-divergence and the score matching between the teacher and student models, and the TV-norm of the student measure, i.e. the $\mathcal{F}_1$ norm of the student energy. We use $N = 2 \cdot 10^3, 2 \cdot 10^4$ particles (resp.), $m = 64$ student neurons, and a testing set of $n^* = 10^4$ to compute the KL-divergence and score matching metric. In this setting we observe that $p_R = 0$ and $p_R = 1/60$ perform similarly, while score matching ($p_R = 1$) has a slower convergence and has larger terminal values for both the KL divergence and the score matching metric. As expected, the test metrics improve with more training data n, and we observe that the relative gap between the methods becomes smaller; score matching becomes more competitive.

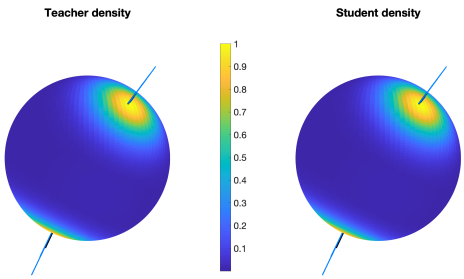

Figure 2: Comparison between the teacher and student density in $d = 2$, after training ($n = 10^4$, $N = 2 \cdot 10^4, m = 62, \alpha = 10, p_R = 0$). The location of the 2 teachers neurons are shown in black, and that of the 64 students neurons in blue.

Bimodality vs. monomodality in $d = 14$. Figure 5 shows the histograms for the cosines of the angles between the samples and each teacher neuron. We see that when the two teacher neurons are at an angle of 2.87 rad (almost opposite), the distribution is bimodal. When they are at an angle of 1.37 rad, the distribution is monomodal.

Comparing different values of p_R for $d = 14$. In Figure 6 (top, middle), which correspond to the bimodal case with angle 2.87 rad, we observe that the three variants have similar performance but $p_R = 1/40$ achieves the best metrics, followed very closely by $p_R = 0$ and $p_R = 1$ (score matching) a bit behind. Remark that early stopping might be beneficial in terms of the test error; the best test metrics are achieved roughly at the iteration at which the $\mathcal{F}_1$ norm of the trained energy reaches the $\mathcal{F}_1$ norm of the teacher energy. Interestingly, in the monomodal case with angle 1.37 rad (bottom of Figure 6), the best value for the KL divergence is achieved by $p_R = 1$ with early stopping, which beats the other two alternatives by a narrow margin. Unlike in the bimodal case, in the monomodal setting the training curves for the three methods display a change of behavior (a bump) slightly after initialization, and before the metrics reach values close to the final ones. This observation seems at odds with the common intuition that monomodal distributions are "easier" to deal with. To assess that the planted model defines a challenging high-dimensional density estimation problem, we consider a kernel density estimator baseline using an RBF kernel projected in the unit sphere. We report the KL divergences obtained in 6, and they are much higher than the EBM ones.

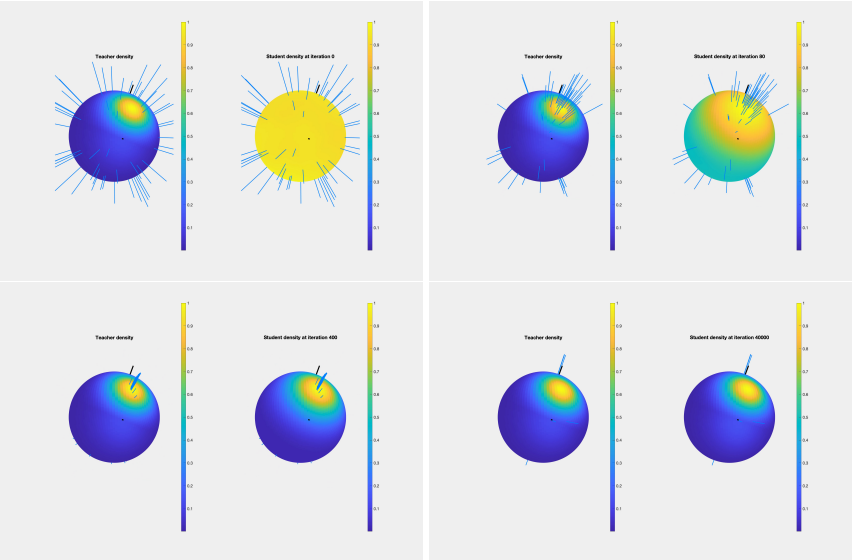

Figure 3: *Experiments in* $d = 2$*:* Selected frames of the video `KLdual_1e4_datapoints_monomodal.mp4` at iterations 0 (top left), 80 (top right), 400 (bottom left) and 40000 (bottom right). The parameters are $d = 2,\ m = 64,\ p_R = 0,\ n = 10^4,\ N = 2 \cdot 10^4,\ w_1^* = w_2^* = -10$. The teacher neurons, shown as black sticks, are almost perpendicular, and hence the teacher distribution is monomodal. The 64 student neurons are shown in blue. The two stages of training mentioned in text are clearly visible.

