Methods for optimization and regularization of Generative Models

Michael Arbel

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Gatsby Computational Neuroscience Unit

University College London

April 18, 2021

I, Michael Arbel, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

This thesis studies the problem of regularizing and optimizing generative models, often using insights and techniques from kernel methods. The work proceeds in three main themes.

Conditional score estimation

We propose a method for estimating conditional densities based on a rich class of RKHS exponential family models. The algorithm works by solving a convex quadratic problem for fitting the gradient of the log density, the *score*, thus avoiding the need for estimating the normalizing constant. We show the resulting estimator to be consistent and provide convergence rates when the model is well-specified.

Structuring and regularizing implicit generative models

In a first contribution, we introduce a method for learning Generative Adversarial Networks, a class of Implicit Generative Models, using a parametric family of Maximum Mean Discrepancies (MMD). We show that controlling the gradient of the critic function defining the MMD is vital for having a sensible loss function. Moreover, we devise a method to enforce exact, analytical gradient constraints.

As a second contribution, we introduce and study a new generative model suited for data with low intrinsic dimension embedded in a high dimensional space. This model combines two components: an implicit model, which can learn the low-dimensional support of data, and an energy function, to refine the probability mass by importance sampling on the support of the implicit model. We further introduce algorithms for learning such a hybrid model and for efficient sampling.

Optimizing implicit generative models

We first study the Wasserstein gradient flow of the Maximum Mean Discrepancy in a non-parametric setting and provide smoothness conditions on the trajectory of the flow to ensure global convergence. We identify cases when this condition does not hold and propose a new algorithm based on noise injection to mitigate this problem.

In a second contribution, we consider the Wasserstein gradient flow of generic loss functionals in a parametric setting. This flow is invariant to the model's parameterization, just like the Fisher gradient flows in information geometry. It has the additional benefit to be well defined even for models with varying supports, which is particularly well suited for implicit generative models. We then introduce a general framework for approximating the Wasserstein natural gradient by leveraging a dual formulation of the Wasserstein pseudo-Riemannian metric that we restrict to a Reproducing Kernel Hilbert Space. The resulting estimator is scalable and provably consistent as it relies on Nyström methods.

Impact Statement

Powerful generative models for high dimensional data have found application in a wide variety of domains, notably in computer imaging, with GANs being a particular success story. In the entertainment industry, GANs have been successful in image painting and style transfer. In medicine, GANs have been used to augment datasets to improve generalization performance in supervised learning for medical image classification, which can improve ML-based medical diagnosis tools. In climate science, GANs have been used to visualize the effects of climate change: seeing the effect of rising sea levels on one's own city can bring home the impact of global warming in a more accessible way than abstract figures.

Given that, all else being equal, the quality of images generated by our Generalized Energy Based Model (GEBM) improves over images generated by a GAN with the same generator and critic networks, we anticipate that GEBMs will be broadly applicable in improving generative modeling across a range of applications.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Arthur Gretton, for his continued support and guidance throughout my Ph.D. Arthur was always patient, understanding, and helped me evolve throughout this journey. Arthur also gave me the freedom to explore questions I found the most exciting and constantly showed enthusiasm for my projects and ideas. I would like to thank the Gatsby unit for providing such a stimulating environment where I constantly learned from all the members. I am also thankful to Francis Bach and John Shawe-Taylor for agreeing to be my thesis examiners and for their valuable feedback and stimulating discussions.

I am also very grateful to all my collaborators from whom I learned so much. I thank Heiko for his helpful advices when I started my PhD. I especially thank Danica for her great help in introducing me to GANs, for the fruitful scientific discussions we had and all the amazing coding tricks she showed me! I thank Mikolaj for being of great help with the MMD-GAN project. I thank Anna and Adil for their multiple insights in Optimal transport and for fruitful discussions. I thank Wuchen Li, Guido Montufar and Ferenc for introducing me to Natural gradient methods and for their inspiring ideas. I thank Tolga and Umut for introducing me to the camera synchronization problem and for their enthusiasm about applying MMD gradient flows to it. I thank Liang for being of great help with the GEBM project and Mihaela Rosca for fruitful discussions on the generalization and optimization of GANs. I thank Pierre Glaser for his enthusiasm and fruitful discussions about optimal transport and many other things. I thank Samuel for introducing me to the problem of Barycenter of measures. I thank Edouard, Eugene, and Louis for showing me how to re-think the fundamentals of Deep Learning. I thank Ted for

his inspiring ideas, for his enthusiasm about connecting the Wasserstein Natural gradient method and Reinforcement Learning. I am particularly grateful to Arnaud Doucet for introducing me to the topic of Sequential Monte Carlo and for his patience and kindness while mentoring me. I also thank Alex Matthews for being a great collaborator on this project and for his patience and enthusiasm. I thank Barry, Ana and Mike for all their help and support and for making my time at Gatsby such a great experience. I am also extremely lucky I have made several friends during my time at the Gatsby Unit: Heishiro, Sanjeevan, Kevin, Ira, Yash, Wenkai, Lea, Eszter, Jorge, Ilyes, Kirsty and many others who made these years memorable. I am immensely grateful to Julia Peyre for providing me emotional support and for her patience during this journey. Finally, I'd like to thank my family for their love and support and for constantly encouraging me ever since I can remember myself.

Contents

Chapter 1

Introduction

This thesis addresses the problem of learning an unknown probability distribution from data. More specifically, we are interested in modeling, regularizing, and optimizing generative models, often using insights and techniques from kernel methods.

Modeling unknown probability distributions from data is one of the most general problems in machine learning. Once an estimator of a probability distribution is learned, it can be used to solve various problems such as classification, regression, matrix completion, and other prediction tasks. Framing those prediction tasks as a distribution estimation problem can account for multiple effects, such as multimodality or heteroscedasticity of the data, which are increasingly common given the diversity and complexity of modern machine learning problems.

At a high level, a generative model defines a probability distribution $\mathbb{Q}_\theta$ from a class of models $\mathcal{Q}$. The goal is to select a distribution from the set $\mathcal{Q}$ that minimizes some discrepancy measure between the data distribution and the model. In a high dimensional setting, successful models need to incorporate additional knowledge about the data to overcome the curse of dimensionality [Donoho, 2000]. Such knowledge can be, for instance, in the form of independence assumptions between variables (a graphical model) or a low intrinsic dimensionality of their support.

We first show how to incorporate independence structure in an expressive class of non-parametric generative models based on Reproducing Kernel Hilbert Spaces (RKHS). Those models fall in the category of *explicit models* since they directly

specify a density model over the data, and can thus be estimated using methods such as Maximum Likelihood or Score Matching [Hyvärinen, 2005]. We show how to learn them using a *score matching* approach, which results in a convex quadratic optimization problem. Unlike Maximum likelihood estimation, this method avoids the need for estimating the normalizing constant.

In the second part of this thesis, we turn to another class of models, *implicit generative models* (IGMs), which enforce a low dimensional structure in the sampling mechanism. Implicit models do not necessarily admit a density defined over the whole data-space. Consequently, learning these models is challenging since likelihood methods or score matching cannot be used. We introduce a method for learning them, using a family of Maximum Mean Discrepancies (MMD) parametrized by a reproducing kernel. We show that constraining the smoothness of the parametric kernel results in a sensible loss for training IGMs. In a second contribution, we introduce a new class of models that combines both *implicit* and *explicit* models. The implicit model is in charge of learning the low-dimensional support of the data. The explicit model then provides importance weights used to refine the mass on the support defined by the implicit model, thus capturing effects such as multimodality. We propose a method for learning this hybrid model, and derive sampling schemes that exploit the model's latent structure to increase efficiency.

Once a model is specified, and a loss is selected, there remains the question of effectively optimizing such models. In many situations, the resulting loss is non-convex, making the optimization more likely reach unsatisfactory local optima. Moreover, usual optimization procedures might be sensitive to the parameterization of the generative model. When such parametrization is less favorable, this can result in a more challenging optimization problem. In the third part of this thesis, we propose optimizers for implicit models to address such challenges using the Wasserstein geometry.

1 Conditional score estimation

In Chapter 3, we consider the problem of density estimation for high dimensional data. A classical approach assumes the density belongs to a suitable hypothesis class and then defines a learning rule for selecting a suitable estimator within such class.The choice of the hypothesis class impacts both the generalization and optimization properties of the learning algorithm.

When choosing a parametric model p_θ with finite-dimensional parameter θ, the density estimation task consists of estimating an optimal parameter θ that best fits the data, using, for instance, maximum likelihood estimation. The first challenge with many such models comes from the normalizing constant's intractability, making learning with maximum likelihood particularly challenging. Thus, various methods propose either to estimate the gradient of the log-likelihood such as contrastive divergence [Hinton, 2002] or to bypass the need for estimating the normalizing constant [Hyvärinen, 2005, Gutmann and Hyvärinen, 2012]. Despite those developments, the non-convexity of the resulting losses can lead to sub-optimal local minima even in simple cases such as Mixtures of Gaussians [Jin et al., 2016]. One notable exception is the class of exponential families described by a finite-dimensional parameter vector T called the natural parameter and a predetermined function $\phi(x)$ called the sufficient statistic. An exponential model's density is proportional to $\exp(\langle T, \phi(x)\rangle)$ and generally yields a concave log-likelihood, thus suitable for maximum likelihood estimation. Unfortunately, when the data distribution does not belong to such a parametric class, there is no possibility to control the bias introduced by the model.

To increase the estimator's expressivity, one could turn to non-parametric estimation. One popular setting is to consider hypothesis classes as large as β-Sobolev or β-Hölder classes with some smoothness parameter β. In this setting, and provided the correct smoothness parameter β is known, estimators as simple as Kernel Density Estimation (KDE) achieve the minimax optimal rate of $O(n^{-\frac{2\beta}{2\beta+d}})$ over those classes [Tsybakov, 2009]. When the smoothness parameter β is unknown, it is still possible to construct adaptive estimators, for instance, using the method of aggregation [Rigollet and Tsybakov, 2007]. However, in either case, the estimator's

convergence rate exhibits an exponential dependence on the dimension of the data. This dependence is a consequence of a large hypothesis class, making any systematic 'search' intractable in large dimensions.

Luckily, the data often exhibits additional structure and, if encoded in the hypothesis class, one can construct estimators that have improved rates of convergence on these restricted classes. Kernel methods offer a general framework for encoding a priori structure while still retaining flexibility and tractability of the estimation [Shawe-Taylor and Cristianini, 2000, 2004, Steinwart and Christmann, 2008]. To achieve this in the context of density estimation, several works proposed to extend the exponential family models to the case where the natural parameter T is a function in a Reproducing Kernel Hilbert Space $\mathcal{H}$ with a reproducing kernel k [Gu and Qiu, 1993, Barron and Sheu, 1991, Pistone and Sempi, 1995, Canu and Smola, 2006, Fukumizu, 2009]. The sufficient statistic $\phi(x)$ is then given by the feature map $k(x, .)$. Although the normalizing constant becomes intractable in general, Sriperumbudur et al. [2017] proposed to use the *score matching* approach introduced in Hyvärinen [2005] to circumvents the need for estimating the normalizing constant. The natural parameter can then be estimated in closed-form using a generalized Representer Theorem [Sriperumbudur et al., 2017, Theorem A.1] and solving a strongly convex quadratic problem. The estimator's convergence rate depends only on the smoothness of the solution. However, there is still a hidden curse of dimensionality since the smoothness requirement becomes more stringent as the dimension increases [Sriperumbudur et al., 2017, Example 3].

Contribution We introduce a method for density estimation using non-parametric exponential families. To reduce the complexity of the modeling task, we allow the method to take advantage of a graphical model [Pearl, 2001, Jordan, 1999] where each variable depends only on a subset of parent variables. This graphical model imposes a factorization of the density into a product of conditional densities. To estimate each factor independently, we extend the non-parametric family framework of Sriperumbudur et al. [2017] to conditional distributions $p(x|z)$ conditionally on an observed variable z. The natural parameter T_z is still an element of an RKHS

$\mathcal{H}$. However, to include information from samples in the neighborhood of z, we require the map $T : z \mapsto T_z$ to belong to a vector-valued RKHS [Micchelli and Pontil, 2005]. This space contains functions of z that take values in the RKHS $\mathcal{H}$ while still having some smooth dependence on z. The goal becomes to estimate the map T. For this purpose, we extend the score matching framework proposed in Hyvärinen [2005] to the vector-valued kernel setting. This approach results in a strongly convex optimization problem in the vector-valued RKHS space that can be solved in closed form by duality using the Representer theorem.

We then establish consistency of the estimator of T in the well-specified case by generalizing the arguments of Sriperumbudur et al. [2017] to the vector-valued setting. The distance between the estimator and the true unknown natural parameter provably converge at a rate of $\mathcal{O}(n^{-\frac{1}{4}\min(1,\frac{2\gamma}{\gamma+1})})$ where γ is a parameter controlling the smoothness of the true unknown natural parameter. While our proof allows for general vector-valued RKHSs, we provide a practical algorithm for a specific case, which takes the form of a linear system of size $n \times d_x$ with d_x being the dimension of x. Hence, given a factorization of the joint density into a product of conditional densities, the proposed method can estimate the joint density by first estimating each conditional density independently and then combining the estimates.

2 Structuring and regularizing implicit generative models

In Chapters 4 and 5, we consider the problem of modeling data supported on a set with a low intrinsic dimension. This notion of dimension can be formalized in the Minkowski sense without necessarily requiring a tangent structure for the data support, such as a smooth manifold structure [Nakada and Imaizumi, 2020]. One significant implication for the data distribution is the absence of a well-defined density w.r.t. the Lebesgue measure. In this setting, explicit models tend to spread mass over the different dimensions of the data-space since, by construction, they possess a density.