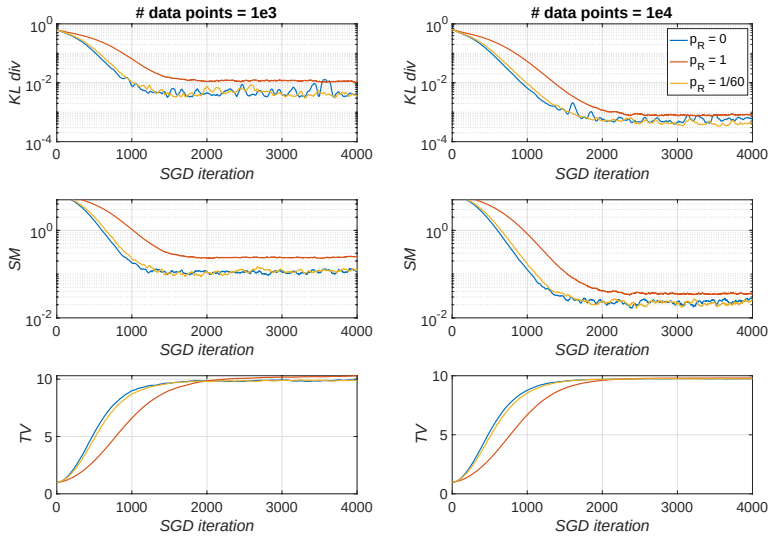

Figure 4: *Experiments in d=2:* The evolution of the KL divergence, the score matching metric and the TV norm of the trained measure (i.e., the $\mathcal{F}_1$ norm) during training for Algorithm 1 with $\mathcal{X} = \mathbb{S}^2$, $m = 64$, $p_R \in \{0, 1, 1/60\}$, $s = 0.02$, $\alpha = 2 + 10p_R$, and (left) $n = 10^3$, $N = 2 \cdot 10^3$, (right) $n = 10^4$, $N = 2 \cdot 10^4$. In comparison, the non-parametric kernel density estimator reaches a KL error of $2 \cdot 10^{-2}$ for $n = 10^3$ and $6 \cdot 10^{-3}$ for $n = 10^4$.

In the bottom row of Figure 6, we observe that when the teacher distribution is monomodal, which happens when the teacher neurons are close to perpendicular, the training curves present a "bumpy" shape unlike in the bimodal case. Figure 7 shows plots in the same setting, but with 10 times more training data points. As already observed in Figure 4 and Figure 6, taking larger n improves the

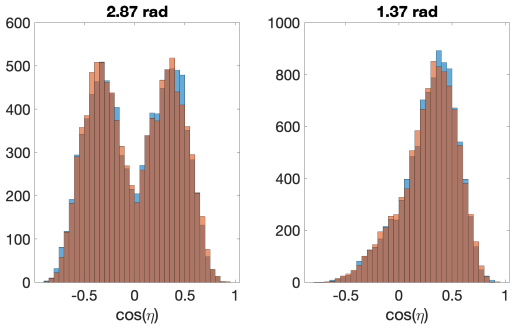

Figure 5: *Experiments in d=14:* Histograms for the cosines of the angles between each teacher neuron and samples from the target distribution, when the angle between teacher neurons is 2.87 and 1.37 rad.

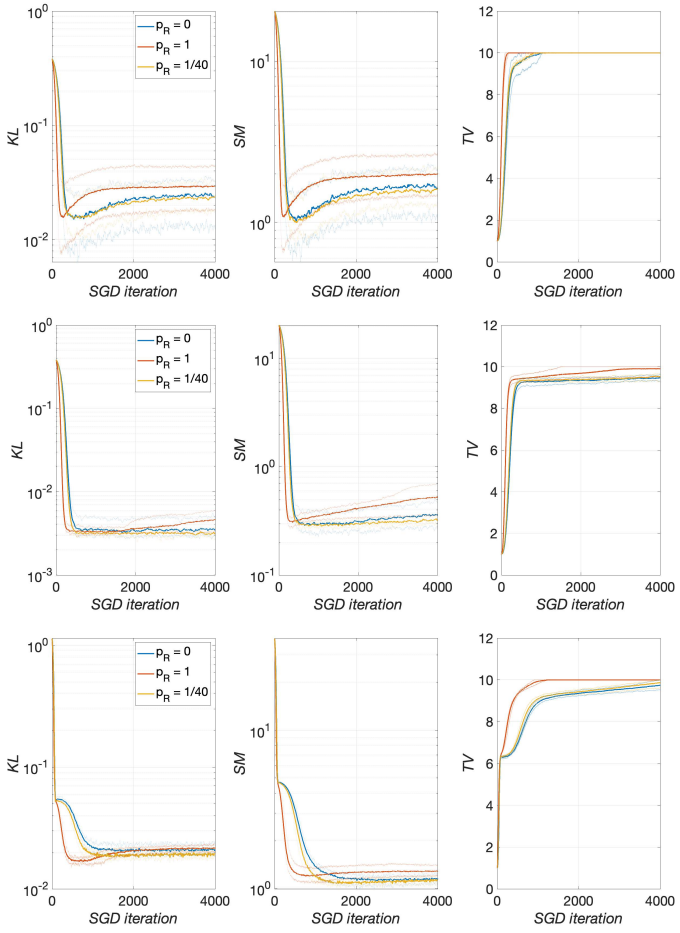

Figure 6: *Experiments in d=14:* (Top) The evolution of the KL divergence, the score matching metric and the TV norm of the trained measure (i.e., the $\mathcal{F}_1$ norm) during training for Algorithm 1 with $\mathcal{X} = \mathbb{S}^{14}$, $m = 64$, $p_R = 0, 1, 1/40$, $s = 0.02$, $\alpha = 10 + 50p_R$, $n = 10^3$, $N = 2 \cdot 10^3$. The plots show the average, maxima and minima over six runs with different training and test samples, initializations and noise realizations, but with the same teacher network with an angle of 2.87 rad between neurons. In comparison, the non-parametric kernel density estimator reaches a KL divergence of 0.18. (Middle) Same experiments with $n = 10^4$ and $N = 2 \cdot 10^4$. The non-parametric kernel density estimator reaches a KL divergence of 0.11. (Bottom) Same experiments with $n = 10^4$ and $N = 2 \cdot 10^4$, and angle of 1.37 rad between teacher neurons. In comparison, the non-parametric kernel density estimator reaches a KL divergence of 0.15.

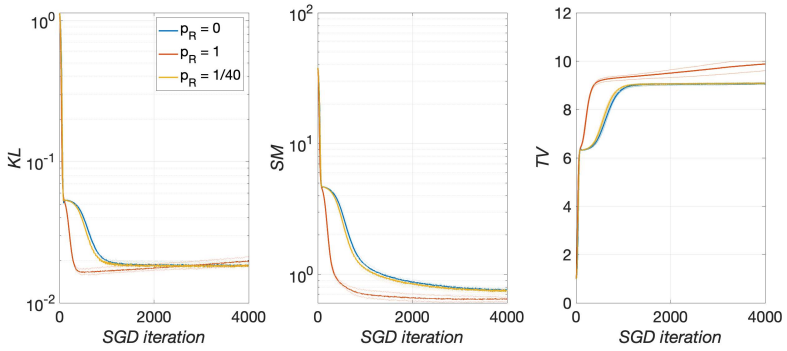

Figure 7: *Experiments in* $d = 14$*:* Same setting as bottom row of Figure 6 (i.e., angle 1.37 rad), but with $n = 10^5, N = 2 \cdot 10^5$.

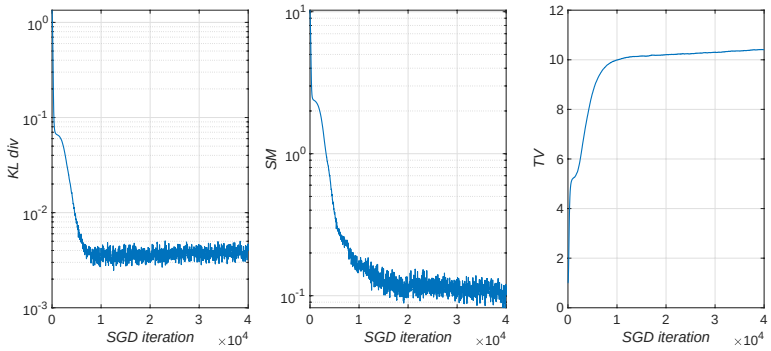

Figure 8: *Experiments in* $d = 2$*:* Evolution of the KL divergence, score matching and TV norm for the training dynamics of `KLdual_1e4_points_monomodal.mp4` and Figure 3.

relative performance of score matching ($p_R = 1$) against the other two choices. It is also remarkable that the value of the KL divergence at the end of training in Figure 7 is about $2 \cdot 10^{-2}$, which is very similar to the value obtained in the bottom row of Figure 6 despite the increase in n. This is at odds with the statistical analysis of Domingo-Enrich et al. (2021), which predicts a decrease of the KL test error as $O(1/\sqrt{n})$ in the case where the approximation error is null. Hence, even though the KL values achieved are low, there is some effect at play which hinders optimization in the monomodal case.

To further understand the "bumpy" curves observed in the monomodal case, we return to experiments in $d = 2$, this time with almost perpendicular teacher neurons. The results are shown in Figure 3. We observe similar trends in the curves of Figure 8. In Figure 3, we see that the training occurs in two stages: first the student neurons first concentrate rather quickly near the mode of the teacher distribution: second, they slowly converge toward the teacher neurons. The bump in the KL and SM curves occurs when the first training stage ends and the second one sets in.

These findings seem to suggest an interesting dichotomy: when the two teacher neurons are far away and the distribution is bimodal, sampling is hard but training is easier; when the teacher neurons are closer and the distribution is monomodal, the opposite is true. In a generic situation, both issues may be present. More experiments are required to formulate concrete statements.