Unlike explicit models, Implicit Generative Models (IGMs), popularized by

Goodfellow et al. [2014], do not directly specify a density function. Instead, they assume each observation x to be generated by a possibly low-dimensional latent variable z with a pre-defined distribution η. Producing a sample from an IGM requires mapping a latent sample z to the data space using a function G_θ selected from some parametric family: $x = G_\theta(z)$. When the latent variable is lower-dimensional, the support of the IGM is also lower-dimensional and thus does not admit a density w.r.t. Lebesgue measure on the data-space [Arjovsky and Bottou, 2017]. Consequently, IGMs seem natural to use when the data support has a small intrinsic dimension as they can concentrate the mass where the data is supported.

As a first contribution, we will consider the challenge of learning IGMs. The absence of a density in IGMs makes likelihood methods ill-suited for learning them, opening the way to other methods, such as *Adversarial training* [Goodfellow et al., 2014]. Those methods often require approximately solving a challenging bilevel optimization problem [Liang, 2017]. In a second contribution, we will consider the problem of expressiveness of IGMs. Such models often rely on a pre-determined simple latent distribution η, such as a gaussian, that is transformed by the generator G into a potentially multi-modal high-dimensional distribution. This restriction often requires high modeling and learning complexity for the generator function [Cornish et al., 2020].

2 .1 Optimized MMD under gradient constraints

In Chapter 4, we consider the problem of learning IGMs. In the absence of a density, *adversarial training* [Goodfellow et al., 2014] offers an alternative to maximum likelihood methods for leaning IGMs. This method introduces an auxiliary model called the critic function E that is selected from a model class $\mathcal{E}$ to maximize a discrepancy $\mathcal{L}$ between samples from the data distribution $\mathbb{P}$ and samples from the $\mathbb{Q}_\theta$:

$$E^\star(\theta) \in \arg\max_{E \in \mathcal{E}} \mathcal{L}(E, \mathbb{P}, \mathbb{Q}_\theta). \tag{1.1}$$

Goodfellow et al. [2014] considered a particular choice for the discrepancy $\mathcal{L}$ which results in the Jensen-Shanon divergence when the set $\mathcal{E}$ consists in all measurable functions, a computationally intractable case. For general choices of discrepancies, their maximum value over a class $\mathcal{E}$ results in other divergences, such as f-divergences [Nowozin et al., 2016]. Those divergences can be minimized with respect to the parameters θ of the IGM $\mathbb{Q}_\theta$:

$$\theta^\star \in \arg\min_{\theta\in\Theta} \mathcal{L}(E^\star(\theta), \mathbb{P}, \mathbb{Q}_\theta). \tag{1.2}$$

A notable class of symmetric divergences called Integral Probability Metrics (IPMs) are obtained when the discrepancy $\mathcal{L}(E, \mathbb{P}, \mathbb{Q}_\theta)$ is the difference in moments $\int E \,\mathrm{d}\mathbb{P} - \int E \,\mathrm{d}\mathbb{Q}_\theta$ [Müller, 1997]. IPMs are thus only determined by the set $\mathcal{E}$. All those methods have the common property of approximately solving the bilevel optimization defined by (1.1) and (1.2) and have yielded impressive empirical results, particularly for image generation, that are far beyond the quality of samples seen from most earlier generative models [Karras et al., 2018, Radford et al., 2016, Gulrajani et al., 2017, Huang et al., 2018a, Jin et al., 2017]. Despite this success, the earliest attempts in training IGMs resulted in instabilities, and pathological behavior [Liang, 2017].

These challenges inspired further research to stabilize training by adding various regularizations to the optimization [Nagarajan and Kolter, 2017, Kodali et al., 2017]. Alternatively, Arjovsky et al. [2017] proposed to use the 1-Wasserstein distance as a loss function for training IGMs. This distance is an IPM with the set $\mathcal{E}$ consisting of all 1-Lipschitz functions. It metrizes the weak topology [Villani, 2009, Theorem 6.9], unlike the Kullback-Leibler or Jensen-Shanon divergence, which define a stronger topology. This property implies the continuity of the loss in the parameters of the IGMs despite its density being ill-defined [Arjovsky and Bottou, 2017]. Unfortunately, estimating the Wasserstein distance suffers from the curse of dimensionality [Weed et al., 2019]. In practice, Arjovsky et al. [2017] proposed to approximate it by restricting the set $\mathcal{E}$ to a parametric subset of neural networks under a Lipschitz constraint. To approximately enforce this constraint, subsequent work developed

regularization techniques such as the *gradient penalty* [Gulrajani et al., 2017] and the *spectral normalization* [Miyato et al., 2018].

Another line of work proposes losses for which the inner optimization (1.1) can be solved, at least partially, in closed form, thus simplifying the bilevel optimization and allowing for easier control over its smoothness. The Maximum Mean Discrepancy (MMD) is one particular instance that falls in the category of IPMs with the critic class $\mathcal{E}$ being a unit ball of Reproducing Kernel Hilbert Space with kernel k [Gretton et al., 2012]. It is appealing since it admits tractable unbiased estimators. However, using MMDs with simple kernels, such as a Gaussian kernel, does not achieve state-of-the-art performance on high dimensional data such as images [Li et al., 2017]. Bińkowski* et al. [2018] proposed to overcome this limitation by using features of the samples as inputs to the kernel k. Those features are computed using a neural network $\phi(x)$ and define a parametric family of kernels $k_\phi(x, y) = k(\phi(x), \phi(y))$. The network parameters are optimized so as to maximize the MMD between the data distribution and the IGM, thus falling back to the original bi-level optimization problem. Indeed, in this case, the model class $\mathcal{E}$ consists of a union of unit balls from different RKHS's each corresponding to a kernel k_ϕ. The optimization problem in (1.1) can no longer be solved in closed form over $\mathcal{E}$. However, it can be solved on each RKHS ball, resulting in an MMD with a particular choice for the kernel. This kernel's parameters are then selected to maximize the MMD between the data and the model $\mathbb{Q}_\theta$. This approach yields good results when using an additional *gradient penalty* as a training trick, even though such penalty was originally introduced so as to approximate Wasserstein distances [Arjovsky et al., 2017, Gulrajani et al., 2017].

Contribution We introduce a new loss for training IGMs using a family of MMDs parametrized by reproducing kernels. This loss is obtained by choosing the model class $\mathcal{E}$ to be a union of RKHS balls. However, unlike in Bińkowski* et al. [2018], we allow the radii of those balls to adjust to the kernel's smoothness. This flexibility favors critic functions that are flat in regions where the data distribution $\mathbb{P}$ has high mass. The proposed method takes inspiration from an approach widely used in

semi-supervised learning [Bousquet et al., 2004, Section 2]. In such approaches, a classifier/critic E is encouraged to be smooth on the data support by constraining the sum of its variance and expected gradient norm under the data distribution. In our case, this constraint is encoded in the radii of the RKHS balls. We show that the resulting loss is continuous in the weak topology, thus providing a sensible loss for learning IGMs.

2 .2 Generalized energy based models

By defining low-dimensional support that can vary during training, implicit models are well suited for modeling data with low-intrinsic dimensionality and often result in impressivly sharp samples. However, IGMs still suffer from at least three limitations related to their capacity and the efficiency of their training methods.

As discussed in Section 2 .1, the success of IGMs comes at the cost of introducing auxiliary models, in the form of critic functions E, that are solely used for training and are not part of the final model. This *waste* in modeling motivated further research to use the critic model during sampling in order to improve the sample quality of the IGM [Azadi et al., 2019, Turner et al., 2019, Neklyudov et al., 2019]. These earlier approaches interpreted the critic function as defining a density ratio between the IGM and the data distribution. This ratio is then used in sampling algorithms such as rejection sampling or the Hasting-Metropolis algorithm, with the IGM acting as a proposal distribution. However, the critic's interpretation as a density ratio becomes problematic when the IGM and the data distribution do not share the same support, which is the usual scenario when the generator maps low-dimensional noise into a high dimensional sample space.

A second limitation of IGMs is the use of pre-determined simple latent distributions η, such as a Gaussian, that are transformed by the generator G into a potentially multi-modal high-dimensional distribution. This choice often requires high modeling complexity for the generator function, especially when the data distribution is multi-modal [Cornish et al., 2020]. On the other hand, explicit models can easily capture multi-modality using mixture-models, for instance.

Finally, as discussed earlier in Section 2 .1, when it comes to learning IGMs,

the losses need to be sensitive to the mismatch between the support of the model and the data distribution. This requirement can be satisfied when the divergences used for training are continuous in the topology of weak convergence of probability measures. However, with this requirement, the benefits of using stronger topologies, such as the topology induced by likelihood methods, are not guaranteed anymore. For instance, statistical efficiency is a crucial property of estimators obtained using likelihood methods [Daniels, 1961]. Consequently, explicit models trained by maximum likelihood benefit from favorable statistical properties, unlike IGMs, which require weaker training losses. A natural question is whether implicit and explicit models can be combined in a way that exploits both of their complementary strengths.

Contribution In Chapter 5, we introduce a new class of models that combines both *implicit* and *explicit* models. The implicit model is in charge of learning the low-dimensional support of the data. The explicit model, called the energy by analogy to statistical physics models, provides importance weights used to refine mass on the implicit model's support, therefore capturing multi-modality. Crucially, both parts of the model are trained together without the need for additional modeling. The energy is learned by maximizing a generalized relative likelihood to the support of the implicit model. This energy allows computing a provably weak divergence between the data and the implicit model. We further show that the resulting divergence is smooth enough in the parameters of the IGM so that stochastic gradient methods are guaranteed to converge to a local optimum. Hence, the explicit part of the model benefits from the strong topology, while the implicit part relies on the weak topology to capture the data support. Additionally, we show that sampling from such a model can be achieved by leveraging the latent structure and performing MCMC in the lower dimensional latent space. This model results in a framework that generalizes and unifies two classes of models a priori orthogonal to each other.

3 Optimizing implicit generative models

Characterizing and finding optimal solutions for IGMs is of particular importance to better understand their generalization properties and devise more efficient optimiza-

tion algorithms.

Central to this question is the choice of the geometry used to study this problem. Using the euclidean geometry defined by the parameters of the IGM does not account for the probabilistic structure of the problem. As a result, optimization algorithms such as usual gradient descent or even adaptive optimizers [Kingma and Ba, 2015] are still sensitive to the model's parameterization. Alternatively, Amari [1985] introduced the Fisher natural gradient to account for the probabilistic structure of the problem. This gradient relies on the geometry defined by the Fisher-Rao metric on the space of probability distributions [Holbrook et al., 2017], and can thus result in methods robust to parameterization. The geometry defined by the Fisher-Rao metric requires the model to have a well-defined density and is thus well adapted to explicit models. However, it does not apply to IGMs since they usually include mutually singular distributions.

An alternative geometry in probability space, the Wasserstein geometry, relies on optimal transport between probability distributions and was first introduced in Jordan et al. [1998] and Otto and Villani [2000]. It remains well-defined even for mutually singular distributions and is therefore well suited for implicit models [Mroueh et al., 2019]. This geometry was formalized in Jordan et al. [1998], Otto and Villani [2000], Ambrosio et al. [2004] using the idea of a minimizing movement scheme of a loss functional $\rho \mapsto \mathcal{L}(\rho)$ defined over the set $\mathcal{P}_2$ of probability distributions with finite second order moments. This scheme constructs a sequence of probability distributions ρ_k by iteratively solving the minimization problem:

$$\rho_{k+1} \in \arg\min_{\rho \in \mathcal{P}_2} \mathcal{L}(\rho) + \frac{1}{2\tau} W_2^2(\rho, \rho_k) \tag{1.3}$$

with τ being a step-size parameter controlling how close the next iterate ρ_{k+1} must be to ρ_k in terms of the Wasserstein-2 distance. Under mild assumption on the functional $\mathcal{L}$ and formally taking the limit of infinitely small step-size ($\tau \to 0$) results in continuous time path in probability space ρ_t known as the Wasserstein gradient flow of $\mathcal{L}$. This gradient flow describes the variations of probability distributions in terms of displacement of the probability mass from a location in space to another, also

called *horizontal movement* [Santambrogio, 2010]. This gradient flow shares some similarities with the optimization trajectories of implicit models in probability space. Indeed training IGMs requires producing samples from the model and transporting them to a new location by varying the model's parameters, hence mimicking a *horizontal movement*.

In Chapter 6, we will study the convergence properties of the gradient flow of the Maximum Mean Discrepancy functional in a non-parametric setting. In Chapter 7, we will propose an efficient optimizer that approximates the Wasserstein gradient flow when constrained to a family of parametric models.

3 .1 Wasserstein gradient flow of the MMD

Despite the empirical success of IGMs, providing a complete description of their generalization properties remains a challenge. Part of this challenge is related to the highly non-convex nature of the optimization problem. Here, our goal is to investigate further some properties of the optimization trajectories that result from learning IGMs.

An interesting starting point is the non-parametric setting, where the model consists of a (potentially infinite) set of particles free to move in space. Thus, the goal becomes to minimize a loss functional over a non-parametric class of probability distributions over those particles. The trajectory of those particles can be described by a Wasserstein gradient flow of the functional in probability space [Chizat and Bach, 2018a, Mroueh et al., 2019]. The parametric setting used in practice is a constrained version of the Wasserstein gradient flow where the set of probability distributions is restricted to a parametric family. This constraint causes a departure from a pure particle flow. Despite this fact, we anticipate that the theoretical study of particle flow convergence will provide helpful insights into conditions for convergence of IGMs, and ultimately, improvements to their training algorithms. The Wasserstein flow of the MMD functional is of particular interest since it yields competitive empirical results when used for training IGMs [Bińkowski* et al., 2018]. Although those works rely on optimized kernels, considering the MMD with a fixed kernel results in a closed-form expression and makes a mathematical analysis more accessible.

Several works studied the Wasserstein flows of functionals similar to the MMD. Carrillo et al. [2006] studied such flows under convexity assumptions of the functional and proved their convergence to a global minimum. However, these convexity assumptions are unlikely to hold in the case of the MMD because of the opposing terms resulting from the attractive and repulsive forces that will typically not be simultaneously convex. More recently, Mei et al. [2018] proved that adding entropic regularization to functionals similar to the MMD results in global convergence of the particle flow. In this case, the optimal distribution is in general different from the one obtained without adding entropic regularization. In the context of optimization of neural networks, Rotskoff and Vanden-Eijnden [2018] viewed the parameters of a one hidden layer network as particles that are drawn from a probability distribution and considered the *infinite width limit* of such networks. They provided an informal global convergence result of the gradient flow of a representative particle. The work of Chizat and Bach [2018a] generalized and provided rigorous proofs of global convergence using the Wasserstein geometry formalism. Their results rely on a homogeneity structure of the loss function, which holds systematically in the context of deep neural networks. In the case of the MMD, this structure can hold for kernels that are linear in at least one of their variables. However, this requirement is incompatible with the MMD being a strict distance since the kernel is not characteristic anymore [Sriperumbudur et al., 2009].

Contribution In Chapter 6, we study the Wasserstein gradient flow of the MMD functional in a setting that is agnostic to the specific form of the kernel. We provide a smoothness condition on the trajectory path in probability space that guarantees convergence to a global minimum. This condition states that the negative Sobolev distance between the current particles' distribution and the global minimizer remains finite. While this condition is trajectory dependent, it indicates that a common reason for failure to attain global optimality is when the particles' support collapses. We exhibit simple cases when this condition does not hold, thus resulting in pathological trajectories. We then propose a modified algorithm for the particle flow that adds a Gaussian perturbation to the particles' location when evaluating its gradient.

This procedure maintains the global optimum as a stationary point, unlike entropic regularization, and is shown empirically to mitigate the collapse problem.

3 .2 Scalable Wasserstein natural gradient

Usually, when optimizing generative models, the probability distributions matter more than the particular parameters used to describe them. This principle is central to natural gradient methods which construct optimization algorithm based on a proximity measure between probability distributions [Amari, 1985, 1998]. When the Kullback-Leibler (KL) is used as a proximity measure, this results in the so-called Fisher Natural Gradient (FNG) descent, which received particular attention in the context of variational inference [Khan and Lin, 2017, Zhang et al., 2018]. This approach has the particular advantage that the resulting optimization trajectory in probability space is invariant to parametrization in the continuous-time limit [Ollivier et al., 2011]. Because of its close connection to the KL, this method is only applicable when the generative model has a well-defined probability density relative to some reference measure. Unfortunately, this is rarely the case in the context of IGMs.

On the other hand, the Wasserstein distance remains well-defined for IGMs and appears to be a viable alternative proximity measure to the KL. This amounts to considering a parametric Wasserstein Gradient flow ρ_{θ_t}. Such a flow would be obtained by taking the continuous-time limit of a proximal scheme similar to (1.3), albeit constrained to the parametric set of IGMs:

$$\rho_{\theta_{k+1}} \in \arg\min_{\rho_\theta \in \mathcal{Q}} \mathcal{L}(\rho_\theta) + \frac{1}{2\tau} W_2^2(\rho_\theta, \rho_{\theta_k}). \tag{1.4}$$

By doing so, Li and Montufar [2018a,b], Li [2018] introduce the notion of Wasserstein natural gradient (WNG) which drives the dynamics of the parameter θ_t describing the trajectory ρ_{θ_t}. Operating directly on the parametric 'manifold' of probability distributions defined by the IGMs, the Wasserstein Natural gradient flow yields the same trajectory in probability space regardless of the model's parameterization. Similarly to the FNG, the WNG can also be computed by pre-conditioning the usual *euclidean gradient* with the inverse a symmetric positive matrix: the *Wasserstein*

Information matrix. Because of this inversion, direct estimation of either FNG or WNG becomes quickly infeasible for current large models with typically millions of parameters.

So far, prior works focused on finding efficient algorithms to estimate the Fisher natural gradient thus achieving a good trade-off between computation and empirical performance [Martens and Grosse, 2015, Grosse and Martens, 2016, George et al., 2018, Heskes, 2000, Bernacchia et al., 2018]. However, these methods exploit the particular structure of the FNG and do not apply to estimating the Wasserstein Natural gradient, to the best of our knowledge. Recently, Li et al. [2019] proposed to use a proximal scheme similar to (1.4) to compute approximate updates of the WNG flow without directly estimating the WNG. They propose to replace the Wasserstein penalty in (1.4) by its non-parametric linearization known as the Negative Sobolev Distance $\|\rho_\theta - \rho_{\theta_k}\|^2_{H^{-1}(\rho_{\theta_k})}$ [Peyre, 2018, Villani, 2003]. While such penalty is still intractable to compute, Li et al. [2019] express it as the optimal value of some convex functional optimization problem obtained by duality and approximated in practice using a finite set of basis functions. However, the Negative Sobolev Distance between two probability distributions can be infinite, in theory, when the supports of those distributions do not overlap, limiting the applicability to IGMs. However, this pathology disappears when directly considering the infinitesimal limit of (1.4) as described by the Wasserstein Natural Gradient.

Contribution. In Chapter 7, we introduce a scalable and provably consistent estimator of the Wasserstein Natural Gradient. To achieve this, we first provide a dual formulation of the Wasserstein Information Matrix (WIM) as the optimal value of some convex functional optimization problem. Unlike in Li et al. [2019], this formulation is provably well-defined since the WIM represents only elements that are in the tangent space of the current probability distribution ρ_θ. The gradient of the optimal solution provides a vector field ϕ_u allowing to perform an *infinitesimal* transport of mass from ρ_θ in a parameter direction u. We then express the Wasserstein Natural Gradient as the optimal solution of some convex-concave saddle problem using the dual formulation used to express the WIM. This formulation allows us to derive an

estimator of the WNG by restricting the functional space to a Reproducing Kernel Hilbert Space. The WNG may then be approximated using a Nyström method. We provide a convergence rate for the estimator in two settings: a well-specified and a misspecified setting. The well-specified setting considers the case when the optimal vector field ϕ_u is a gradient of a function in the RKHS space. The misspecified case assumes that the vector field can be approximated to arbitrary precision in $L_2(\rho_\theta)$ by the gradient of a function in the RKHS. The RKHS norm of the approximating function is allowed to grow polynomially with the precision. This growth is precisely what determines the estimator's convergence rate. An experimental evaluation confirms that our kernel approach can accurately estimate the WNG while being scalable and robust to the model's parametrization.

4 Structure of the Thesis

The five main thesis chapters are based on the following publications. Source code for all proposed methods in this is publicly available.

1. Chapter 3: Conditional score estimation

   Arbel, Michael and Arthur Gretton. Kernel Conditional Exponential Family. *In AISTATS*, 2018

   Code: https://github.com/MichaelArbel/KCEF

2. Chapter 4: Optimized MMD under gradient constraints

   Arbel*, Michael, Sutherland*, Danica J., Mikołaj Bińkowski, and Arthur Gretton. On gradient regularizers for MMD GANs. *In NeurIPS*, 2018

   Code: https://github.com/MichaelArbel/Scaled-MMD-GAN

3. Chapter 5: Generalized energy based models

   Arbel, Michael, Liang Zhou, and Arthur Gretton. Generalized energy based models. In *ICLR*, 2021

   Code: https://github.com/MichaelArbel/GeneralizedEBM

4. Chapter 6: Wasserstein gradient flow of the Maximum Mean Discrepancy

   Arbel, Michael, Anna Korba, Adil Salim, and Arthur Gretton. Maximum Mean Discrepancy Gradient Flow. *In NeurIPS*, 2019a

   Code: https://github.com/MichaelArbel/MMD-gradient-flow

5. Chapter 7: Scalable Wasserstein natural gradient

   Arbel, Michael, Arthur Gretton, Wuhen Li, and Guido Montufar. Kernelized Wasserstein Natural Gradient. *In ICLR*, 2019b

   Code: https://github.com/MichaelArbel/KWNG

Other Contributions Works published over the course of this thesis that are not included are

- Moskovitz*, Ted, **Arbel*, Michael**, Ferenc Huszar, and Arthur Gretton. Efficient wasserstein natural gradients for reinforcement learning. In *ICLR*, 2021

- Mikołaj Bińkowski*, Danica J. Sutherland*, **Arbel, Michael**, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018

  Code: https://github.com/mbinkowski/MMD-GAN

- Danica Sutherland, Heiko Strathmann, **Arbel, Michael**, and Arthur Gretton. Efficient and principled score estimation with Nystrom kernel exponential families. In *International Conference on Artificial Intelligence and Statistics*, pages 652–660, March 2018

  Code: https://github.com/karlnapf/nystrom-kexpfam

- Tolga Birdal, **Arbel, Michael**, Umut Simsekli, and Leonidas J Guibas. Synchronizing probability measures on rotations via optimal transport. In *In CVPR*, 2020

  Code: https://synchinvision.github.io/probsync

- Anna Korba, Adil Salim, **Arbel, Michael**, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for stein variational gradient descent. *In NeurIPS*, 2020

- Louis Thiry, **Arbel, Michael**, Eugene Belilovsky, and Edouard Oyallon. The unreasonable effectiveness of patches in deep convolutional kernels methods. In *ICLR*, 2021

- Samuel Cohen, **Arbel, Michael**, and Marc Peter Deisenroth. Estimating barycenters of measures in high dimensions. *Under review*, 2020

Chapter 2

Background

We introduce some concepts used throughout the rest of the thesis. We define $\mathcal{X} \subset \mathbb{R}^d$ as the closure of a convex open set, and $\mathcal{P}_2(\mathcal{X})$ as the set of probability distributions on $\mathcal{X}$ with finite second moment, equipped with the 2-Wassertein metric denoted W_2. For any $\nu \in \mathcal{P}_2(\mathcal{X})$, $L_2(\nu)$ is the set of square integrable functions w.r.t. ν.

1 Reproducing Kernel Hilbert Spaces

We recall here some fundamental definitions and properties of reproducing kernel Hilbert spaces (RKHS) (see Smola and Scholkopf [1998]). Given a positive semi-definite kernel $(x, y) \mapsto k(x, y) \in \mathbb{R}$ defined for all $x, y \in \mathcal{X}$, we denote by $\mathcal{H}$ its corresponding RKHS (see Smola and Scholkopf [1998]). The space $\mathcal{H}$ is a Hilbert space with inner product $\langle ., . \rangle_{\mathcal{H}}$ and corresponding norm $\|.\|_{\mathcal{H}}$. A key property of $\mathcal{H}$ is the reproducing property: for all $f \in \mathcal{H}, f(x) = \langle f, k(x, .) \rangle_{\mathcal{H}}$. Moreover, if k is m-times differentiable w.r.t. each of its coordinates, then any $f \in \mathcal{H}$ is m-times differentiable and $\partial^\alpha f(x) = \langle f, \partial^\alpha k(x, .) \rangle_{\mathcal{H}}$ where α is any multi-index with $\alpha \leq m$ [Steinwart and Christmann, 2008, Lemma 4.34]. When k has at most quadratic growth, then for all $\mu \in \mathcal{P}_2(\mathcal{X})$, $\int k(x, x) \, \mathrm{d}\mu(x) < \infty$. In that case, for any $\mu \in \mathcal{P}_2(\mathcal{X})$, $\phi_\mu := \int k(., x) \, \mathrm{d}\mu(x)$ is a well defined element in $\mathcal{H}$ called the mean embedding of μ. The kernel k is said to be characteristic when such mean embedding is injective, that is any mean embedding is associated to a unique probability distribution.

Maximum Mean Discrepancy When k is characteristic, it is possible to define a distance between distributions in $\mathcal{P}_2(\mathcal{X})$ called the Maximum Mean Discrepancy:

$$MMD(\mu, \nu) = \|\phi_\mu - \phi_\nu\|_\mathcal{H} \qquad \forall\ \mu, \nu \in \mathcal{P}_2(\mathcal{X}).$$

The difference between the mean embeddings of μ and ν is an element in $\mathcal{H}$ called the unnormalised witness function between μ and ν: $f_{\mu,\nu} = \phi_\nu - \phi_\mu$. The MMD can also be seen as an *Integral Probability Metric*:

$$MMD(\mu, \nu) = \sup_{g \in \mathcal{B}} \left\{ \int g \, \mathrm{d}\mu - \int g \, \mathrm{d}\nu \right\}$$

where $\mathcal{B} = \{g \in \mathcal{H} :\ \|g\|_\mathcal{H} \leq 1\}$ is the unit ball in the RKHS.

2 2-Wasserstein geometry

For two given probability distributions ν and μ in $\mathcal{P}_2(\mathcal{X})$, we denote by $\Pi(\nu, \mu)$ the set of possible couplings between ν and μ. In other words $\Pi(\nu, \mu)$ contains all possible distributions π on $\mathcal{X} \times \mathcal{X}$ such that if $(X, Y) \sim \pi$ then $X \sim \nu$ and $Y \sim \mu$. The 2-Wasserstein distance on $\mathcal{P}_2(\mathcal{X})$ is defined by means of an optimal coupling between ν and μ in the following way:

$$W_2^2(\nu, \mu) := \inf_{\pi \in \Pi(\nu,\mu)} \int \|x - y\|^2 \, \mathrm{d}\pi(x, y) \qquad \forall \nu, \mu \in \mathcal{P}_2(\mathcal{X})$$

It is a well established fact that such optimal coupling π^* exists Villani [2009], Santambrogio [2015] . Moreover, it can be used to define a path $(\rho_t)_{t \in [0,1]}$ between ν and μ in $\mathcal{P}_2(\mathcal{X})$. For a given time t in $[0, 1]$ and given a sample (x, y) from π^*, it is possible to construct a sample z_t from ρ_t by taking the convex combination of x and y: $z_t = s_t(x, y)$ where s_t is given by:

$$s_t(x, y) = (1 - t)x + ty \qquad \forall x, y \in \mathcal{X},\ \forall t \in [0, 1].$$

The function s_t is well defined since $\mathcal{X}$ is a convex set. More formally, ρ_t can be written as the projection or push-forward of the optimal coupling π^* by s_t:

$$\rho_t = (s_t)_{\#}\pi^* \tag{2.1}$$

We recall that for any $T : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$ a measurable map, and any $\rho \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$, the push-forward measure $T_{\#}\rho$ is characterized by:

$$\int_{y\in\mathcal{X}} \phi(y)\,\mathrm{d}T_{\#}\rho(y) = \int_{x,x'\in\mathcal{X}\times\mathcal{X}} \phi(T(x,x'))\,\mathrm{d}\rho(x,x'),$$

for every measurable and bounded function ϕ. It is easy to see that (2.1) satisfies the following boundary conditions at $t = 0, 1$:

$$\rho_0 = \nu \qquad \rho_1 = \mu. \tag{2.2}$$

Paths of the form of (2.1) are called *displacement geodesics*. They can be seen as the shortest paths from ν to μ in terms of mass transport (Santambrogio [2015] Theorem 5.27).

2 .1 Dynamical formulation

It can be shown that there exists a *velocity vector field* $(t, x) \mapsto \phi_t(x)$ with values in $\mathbb{R}^d$ such that ρ_t satisfies the continuity equation:

$$\partial_t\rho_t + div(\rho_t\phi_t) = 0 \qquad \forall t \in [0, 1]. \tag{2.3}$$

While there could be multiple vector fields ϕ_t satisfying (2.3), it is possible to consider only those that are in the closure in $L_2(\rho_t)$ of gradient of smooth compactly supported functions:

$$\phi_t \in \overline{\{\nabla f, \quad f \in \mathcal{C}_c^\infty\}}_{L_2(\rho_t)}. \tag{2.4}$$

This additional condition ensures uniqueness of the vector field ϕ_t [Ambrosio et al., 2004, Section 3.2].

Equation (2.3) expresses two facts, the first one is that $-div(\rho_t\phi_t)$ reflects the infinitesimal changes in ρ_t as dictated by the vector field (also referred to as velocity field) ϕ_t, the second one is that the total mass of ρ_t does not vary in time as a consequence of the divergence theorem. Equation (2.3) is well defined in the distribution sense even when ρ_t does not have a density. At each time t, ϕ_t can be interpreted as a tangent vector to the curve $(\rho_t)_{t\in[0,1]}$ so that the length $l((\rho_t)_{t\in[0,1]})$ of the curve $(\rho_t)_{t\in[0,1]}$ would be given by:

$$l((\rho_t)_{t\in[0,1]})^2 = \int_0^1 \|\phi_t\|^2_{L_2(\rho_t)}\,\mathrm{d}t \quad \text{where} \quad \|\phi_t\|^2_{L_2(\rho_t)} = \int \|\phi_t(x)\|^2\,\mathrm{d}\rho_t(x)$$

This perspective allows to provide a dynamical interpretation of the W_2 as the length of the shortest path from ν to μ and is summarized by the celebrated Benamou-Brenier formula (Benamou and Brenier [2000]):

$$W_2^2(\nu,\mu) = \inf_{\rho_t,\phi_t} \int_0^1 \int \|\phi_l(x)\|^2\,\mathrm{d}\rho_l(x)\,\mathrm{d}l \tag{2.5}$$

where the infimum is taken over all couples ρ and ϕ satisfying (2.3) with boundary conditions given by (2.2). If $(\rho_t,\phi_t)_{t\in[0,1]}$ satisfies (2.3) and (2.2) and realizes the infimum in (2.5), it is then simply called a geodesic between ν and μ; moreover it is called a constant-speed geodesic if, in addition, the norm of ϕ_t is constant for all $t\in[0,1]$. As a consequence, (2.1) is a constant-speed displacement geodesic.

2 .2 Pseudo-Riemannian structure

The formulation in (2.5) suggests that $W_2(\nu,\mu)$ corresponds in fact to the shortest path from ν to μ. Indeed, given a path ρ_l from μ to ν, the infinitesimal displacement direction is given by the distribution $\partial_l\rho_l$. The length $|\partial_l\rho_l|$ of this direction is measured by: $|\partial_l\rho_l|^2 := \int \|\phi_l(x)\|^2\,\mathrm{d}\rho_l(x)$. It is therefore possible to express $W_2^2(\nu,\mu)$ as:

$$W_2^2(\nu,\mu) = \inf_{\rho_l} \int_0^1 |\partial_l\rho_l|^2\,\mathrm{d}l.$$

In fact, $\partial_l\rho_l$ can be seen as an element in the tangent space $T_{\rho_l}\mathcal{P}_2$ to $\mathcal{P}_2$ at point ρ_l. To ensure that (2.3) is well defined, $T_{\rho_l}\mathcal{P}_2$ can be taken as the set of distributions σ

satisfying $\sigma(1) = 0$.

$$|\sigma(f)| \leq C\|\nabla f\|_{L_2(\rho)}, \qquad \forall f \in C_c^\infty(\Omega) \tag{2.6}$$

for some positive constant C. Indeed, the condition in (2.6) guarantees the existence of a vector field ϕ_σ in (2.4) that is a solution to the PDE: $\sigma = -div(\rho_l \phi_\sigma)$.

Moreover, $|\partial_l \rho_l|^2$ can be seen as an inner product of $\partial_l \rho_l$ with itself in $T_{\rho_l}\mathcal{P}_2$. This inner product defines in turn a metric tensor g^W on $\mathcal{P}_2$ called the Wasserstein metric tensor (see Otto and Villani [2000], Ambrosio et al. [2004]):

Definition 1. *The Wasserstein metric g^W is defined for all $\rho \in \mathcal{P}_2$ as the inner product over $T_\rho\mathcal{P}_2$ of the form:*

$$g_\rho^W(\sigma, \sigma') := \int \phi_\sigma(x)^\top \phi_{\sigma'}(x)\, \mathrm{d}\rho(x), \qquad \forall \sigma, \sigma' \in T_\rho\mathcal{P}_2$$

where ϕ_σ and $\phi_{\sigma'}$ are solutions to the partial differential equations:

$$\sigma = -div(\rho\phi_\sigma), \qquad \sigma' = -div(\rho\phi_{\sigma'}).$$

Moreover, ϕ_σ and $\phi_{\sigma'}$ are required to be in the closure of gradient of smooth and compactly supported functions w.r.t. $L_2(\rho)^d$.

Definition 1 allows to endow $\mathcal{P}_2$ with a pseudo-Riemannian[1] structure with W_2 being its geodesic distance:

$$W_2^2(\rho, \rho') = \inf_{\rho_l} \int_0^1 g_{\rho_l}(\partial_l \rho_l, \partial_l \rho_l)\, \mathrm{d}l,$$

where the infimum is taken over absolutely continuous paths $\rho_l : [0, 1] \to \mathcal{P}_2$ with boundary conditions $\rho_0 = \rho$ and $\rho_1 = \rho'$.

[1] We used the term pseudo since the geometric structure is not exactly Riemannian, as discussed extensively in [Ambrosio et al., 2004, Section 3.2]

2 .3 Gradient flows on the space of probability measures

Consider a real valued functional $\mathcal{F}$ defined over $\mathcal{P}_2(\mathcal{X})$. We call $\frac{\partial \mathcal{F}}{\partial \nu}$ if it exists, the unique (up to additive constants) function such that $\frac{d}{d\epsilon}\mathcal{F}(\nu + \epsilon(\nu' - \nu))|_{\epsilon=0} = \int \frac{\partial \mathcal{F}}{\partial \nu}(\nu)(\mathrm{d}\nu' - \mathrm{d}\nu)$ for any $\nu' \in \mathcal{P}_2(\mathcal{X})$. The function $\frac{\partial \mathcal{F}}{\partial \nu}$ is called the first variation of $\mathcal{F}$ evaluated at ν. We consider here functionals $\mathcal{F}$ of the form:

$$\mathcal{F}(\nu) = \int U(\nu(x))\nu(x)dx + \int V(x)\nu(x)dx + \int W(x,y)\nu(x)\nu(y)dxdy$$

where U is the internal potential, V an external potential and W an interaction potential. The formal gradient flow equation associated to such functional can be written (see Carrillo et al. [2006], Lemma 8 to 10):

$$\frac{\partial \nu}{\partial t} = div(\nu \nabla \frac{\partial \mathcal{F}}{\partial \nu}) = div(\nu \nabla (U'(\nu) + V + W * \nu)) \tag{2.7}$$

where div is the divergence operator and $\nabla \frac{\partial \mathcal{F}}{\partial \nu}$ is the strong subdifferential of $\mathcal{F}$ associated to the W_2 metric (see Ambrosio et al. [2008], Lemma 10.4.1). Indeed, for some generalized notion of gradient ∇_{W_2}, and for sufficiently regular ν and $\mathcal{F}$, the r.h.s. of (2.7) can be formally written as $-\nabla_{W_2}\mathcal{F}(\nu)$. The dissipation of energy along the flow is then given by:

$$\frac{d\mathcal{F}(\nu_t)}{dt} = -D(\nu_t) \quad \text{with } D(\nu) = \int \|\nabla \frac{\partial \mathcal{F}(\nu_t(x))}{\partial \nu}\|^2 \nu_t(x)dx \tag{2.8}$$

Such expression can be obtained by the following formal calculations:

$$\frac{d\mathcal{F}(\nu_t)}{dt} = \int \frac{\partial \mathcal{F}(\nu_t)}{\partial \nu_t}\frac{\partial \nu_t}{\partial t} = \int \frac{\partial \mathcal{F}(\nu_t)}{\partial \nu} div(\nu_t \nabla \frac{\partial \mathcal{F}(\nu_t)}{\partial \nu}) = -\int \|\nabla \frac{\partial \mathcal{F}(\nu_t)}{\partial \nu}\|^2 d\nu_t.$$

2 .4 Displacement convexity

Just as for Euclidian spaces, an important criterion to characterize the convergence of the Wasserstein gradient flow of a functional $\mathcal{F}$ is given by displacement convexity (see [Villani, 2004, Definition 16.5 (1st bullet point)])):

Definition 2. *[Displacement convexity] We say that a functional $\nu \mapsto \mathcal{F}(\nu)$ is*

displacement convex if for any ν and ν' and a constant speed geodesic $(\rho_t)_{t\in[0,1]}$ between ν and ν' with velocity vector field $(\phi_t)_{t\in[0,1]}$ as defined by (2.3), the following holds:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\nu_0) + t\mathcal{F}(\nu_1) \qquad \forall\, t \in [0,1].$$

Definition 2 can be relaxed to a more general notion of convexity called Λ-displacement convexity (see [Villani, 2009, Definition 16.5 (3rd bullet point)]). We first define an admissible functional Λ:

Definition 3. *[Admissible Λ functional] Consider a functional $(\rho, v) \mapsto \Lambda(\rho, v) \in \mathbb{R}$ defined for any probability distribution $\rho \in \mathcal{P}_2(\mathcal{X})$ and any square integrable vector field v w.r.t ρ. We say that Λ is admissible, if it satisfies:*

- *For any $\rho \in \mathcal{P}_2(\mathcal{X})$, $v \mapsto \Lambda(\rho, v)$ is a quadratic form.*
- *For any geodesic $(\rho_t)_{0\leq t\leq 1}$ between two distributions ν and ν' with corresponding vector fields $(\phi_t)_{t\in[0,1]}$ it holds that $\inf_{0\leq t\leq 1} \Lambda(\rho_t, \phi_t)/\|\phi_t\|^2_{L_2(\rho_t)} > -\infty$*

We can now define the notion of Λ-convexity:

Definition 4. *[Λ convexity] We say that a functional $\nu \mapsto \mathcal{F}(\nu)$ is Λ-convex if for any $\nu, \nu' \in \mathcal{P}_2(\mathcal{X})^2$ and a constant speed geodesic $(\rho_t)_{t\in[0,1]}$ between ν and ν' with velocity vector field $(\phi_t)_{t\in[0,1]}$ as defined by (2.3), the following holds:*

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\nu_0) + t\mathcal{F}(\nu_1) - \int_0^1 \Lambda(\rho_s, \phi_s) G(s,t)\, ds \qquad \forall\, t \in [0,1]. \quad (2.9)$$

where $(\rho, v) \mapsto \Lambda(\rho, v)$ satisfies Definition 3, and $G(s,t) = s(1-t)\mathbb{I}\{s \leq t\} + t(1-s)\mathbb{I}\{s \geq t\}$. A particular case is when $\Lambda(\rho, v) = \lambda \int \|v(x)\|^2 \,\mathrm{d}\rho(x)$ for some $\lambda \in \mathbb{R}$. In that case, (2.9) becomes:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\nu_0) + t\mathcal{F}(\nu_1) - \frac{\lambda}{2} t(1-t) W_2^2(\nu_0, \nu_1) \qquad \forall\, t \in [0,1]. \quad (2.10)$$

Definition 2 is a particular case of Definition 4, where in (2.10) one has $\lambda = 0$.

3 Fisher-Rao Statistical Manifold

In this section we briefly introduce the non-parametric Fisher-Rao metric defined over the set $\mathcal{P}$ of probability distributions with positive density. More details can be found in Holbrook et al. [2017]. By abuse of notation, an element $\rho \in \mathcal{P}$ will be identified with its density which will also be denoted by ρ. Consider $\mathcal{T}_\rho$, the set of real valued functions f defined over Ω and satisfying

$$\int \frac{f(x)^2}{\rho(x)}\,\mathrm{d}x < \infty; \qquad \int f(x)\rho(x)\,\mathrm{d}x = 0.$$

We have the following definition for the Fisher-Rao metric:

Definition 5 (Fisher-Rao metric)**.** *The Fisher-Rao metric g^F is defined for all $\rho \in \mathcal{P}$ as an inner product over $\mathcal{T}_\rho$ of the form:*

$$g^F_\rho(f,g) := \int \frac{1}{\rho(x)} f(x)g(x)\,\mathrm{d}x, \qquad \forall f,g \in \mathcal{T}_\rho$$

Note that the choice of the set $\mathcal{T}_\rho$ is different from the one considered in Holbrook et al. [2017] which replaces the integrability condition by a smoothness one. In fact, it can be shown that these choices result in the same metric by a density argument.

Part I

Conditional score estimation

Chapter 3

Conditional score estimation

A nonparametric family of conditional distributions is introduced, which generalizes conditional exponential families using functional parameters in a suitable RKHS. An algorithm is provided for learning the generalized natural parameter, and consistency of the estimator is established in the well specified case. In experiments, the new method generally outperforms a competing approach with consistency guarantees, and is competitive with a deep conditional density model on datasets that exhibit abrupt transitions and heteroscedasticity.

1 Introduction

Distribution estimation is one of the most general problems in machine learning. Once an estimator for a distribution is learned, in principle, it allows to solve a variety of problems such as classification, regression, matrix completion and other prediction tasks. With the increasing diversity and complexity of machine learning problems, regressing the conditional mean of y knowing x may not be sufficiently informative when the conditional density $p(y|x)$ is multimodal. In such cases, one would like to estimate the conditional distribution itself to get a richer characterization of the dependence between the two variables y and x. In this work, we address the problem of estimating conditional densities when x and y are continuous and multi-dimensional.

Our conditional density model builds on a generalisation of the exponential family to infinite dimensions [Gu and Qiu, 1993, Barron and Sheu, 1991, Pistone and

Sempi, 1995, Canu and Smola, 2006, Fukumizu, 2009], where the natural parameter is a function in a reproducing kernel Hilbert space (RKHS): in this sense, like the Gaussian and Dirichlet processes, the kernel exponential family (KEF) is an infinite dimensional analogue of the finite dimensional case, allowing to fit a much richer class of densities. While the maximum likelihood solution is ill-posed in infinite dimensions, [Sriperumbudur et al., 2017] have demonstrated that it is possible to fit the KEF via score matching [Hyvärinen, 2005], which entails solving a linear system of size $n \times d$, where n is the number of samples and d is the problem dimension. It is trivial to draw samples from such models using Hamiltonian Monte Carlo [Neal, 2010], since they directly return the required potential energy [Rasmussen, 2003, Strathmann et al., 2015]. In high dimensions, fitting a KEF model to samples becomes challenging, however: the computational cost rises as d^3, and complex interactions between dimensions can be difficult to model.

The complexity of the modelling task can be significantly reduced if a directed graphical model can be constructed over the variables, [Pearl, 2001, Jordan, 1999], where each variable depends on a subset of parent variables (ideally much smaller than the total, as in e.g. a Markov chain). In the present study, we extend the non-parametric family of Sriperumbudur et al. [2017] to fit conditional distributions. The natural parameter of the *conditional* infinite exponential family is now an operator mapping the conditioning variable to a function space of features of the conditioned variable: for this reason, the score matching framework must be generalised to the vector-valued kernel regression setting of Micchelli and Pontil [2005]. We establish consistency in the well specified case by generalising the arguments of Sriperumbudur et al. [2017] to the vector-valued RKHS. While our proof allows for general vector-valued RKHSs, we provide a practical implementation for a specific case, which takes the form of a linear system of size $n \times d$.

A number of alternative approaches have been proposed to the problem of conditional density estimation. Sugiyama et al. [2010] introduced the Least-Square Conditional Density Estimation (LS-CDE) method, which provides an estimate of a conditional density function $p(y|x)$ as a non-negative linear combination of basis

functions. The method is proven to be consistent, and works well on reasonably complicated learning problems, although the optimal choice of basis functions for the method is an open question (in their paper, the authors use Gaussians centered on the samples). Earlier non-parametric methods such as variants of Kernel Density Estimation (KDE) may also be used in conditional density estimation [Fan et al., 1996, Hall et al., 1999]. These approaches also have consistency guarantees, however their performance degrades in high-dimensional settings [Nagler and Czado, 2016]. Sugiyama et al. [2010] found that kernel density approaches performed less well in practice than LS-CDE.

It is possible to represent and learn conditional probabilities without specifying probability densities, via conditional mean embeddings [Song et al., 2010, Grunewalder et al., 2012]. These are conditional expectations of (potentially infinitely many) features in an RKHS, which can be used in obtaining conditional expectations of functions in this RKHS. Such expected features are complementary to the infinite dimensional exponential family, as they can be thought of as conditional expectations of an infinite dimensional sufficient statistic. This statistic can completely characterise the conditional distribution if the feature space is sufficiently rich [Sriperumbudur et al., 2010], and has consistency guarantees under appropriate smoothness assumptions. Drawing samples given a conditional mean embedding can be challenging: this is possible via the Herding procedure [Chen et al., 2010, Bach et al., 2012], as shown in [Kanagawa et al., 2016], but requires a non-convex optimisation procedure to be solved for each sample.

A powerful and recent deep learning approach to modelling conditional densities is the Neural Autoregressive Network (Uria et al. [2013], Raiko et al. [2014] and Uria et al. [2016]). These networks can be thought of as a generalization of the Mixture Density Network introduced by Bishop [2006]. In brief, each variable is represented by a mixture of Gaussians, with means and variances depending on the parent variables through a deep neural network. The network is trained on observed data using stochastic gradient descent. Neural autoregressive networks have shown their effectiveness for a variety of practical cases and learning problems. Unlike the

earlier methods cited, however, consistency is not guaranteed, and these methods require non-convex optimization, meaning that locally optimal solutions are found in practice.

We begin our presentation in Section 2 , where we briefly present the Kernel Exponential Family. We generalise this model to the conditional case, in our first major contribution: this requires the introduction of vector-valued RKHSs and associated concepts. We then show that a generalisation of score matching may be used to fit the conditional density models for general vector valued RKHS, subject to appropriate conditions. We call this model the kernel conditional exponential family (KCEF).

Our second contribution, in Section 3 , is an empirical estimator for the *natural parameter* of the KCEF (Theorem 1), with convergence guarantees in the well-specified case (Theorem 2). In our experiments (Section 4), we empirically validate the consistency of the estimator and compare it to other methods of conditional density estimation. Our method generally outperforms the leading alternative with consistency guarantees (LS-CDE). Compared with the deep approach (RNADE) which lacks consistency guarantees, our method has a clear advantage at small training sample sizes while being competitive at larger training sizes.

2 Kernel Exponential Families

In this section we first present the kernel exponential family, which we then extend to a class of conditional exponential families. Finally, we provide a methodology for unnormalized density estimation within this class.

2 .1 Kernel Conditional Exponential Family

We consider the task of estimating the density $p(y)$ of a random variable Y with support $\mathcal{Y} \subseteq \mathbb{R}^d$ from i.i.d samples $(Y_i)_{i=1}^n$. We propose to use a family of densities parametrized by functions belonging to a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_\mathcal{Y}$ with positive semi-definite kernel k [Canu and Smola, 2006, Fukumizu, 2009, Sriperumbudur et al., 2017]. This exponential family of density functions takes the

form

$$\left\{ p(y) := q_0(y) \frac{\exp \langle f, k(y, .) \rangle_{\mathcal{H}_\mathcal{Y}}}{Z(f)} \middle| f \in \mathcal{F} \right\}, \tag{3.1}$$

where q_0 is a base density function on $\mathcal{Y}$ and $\mathcal{F}$ is the set of functions in the RKHS space $\mathcal{H}_\mathcal{Y}$ such that $Z(f) := \int_\mathcal{Y} \exp \langle f, k(y, .) \rangle_{\mathcal{H}_\mathcal{Y}} q_0(y) dx < \infty$. In what follows, we call this family the *kernel exponential family* (KEF) by analogy to classical exponential family. f plays the role of the natural parameter while $k(y, .)$ is the sufficient statistic. Note that with an appropriate choice of the base distribution q_0 and a finite dimensional RKHS $\mathcal{H}_\mathcal{Y}$, one can recover any finite dimensional exponential family. When $\mathcal{H}_\mathcal{Y}$ is infinite-dimensional, however, the family can approximate a much broader class of densities on $\mathbb{R}^d$: under mild conditions, it is shown in Sriperumbudur et al. [2017] that the KEF approximates all densities of the form $\{q_0(y) \exp(f(y) - A) | f \in C_0(\mathcal{Y})\}$, A being the normalizing constant and $C_0(\mathcal{Y})$ the set of continuous functions with vanishing tails.

Given two subsets $\mathcal{Y}$ and $\mathcal{X}$ of $\mathbb{R}^d$ and $\mathbb{R}^p$ respectively, we now propose to extend the KEF to a family of conditional densities $p(y|x)$. We modify equation (3.1) by making the function f depend on the conditioning variable x. The parameter f is a function of two variables x and y, $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that $y \mapsto f(x, y)$ belongs to the RKHS $\mathcal{H}_\mathcal{Y}$ for all x in $\mathcal{X}$. In all that follows, we will denote by T the mapping

$$T : \mathcal{X} \to \mathcal{H}_\mathcal{Y} \qquad x \mapsto T_x$$

such that $T_x(y) = f(x, y)$ for all y in $\mathcal{Y}$

We next consider how to enforce a smoothness requirement on T to make the conditional density estimation problem well-posed. To achieve this, we will require that the mapping T belongs to a vector valued RKHS $\mathcal{H}$: we now briefly review the associated theory, following [Micchelli and Pontil, 2005]. A Hilbert space $(\mathcal{H}, \langle ., . \rangle_\mathcal{H})$ of functions $T : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ taking values in a vector space $\mathcal{H}_\mathcal{Y}$ is said to be a vector valued RKHS if for all $x \in \mathcal{X}$ and $h \in \mathcal{H}_\mathcal{Y}$, the linear functional $T \mapsto \langle h, T_x \rangle_{\mathcal{H}_\mathcal{Y}}$ is continuous. The reproducing property for vector-valued RKHSs

follows from this definition. By the Riesz representation theorem, for each $x \in \mathcal{X}$ and $h \in \mathcal{H}_\mathcal{Y}$, there exists a linear operator Γ_x from $\mathcal{H}_\mathcal{Y}$ to $\mathcal{H}$ such that for all $T \in \mathcal{H}$,

$$\langle h, T_x \rangle_{\mathcal{H}_\mathcal{Y}} = \langle T, \Gamma_x h \rangle_\mathcal{H} \tag{3.2}$$

Considering the dual operator Γ_x^* from $\mathcal{H}$ to $\mathcal{H}_\mathcal{Y}$, we also get

$$\Gamma_x^* T = T_x.$$

We can define a vector-valued reproducing kernel by composing the operator Γ_x with its dual,

$$\Gamma(x, x') = \Gamma_x^* \Gamma_{x'},$$

where for all x and $x^{'}$, $\Gamma(x, x^{'})$ is a bounded linear operator from $\mathcal{H}_\mathcal{Y}$ to $\mathcal{H}_\mathcal{Y}$, i.e., $\Gamma(x, x^{'}) \in \mathcal{L}(\mathcal{H}_\mathcal{Y})$. The space $\mathcal{H}$ is said to be generated by an operator valued reproducing kernel Γ. One practical choice for Γ is to define it as:

$$\Gamma(x, x^{'}) = k_\mathcal{X}(x, x^{'}) I_{\mathcal{H}_\mathcal{Y}} \qquad \forall x, x^{'} \in \mathcal{X}, \tag{3.3}$$

where $I_{\mathcal{H}_\mathcal{Y}}$ the identity operator on $\mathcal{H}_\mathcal{Y}$ and $k_\mathcal{X}$ is now a real-valued kernel which generates a real valued RKHS $\mathcal{H}_\mathcal{X}$ on $\mathcal{X}$ [as in the conditional mean embedding; see Grunewalder et al., 2012]. A simplified form of the estimator of T will be presented in Section 3 for this particular choice for Γ and will be used in the experimental setup in Section 4 .

We will now express $T_x(y)$ in a convenient form that will allow to extend the KEF. For a given x, recalling that T_x belongs to $\mathcal{H}_\mathcal{Y}$, one can write $T_x(y) = \langle T_x, k(y, .) \rangle_{\mathcal{H}_\mathcal{Y}}$ for all y in $\mathcal{Y}$. Using the reproducing property in (3.2), one further gets $T_x(y) = \langle T, \Gamma_x k(y, .) \rangle_\mathcal{H}$. By considering the subset $\mathcal{T}$ of elements T in $\mathcal{H}$ such

that for all x in $\mathcal{X}$ the integral

$$Z(T_x) := \int_{\mathcal{Y}} q_0(y) \exp(\langle T, \Gamma_x k(y, .)\rangle_{\mathcal{H}}) \mathrm{d}y < \infty$$

is finite, we define the *kernel conditional exponential family* (KCEF) as the set of conditional densities

$$\left\{ p_T(y|x) := q_0(y) \frac{\exp \langle T, \Gamma_x k(y, .)\rangle_{\mathcal{H}}}{Z(T_x)} \middle| T \in \mathcal{T} \right\}. \tag{3.4}$$

Here T plays the role of the natural parameter while $\Gamma_x k(y, .)$ is the sufficient statistic. When T is restricted to be constant with respect to x, we recover the *kernel exponential family* (KEF). The KCEF is therefore an extension of the KEF introduced in Sriperumbudur et al. [2017]. It is also a special case of the family introduced in Canu and Smola [2006]. In the latter, the inner product is given by $\langle T, \phi(x, y)\rangle_{\mathcal{H}}$ where ϕ is a general feature of x and y. In the present work, ϕ has the particular form $\phi(x, y) = \Gamma_x k(y, .)$. This allows to further express $p_T(y|x)$ for a given x as an element in a KEF with sufficient statistic $k(y, .)$, by using the identity $\langle T, \Gamma_x k(y, .)\rangle_{\mathcal{H}} = \langle T_x, k(y, .)\rangle_{\mathcal{H}_{\mathcal{Y}}}$. This is desirable since $p_T(y|x)$ remains in the same KEF family as x varies and only the natural parameter T_x changes.

2 .2 Unnormalized density estimation

Given i.i.d samples $(X_i, Y_i)_{i=1}^n$ in $\mathcal{X} \times \mathcal{Y}$ following a joint distribution $\pi(x)p_0(y|x)$, where π defines a marginal distribution over $\mathcal{X}$ and $p_0(y|x)$ is a conditional density function, we are interested in estimating p_0 from the samples $(X_i, Y_i)_{i=1}^n$. Our goal is to find the optimal conditional density p_T in the KCEF that best approximates p_0. The intractability of the normalizing constant $Z(T_x)$ makes maximum likelihood estimation difficult. Sriperumbudur et al. [2017] used a score-matching approach (see Hyvärinen [2005]) to avoid this normalizing constant; in the case of the KCEF, however, the score function between $\pi(x)p_0(y|x)$ and $\pi(x)p_T(y|x)$ contains additional terms that involve the derivatives of the log-partition function $\log Z(T_x)$ with respect to x. Instead, we now propose a different approach with a modified version of the score-matching objective.

We define the expected conditional score between two conditional densities $p_0(y|x)$ and $q(y|x)$ under a marginal density π on x to be:

$$J(p_0|q) := \int_{\mathcal{X}} \pi(x)\mathcal{J}(p_0(.|x)||q(.|x))dx$$

where:

$$\mathcal{J}(p_0(.|x)\|q(.|x)) = \frac{1}{2}\int_{\mathcal{Y}} p_0(y|x)\left\|\nabla_y \log \frac{p_0(y|x)}{q(y|x)}\right\|^2 dy.$$

For a fixed value x in $\mathcal{X}$, $\mathcal{J}(p_0(.|x)||q(.|x))$ is the score-matching function between $p_0(.|x)$ and $q(.|x)$ as defined by Hyvärinen [2005]. We further take the expectation over x to define a divergence over conditional densities. The normalizing constant of $q(y|x)$, which is a function of x, is never involved in this formulation, as we take the gradient of the log-densities over y only. For a conditional density $p_0(y|x)$ that is supported on the whole domain $\mathcal{Y}$ for all x in $\mathcal{X}$, the expected conditional score is well behaved in the sense that $J(p_0|q)$ is always non-negative, and reaches 0 if and only if the two conditional distributions $p_0(y|x)$ and $q(y|x)$ are equal for π-almost all x. The goal is then to find a conditional distribution p_T in the KCEF for a given $T \in \mathcal{T}$ that minimizes this score over the whole family.

Under mild regularity conditions on the densities [see Hyvärinen, 2005, Sriperumbudur et al., 2017, and below], the score can be rewritten

$$\begin{aligned} J(p_0||p_T) =& \mathbb{E}\Big[\sum_{i=1}^{d} \partial_i^2 T_x(y) + \frac{1}{2}(\partial_i T_x(y))^2\Big] \\ &+ \mathbb{E}\Big[\sum_{i=1}^{d} \partial_i T_x(y)\partial_i \log q_0(y)\Big] + J(p_0||q_0) \end{aligned}$$

where $J(p_0||q_0)$ is a constant term for the optimization problem and the expectation is taken over $\pi(x)p_0(y|x)$. All derivatives are with respect to y, and we used the notation $\partial_i f(y) = \frac{\partial}{\partial y_i} f(y)$. In the case of KCEF, conditions to obtain this expression are satisfied under assumptions in Section A .4, as proved in Theorem 3 of Section B

.1 . The expression is further simplified using the reproducing property for the derivatives of functions in an RKHS (Lemma 11 of Section C),

$$\partial_i T_x(y) = \langle T, \Gamma_x \partial_i k(y, .)\rangle_{\mathcal{H}}$$
$$\partial_i^2 T_x(y) = \langle T, \Gamma_x \partial_i^2 k(y, .)\rangle_{\mathcal{H}}$$

which leads to:

$$J(T) = \mathbb{E}\Big[\sum_{i=1}^{d} \frac{1}{2}\big\langle T, \Gamma_x \partial_i k(y, .)\big\rangle_{\mathcal{H}}^2 + \big\langle T, \Xi_i(x, y)\big\rangle_{\mathcal{H}}\Big]$$

with:

$$\Xi_i(x, y) = \Gamma_x(\partial_i^2 k(y, .) + \partial_i \log q_0(y)\partial_i k(y, .)). \tag{3.5}$$

We introduced the notation $J(T) := J(p_0||p_T) - J(p_0||q_0)$ for convenience. This formulation depends on $p_0(y|x)$ only through an expectation, therefore a Monte Carlo estimator of the score can be derived as a quadratic functional of T in the RKHS $\mathcal{H}$,

$$\hat{J}(T) = \frac{1}{n}\sum_{\substack{b\in[n]\\ i\in[d]}} \frac{1}{2}\big\langle T, \Gamma_{X_b} \partial_i k(Y_b, .)\big\rangle_{\mathcal{H}}^2 + \langle T, \Xi_i(X_b, Y_b)\big\rangle_{\mathcal{H}}.$$

Note that the objective functions $J(T)$ and $\hat{J}(T)$ can be defined over the whole space $\mathcal{H}$, whereas $J(p_0|p_T)$ is meaningful only if T belongs to $\mathcal{T}$.

3 Empirical KCEF and consistency

In this section, we will first estimate the optimal $T^* = \operatorname{argmin}_{T\in\mathcal{H}} J(T)$ over the whole space $\mathcal{H}$ by minimizing a regularized version of the quadratic form in equation $\hat{J}(T)$, then we will state conditions under which all of the obtained solutions belong to $\mathcal{T}$ defining therefore conditional densities in the KCEF.

Following Sriperumbudur et al. [2017], we define the kernel ridge estimator to be $T_{n,\lambda} = \operatorname{argmin}_{T\in\mathcal{H}} \hat{J}(T) + \frac{\lambda}{2}\|T\|_{\mathcal{H}}^2$ where $\|T\|_{\mathcal{H}}$ is the RKHS norm of T. $T_{n,\lambda}$

is then obtained by solving a linear system of nd variables as shown in the next theorem:

Theorem 1. *Under assumptions listed in Section A .4, and in particular if* $\|\Gamma(x,x)\|_{Op}$ *is uniformly bounded on* $\mathcal{X}$ *for the operator norm, then the minimizer* $T_{n,\lambda}$ *exists, is unique, and is given by*

$$T_{n,\lambda} = -\frac{1}{\lambda}\hat{\Xi} + \sum_{b\in[n];i\in[d]} \beta_{(b,i)}\Gamma_{X_b}\partial_i k(Y_b,\cdot),$$

where

$$\hat{\Xi} = \frac{1}{n}\sum_{b\in[n];i\in[d]} \Xi_i(X_b,Y_b),$$

and Ξ_i *are given by (3.5).* $\beta_{(b,i)}$ *denotes the* $(b-1)d+i$ *entry of a vector* β *in* $\mathbb{R}^{nd}$*, obtained by solving the linear system*

$$(G+n\lambda I)\beta = \frac{h}{\lambda},$$

where G *is an* nd *by* nd *Gram matrix, and* h *is a vector in* $\mathbb{R}^{nd}$*,*

$$\begin{aligned}(G)_{(a,i),(b,j)} &= \langle \Gamma_{X_a}\partial_i k(Y_a,\cdot), \Gamma_{X_b}\partial_j k(Y_b,\cdot)\rangle_{\mathcal{H}}\\ (h)_{(b,i)} &= \langle \hat{\Xi}, \Gamma_{X_b}\partial_i k(Y_a,\cdot)\rangle_{\mathcal{H}}.\end{aligned}$$

The result is proved in Theorem 4 of Section B .2.

For the particular choice of Γ in (3.3), the estimator takes a simplified form

$$T_{n,\lambda}(x,y) = -\frac{1}{\lambda}\hat{\Xi}(x,y) + \sum_{\substack{b\in[n]\\ i\in[d]}} \beta_{(b,i)}k_{\mathcal{X}}(X_b,x)\partial_i k(Y_b,y),$$

with

$$\begin{aligned}\hat{\Xi}(x,y) =& \frac{1}{n}\sum_{b\in[n];i\in[d]} k_{\mathcal{X}}(X_b,x)\partial_i^2 k(Y_b,y) \\ &+ \frac{1}{n}\sum_{b\in[n];i\in[d]} k_{\mathcal{X}}(X_b,x)\partial_i \log q_0(Y_b)\partial_i k(Y_b,y)\end{aligned}$$

The coefficients β are obtained by solving the same system $(G+n\lambda I)\beta = \frac{h}{\lambda}$, where G and h reduce to

$$\begin{aligned}(G)_{(a,i),(b,j)} =& k_{\mathcal{X}}(X_a,X_b)\partial_i\partial_{j+d}k(Y_a,Y_b), \\ (h)_{(b,i)} =& \partial_i\hat{\Xi}(X_b,Y_b),\end{aligned}$$

and all derivatives are taken with respect to y.

The above estimator generalizes the estimator in Sriperumbudur et al. [2017] to conditional densities. In fact, if one choses the kernel $k_{\mathcal{X}}$ to be a constant kernel, then one exactly recovers the setting of the KEF.

This linear system has a complexity of $\mathcal{O}(n^3d^3)$ in time and $\mathcal{O}(n^2d^2)$ in memory, which can be problematic for higher dimensions d as n grows. However, in practice, if the goal is to estimate a density of the form $p(x_1,...,x_d)$, one can use the general chain rule for distributions, $p(x_1)p(x_2|x_1)....p(x_d|x_1,...,x_{d-1})$, and estimate each conditional density $p(x_i|x_1,...,x_{i-1})$ using the KCEF in (3.4). While this strategy requires the stonger assumption that each conditional density to be in a KCEF, this reduces the complexity of the algorithm to $\mathcal{O}(n^3d)$. A reduction to the cubic complexity in the number of data points n could be managed via a Nyström-like approximation [Sutherland et al., 2018].

In the well-specified case where the true conditional density $p_0(y|x)$ is assumed to be in (3.4) (i.e. $p_0(y|x) = p_{T_0}(y|x)$), we analyze the parameter convergence of the estimator $T_{n,\lambda}$ to T_0 and the convergence of the corresponding density $p_{T_{n,\lambda}}(y|x)$ to the true density $p_0(y|x)$. First, we consider the covariance operator C of the joint feature $\Gamma_x k(y,\cdot)$ under the joint distribution of x and y, as introduced in Theorem 3 of Section B .1, and we denote by $\mathcal{R}(C^\gamma)$ the range space of the operator C^γ. We

then have the following consistency result:

Theorem 2. *Let $\gamma > 0$ be a positive parameter and define $\alpha = \max(\frac{1}{2(\gamma+1)}, \frac{1}{4}) \in (\frac{1}{4}, \frac{1}{2})$. Under the conditions in Section A .4, for $\lambda = n^{-\alpha}$, and if $T_0 \in \mathcal{R}(C^\gamma)$, then*

$$\|T_{n,\lambda} - T_0\| = \mathcal{O}_{p_0}(n^{-\frac{1}{2}+\alpha}).$$

Furthermore, if $\sup_{y\in\mathcal{Y}} k(y,y) < \infty$, then

$$KL(p_0||p_{T_{n,\lambda}}) = \mathcal{O}_{p_0}(n^{-1+2\alpha}).$$

These asymptotic rates match those obtained for the unconditional density estimator in Sriperumbudur et al. [2017]. The smoother the parameter T_0, the closer α gets to $\frac{1}{4}$, which in turns leads to a convergence rate in KL divergence of the order of $\frac{1}{\sqrt{n}}$. The worst case scenario happens when the range-space parameter γ gets closer to 0, in which case convergence in KL divergence happens at a rate close to $\frac{1}{n^\gamma}$. A more technical formulation of this theorem along with a proof is presented in Section B .3 (see Theorems 5 and 6).

The regularity of the conditional density $p(y|x)$ with respect to x is captured by the boundedness assumption on the operator valued kernel Γ; i.e., $\|\Gamma(x,x)\|_{op} \leq \kappa$ for all $x \in \mathcal{X}$ in Assumption **(E)**. This assumption allows to control the variations of the conditional distribution $p(y|x)$ as x changes. Roughly speaking, we may estimate the conditional density $p(y|x_0)$ at a given point x_0 from samples (Y_i, X_i) whenever there are X_i sufficiently close to x_0. The uniformly bounded kernel Γ allows to express the objective function $J(T)$ as a quadratic form $J(T) = \frac{1}{2}\langle T, CT\rangle_{\mathcal{H}} + \langle T, \Xi\rangle_{\mathcal{H}} + c_0$ for constant c_0, where C is the covariance operator introduced in Theorem 3. Furthermore, this boundedness assumption ensures that C is a "well-behaved" operator, namely a positive semi-definite trace-class operator. The population solution of the regularized score objective is then given by $T_\lambda = (C + \lambda I)^{-1} C T_0$ while the estimator is given by: $\hat{T}_{\lambda,n} = -(\hat{C} + \lambda I)^{-1}\hat{\Xi}$ where $\hat{C}$ and $\hat{\Xi}$ are empirical estimators for C and Ξ.

The proof of consistency makes use of ideas from Sriperumbudur et al. [2017],

Caponnetto and De Vito [2007], exploiting the properties of trace-class operators. The main idea is to first control the error $\|T_0 - \hat{T}_{\lambda,n}\|_{\mathcal{H}}$ by introducing the population solution T_λ,

$$\|T_0 - \hat{T}_{\lambda,n}\|_{\mathcal{H}} \leq \|T_0 - T_\lambda\|_{\mathcal{H}} + \|T_\lambda - \hat{T}_{\lambda,n}\|_{\mathcal{H}}$$

The first term $\|T_0 - T_\lambda\|_{\mathcal{H}}$ represents the regularization error which is introduced by adding a regularization term λI to the operator C. This term doesn't depend on n, and can be shown to decrease as the amount of regularization goes to 0 with a rate $\lambda^{min(1,\gamma)}$. The second term represents the estimation error due to the finite number of samples n. This term decreases as $n \to 0$ but also increases when $\lambda \to 0$, therefore a trade-off needs to be made between decreasing the first term $\|T_0 - T_\lambda\|_{\mathcal{H}}$ by setting $\lambda \to 0$ and keeping the term $\|T_\lambda - \hat{T}_{\lambda,n}\|_{\mathcal{H}}$ small enough. Using decompositions similar to those of Sriperumbudur et al. [2017], Caponnetto and De Vito [2007], we apply concentration inequalities on the general Hilbert space $\mathcal{H}$ to get a probabilistic bound on the estimation error of order $\mathcal{O}(\frac{1}{\lambda\sqrt{n}})$.

Concerning the convergence in KL divergence, the requirement that the real-valued kernel k is bounded implies that $\mathcal{T}$ is in fact equal to $\mathcal{H}$. Therefore, minimizing the expected score $J(p_{T_0}||p_T)$ is equivalent to minimizing the quadratic form $J(T)$ over the whole RKHS $\mathcal{H}$. Finally, the rates in KL divergence are obtained from the error rate of $\hat{T}_{\lambda,n}$.

4 Experiments

We perform a diverse set of experiments, on both synthetic and real data, in order to validate our model empirically. In all experiments, the data are centered and rescaled such that the standard deviation for every dimension is equal to 1. Given $(X_1^{(n)}, ..., X_d^{(n)})_{n=1}^N$ i.i.d. samples of dimension d we are interested in approximating the joint distribution $p_0(X_1, ..., X_d)$ of data using different methods:

• The **KEF** model from Sriperumbudur et al. [2017] approximates p by a distribution p_f that belongs to the KEF (3.1) by minimizing the score loss between p and p_f to find the optimal parameter f.

• The **KCEF** model of Theorem 1 approximates p by a distribution $\hat{p}$ that is assumed to factorize according to some Directed Acyclic Graphical model (DAG): $\hat{p}(X_1, ..., X_d) = \Pi_{i=1}^{d} \hat{p}(x_i | x_{\pi(i)})$ where $\pi(i)$ are the parent nodes of i. Note that we do not necessarily make independence assumptions, as the graph can be fully connected. We will consider in particular two graphs, the Full graph (**F**) of the form $\hat{p}(X_1, ..., X_d) = \hat{p}(X_1)\Pi_{i=2}^{d} \hat{p}(X_i | X_1, ..., X_{i-1})$ and the Markov graph (**M**) of the form $\hat{p}(X_1, ..., X_d) = \hat{p}(X_1)\Pi_{i=2}^{d} \hat{p}(X_i | X_{i-1})$. Each of the factors is assumed to belong to the KCEF in (3.4), and is estimated independently from the others by minimizing the empirical loss $\hat{J}(T)$ to find the optimal operator T_i such that $\hat{p}(X_i | X_{\pi(i)}) = p_{T_i}(X_i | X_{\pi(i)})$.

• The **Orderless RNADE** model in Uria et al. [2016], where we train a 2 Layer Neural Autoregressive model with 100 units per layer. The model consists of a product of conditional densities of the form $\Pi_{i=1}^{d} p(X_{o_i} | X_{o_{<i}}, \theta, o)$, where o is a permutation of the dimensions $[1, ..., d]$ and θ is a set of parameters that are shared across the factors regardless of the chosen permutation o. RNADE is trained by minimizing the empirical expected negative log-likelihood, where the expectation is taken over all possible permutations and data,

$$\mathcal{L}(\theta) = \mathbb{E}_{o \in D} \mathbb{E}_{X \in \mathbb{R}^d} \Big[- \log p(X_{o_i} | X_{o_{<i}}, \theta, o) \Big].$$

• The **LSCDE** model in Sugiyama et al. [2010] where we also used the 2 factorizations of the joint distribution (**F**, **M**) and solve a least-squares problem to estimate each of the conditional densities. The approximate densities are of the form $\alpha^T \phi(X_i, X_{\pi(i)})$ where ϕ is a vector of m known non-negative functions and α is obtained by minimizing the squared error between $p(X_i, X_{\pi(i)})$ and $\alpha^T \phi(X_i, X_{\pi(i)})$. Only the non-negative component of the solution α is used.

For all variants of our model, we take the base density q_0 to be a centered gaussian with a standard deviation of 2. The kernel function used for both predicted variable y and conditioning variable x is the anisotropic radial basis function (RBF) with per-dimension bandwidths. The bandwidths and the regularization parameter λ are tuned by gradient descent on the cross validated score.

4 .1 Synthetic data

We consider the 'grid' dataset, which is a d-dimensional distribution with a tractable density that factorizes in the form

$$p(x_i|x_{i-1}) = C_i(1 + \sin(2\pi w_i^a x_i)\sin(2\pi w_i^b x_{i-1}))$$

for all $i \in [d]$. C_i is a tractable normalizing constant. Samples are generated using rejection sampling for each dimension. To study the effect of sample size on the estimator, we generate n training points with n varying from 200 to 2000 and $d = 3$, and estimate the log-likelihood on 2000 newly generated points. To compare the effect of dimension, we generate 2000 datapoints of dimension d varying from 2 to 20, and estimate the log-likelihood on 2000 test points. Unlike in [Sriperumbudur et al., 2017, Sutherland et al., 2018], the score function $\hat{J}(T)$ cannot be used as a metric to compare different factorizations of the estimated distribution, as it is dependent on the specific factorization of the joint distribution. Instead, we estimated the log-likelihood for our proposed model **KCEF**, where the normalizing constants are computed using importance sampling. We discarded the **KEF** in this experiment, since estimating the normalizing constant in high dimensions becomes impractical.

In Figure 3.1(left), we plot the log-likelihood as the number of samples increases. Both variants of **KCEF** (**F**, **M**) performed slightly better than the other methods in terms of speed of convergence as sample size increases. The variants that exploit the Markov structure of data **M** lead to the best performance for both **KCEF** and **LSCDE** as expected. The **NADE** method has comparable performance for large sample sizes, but the performance drops significantly for small sample sizes. This behaviour will also be observed in subsequent experiments on real data. The figure on the right shows the evolution of the log-likelihood per dimension as the dimension increases. In the **F** case, our approach is comparable to **LSCDE** with an advantage in small dimensions. The **F** approaches both use an anisotropic RBF kernel with tuned per-dimension bandwidth which end up performing a kind of automatic relevance determination. This helps getting comparable performance to the **M** methods. A drastic drop in performance can happen when an isotropic kernel is used instead

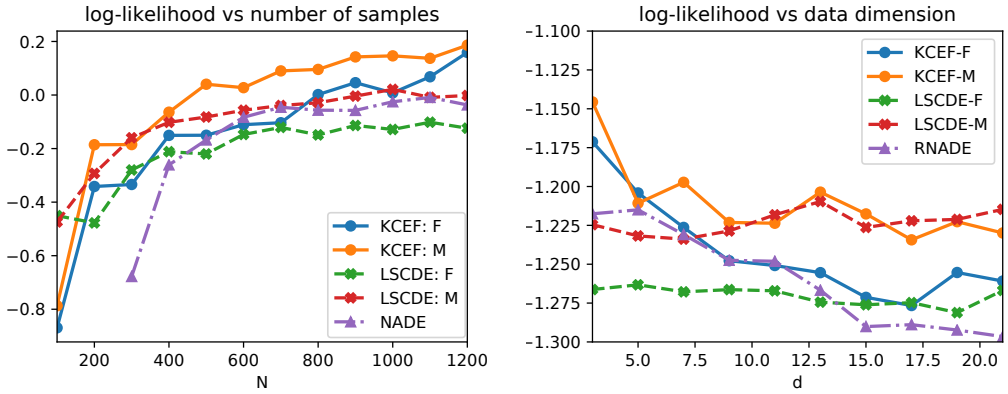


Figure 3.1: Experimental comparison of proposed method KCEF and other methods (LSCDE and NADE) on synthetic *grid* dataset. **LEFT**: log-likelihood vs training samples size, $(d = 3)$. **RIGHT**: log-likelihood per dimension vs dimension, $N = 2000$. The log-likelihood is evaluated on a separate test set of size 2000.

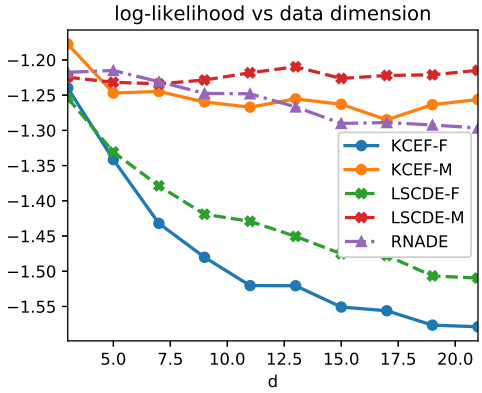


Figure 3.2: Experimental comparison of proposed method KCEF using an isotropic kernel and other methods (LSCDE and NADE) on synthetic *grid* dataset. Log-likelihood per dimension vs dimension, $N = 2000$. The log-likelihood is evaluated on a separate test set of size 2000.

as confirmed by Figure 3.2. Finally, **NADE**, which is also agnostic to the Markov structure of data, seems to achieve comparable performance to the **F** methods with a slight disadvantage in higher dimensions.

4 .2 Real data

We applied the proposed and existing methods to the *R* package benchmark datasets [Team, 2008] (see Table 3.1) as well as three UCI datasets previously used to study the performance of other density estimators (see Uria et al. [2013], Sil [2011], Tang et al. [2012]). In all cases data are centered and normalized.

First, the *R* benchmark datasets are low dimensional with few samples, but with a relatively complex conditional dependence between the variables. This setting allows to compare the methods in terms of data efficiency and overfitting. Each dataset was randomly split into a training and a test set of equal size. The models are trained to estimate the conditional density of a one dimensional variable y knowing x using samples $(x_i, y_i)_{i=1}^n$ form the training set. The accuracy is measured by the negative log-likelihood for the test samples $(\widetilde{x}_i, \widetilde{y}_i)_{i=1}^n$ averaged over 20 random splits of data. We compared the proposed method with **NADE** and **LSCDE** on 14 datasets. For **NADE** we used CV over the number of units per layer $\{2, 10, 100\}$ and number of mixture components $\{1, 2, 5, 10\}$ for a 2 layer network. We also used CV to chose the hyper-parameters for **LSCDE** and the proposed method on a 20×20 grid (for λ and σ).

The experimental results are summmarized in Table 3.1. **LSCDE** worked well in general as claimed in the original paper, however the proposed method substantially improves the results. On the other hand, **NADE** performed rather poorly due to the small sample size of the training set, despite our attempts to improve its performance by reducing the number of parameters to train and by introducing early stopping.

The UCI datasets (Red Wine, White Wine and Parkinsons) represent challenging datasets with non-linear dependencies and abrupt transitions between high and low density regions. This makes the densities difficult to model using standard tools such as mixtures of Gaussians or factor analysis. They also contain enough training sample points to allow a stronger performance by **NADE**. All discrete-valued variables were eliminated as well as one variable from every pair of variables that are highly correlated (Pearson correlation greater than 0.98). Following Uria et al. [2013], 90% of the data were used for training while 10% were held-out for testing. Two different graph factorizations (**F**, **M**) were used for the proposed method and for **LSCDE**.

In Table 3.2, we report the performance of the different models. Our method was among the statistically significant group of best models on Parkinsons dataset according to the two-sided paired $t - test$ at significance level of 5%. On the remaining datasets, it achieved the second best performance after **NADE**.

	KCEF	NADE	LSCDE
caution	$\mathbf{0.99 \pm 0.01}$	4.12 ± 0.02	1.19 ± 0.02
ftcollinssnow	$\mathbf{1.46 \pm 0.0}$	3.09 ± 0.02	1.56 ± 0.01
highway	$\mathbf{1.17 \pm 0.01}$	11.02 ± 1.05	1.98 ± 0.04
heights	$\mathbf{1.27 \pm 0.0}$	2.71 ± 0.0	1.3 ± 0.0
sniffer	$\mathbf{0.33 \pm 0.01}$	1.51 ± 0.04	0.48 ± 0.01
snowgeese	$\mathbf{0.72 \pm 0.02}$	2.9 ± 0.15	1.39 ± 0.05
GAGurine	$\mathbf{0.46 \pm 0.0}$	1.66 ± 0.02	0.7 ± 0.01
geyser	1.21 ± 0.04	1.43 ± 0.07	$\mathbf{0.7 \pm 0.01}$
topo	$\mathbf{0.67 \pm 0.01}$	4.26 ± 0.02	0.83 ± 0.0
BostonHousing	$\mathbf{0.3 \pm 0.0}$	3.46 ± 0.1	1.13 ± 0.01
CobarOre	3.42 ± 0.03	4.7 ± 0.02	$\mathbf{1.61 \pm 0.02}$
engel	$\mathbf{0.18 \pm 0.0}$	1.46 ± 0.02	0.76 ± 0.01
mcycle	$\mathbf{0.56 \pm 0.01}$	2.24 ± 0.01	0.93 ± 0.01
BigMac2003	$\mathbf{0.59 \pm 0.01}$	13.8 ± 0.13	1.63 ± 0.03

Table 3.1: Mean and std. deviation of the negative log-likelihood on benchmark data over 20 runs, with different random splits. In all cases $d_y = 1$. Best method in boldface (two-sided paired *t-test* at 5%).

	white -wine	parkinsons	red wine
KCEF-F	13.05 ± 0.36	$\mathbf{2.86 \pm 0.77}$	11.8 ± 0.93
KCEF-M	14.36 ± 0.37	5.53 ± 0.79	13.31 ± 0.88
LSCDE-F	13.59 ± 0.6	15.89 ± 1.48	14.43 ± 1.5
LSCDE-M	14.42 ± 0.66	10.22 ± 1.45	14.06 ± 1.36
NADE	$\mathbf{10.55 \pm 0.0}$	3.63 ± 0.0	$\mathbf{9.98 \pm 0.0}$

Table 3.2: UCI results: average and standard deviation of the negative log-likelihood over 5 runs with different random splits. Best method in boldface (two-sided paired *t-test* at 5%).

4 .3 Sampling

We compare samples generated from the approximate distribution obtained using different methods (**KEF**, **KCEF**, **NADE**). To get samples $(X_1, ..., X_d)$ from the joint distribution of **KCEF** we performed ancestral sampling, where a sample from the parents $\pi(i)$ of node i is first generated, and then X_i is sampled according to $p(X_i|X_{\pi(i)})$. We used the methodology and code in Strathmann et al. [2015] to sample from each conditional distribution $p(X_i|X_{\pi(i)})$ using an HMC proposal, since we have access to the gradient of the conditional densities and their un-normalized values. We trained the 3 models on Red Wine and Parkinsons datasets as described

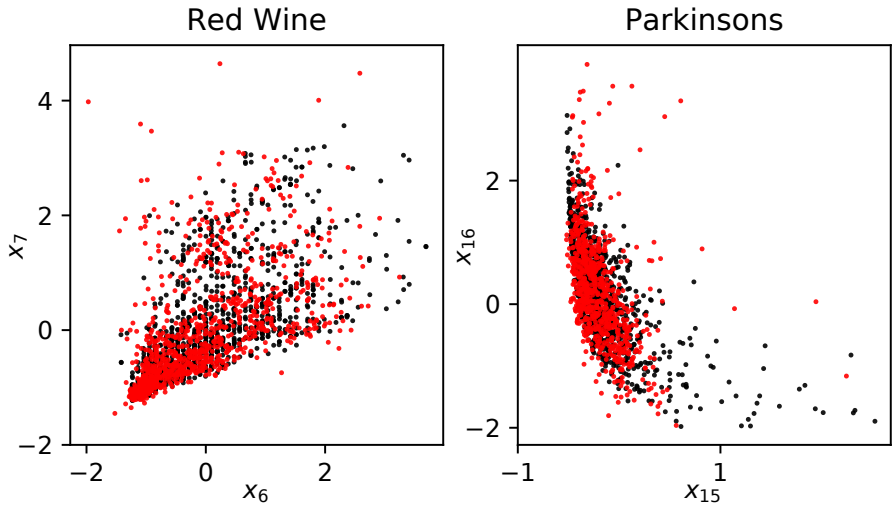


Figure 3.3: Scatter plot of 2-d slices of *red wine* and *parkinsons* data sets. Black points are real data, red are samples from the KCEF.

	$\mathcal{H}_{KEF<KCEF}$	$\mathcal{H}_{NADE<KCEF}$
parkinsons	0.523506	0.011467
red-wine	0.000791	0.326109

Table 3.3: p-values for the relative similarity test. Columns represents the p-values for testing whether samples from KEF (resp. KCEF) model are closer to the data than samples from the KCEF (resp. NADE).

previously, and generated joint samples from two-dimensional slices of data (see Figure 3.3). Since each conditional distribution is low-dimensional, we assumed an idealized scenario where the burn-in is completed after 100 iterations of the HMC sampler. We then run 20 samplers for 1000 and thin by a factor 10, which results in 2000 samples. As shown in Figure 3.3, **KCEF** is able to capture challenging properties of the target distribution, such as heteroscedasticity and sharp thresholds.

We also performed a test of relative similarity between the generated samples and the ground truth data following the methodology and code of Bounliphone et al. [2016]. Given samples from data X_m and generated samples Y_n and Z_r from two different methods, we test the hypothesis that P_x is closer to P_z than P_y according to the MMD metric. The null hypothesis

$$\mathcal{H}_{y<z} : MMD(P_x, P_y) \leq MMD(P_x, P_z)$$

is tested against the alternative at a significance level $\alpha = 5\%$ (see Bounliphone et al. [2016] for details). Table 3.3 shows the p-value for testing **KCEF** vs **KEF** and **NADE** vs **KCEF**. We see that **KCEF** significantly outperforms **NADE** with high confidence for the *parkinsons* dataset, consistent with Table 3.2. Performance of the two methods is not statistically distinguishable for the *red-wine* data. See the scatter plots in Figure 3.4 which visually confirm the result for the Red Wine and Parkinsons datasets. **KCEF** gives significantly better samples than **KEF** on *red-wine*: indeed, **KCEF** generally outperforms **KEF** on distributions where the densities exhibit abrupt transitions, as is clear by inspection of the plots in Figure 3.4 .

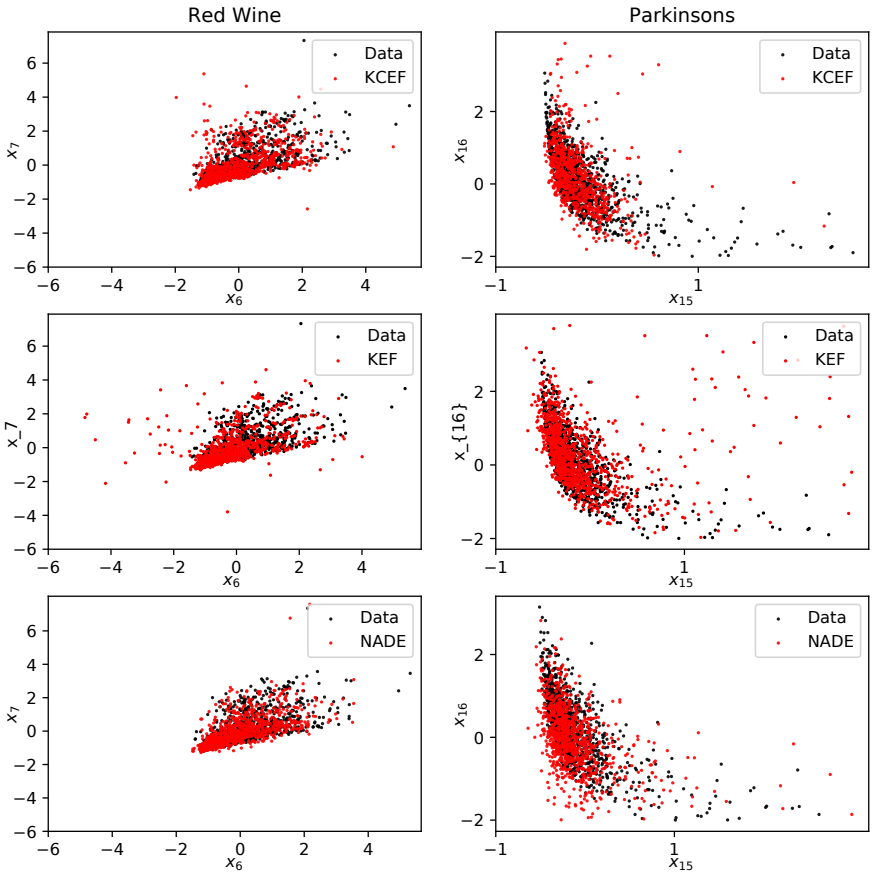


Figure 3.4: Scatter plot of 2-d slices of *red wine* and *parkinsons* data sets, the dimensions are (x_6, x_7) for *red wine* and (x_{15}, x_{16}) for *parkinsons*. The black points represent 1000 data points from the data sets. In red, 1000 samples from each of the three models KEF, KCEF and NADE.

Supplementary

In this section we prove Theorem 1 and Theorem 2.

A Preliminaries

A .1 Notation

We first introduce some relevant concepts from functional analysis. If E is Hilbert space we denote by $\langle ., . \rangle_E$ and $\|.\|_E$ its corresponding inner product and norm, respectively. If E and F are two Hilbert spaces, we use $\|.\|$ to denote the operator norm $\|A\| = \sup_{f:\|f\|\leq 1} \|Af\|$, where A is an operator from E to F. We denote by A^* the adjoint of A.

If E is separable with an orthonormal basis $\{e_k\}_k$, then $\|.\|_1$ and $\|.\|_2$ are the trace norm and Hilbert-Schmidt norm on E and are given by:

$$\|A\|_1 = \sum_k \langle (A^*A)^{\frac{1}{2}} e_k, e_k \rangle$$
$$\|A\|_2 = \|A^*A\|_1.$$

where A is an operator from E to E. $\lambda_{max}(A)$ is used to denote the algebraically largest eigenvalue of A. For f in E and g in F we denote by $g \otimes f$ the tensor product viewed as an application from E to F with $(g \otimes f)h = g\langle f, h \rangle_E$ for all h in E. $C^1(\Omega)$ denotes the space of continuously differentiable functions on Ω and $L^r(\Omega)$ the space of r-power Lebesgues-integrable function. Finally for any vector β in $\mathbb{R}^{nd}$, we use the notation $\beta_{(a,i)} = \beta_{(a-1)d+i}$ for $a \in [n]$ and $i \in [d]$.

A .2 Operator valued kernels and feature map derivatives

Let $\mathcal{X}$ and $\mathcal{Y}$ be two open subsets of $\mathbb{R}^p$ and $\mathbb{R}^d$. $\mathcal{H}_\mathcal{Y}$ is a reproducing kernel Hilbert space of functions $f : \mathcal{Y} \to \mathbb{R}$ with kernel $k_\mathcal{Y}$. We denote by $\mathcal{H}$ a vector-valued reproducing kernel Hilbert space of functions $T : x \mapsto T_x$ from $\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$ and we introduce the feature operator $\Gamma : x \mapsto \Gamma_x$ from $\mathcal{X}$ to $\mathcal{L}(\mathcal{H}_\mathcal{Y}, \mathcal{H})$ where $\mathcal{L}(\mathcal{H}_\mathcal{Y}, \mathcal{H})$ is the set of bounded operators from $\mathcal{H}_\mathcal{Y}$ to $\mathcal{H}$. For every $x \in \mathcal{X}$, Γ_x is an operator defined from $\mathcal{H}_\mathcal{Y}$ to $\mathcal{H}$.

The following reproducing properties will be used extensively:

- Reproducing property of the derivatives of a function in $\mathcal{H}_\mathcal{Y}$ (Steinwart and Christmann [2008], Lemma 4.34): provided that the kernel $k_\mathcal{Y}$ is differentiable m-times with respect to each coordinate, then all $f \in \mathcal{H}_\mathcal{Y}$ are differentiable for every multi-index $\alpha \in \mathbb{N}_0^d$ such that $\alpha \leq m$, and

$$\partial^\alpha f(y) = \langle f, \partial^\alpha k(y, .)\rangle_{\mathcal{H}_\mathcal{Y}} \qquad \forall y \in \mathcal{Y},$$

  where $\partial^\alpha k_y(y, y') = \frac{\partial^\alpha k(y,y')}{\partial^\alpha y}$. In particular we will use the notation

$$\partial_i k(y, y') = \frac{\partial k(y, y')}{\partial y_i}, \qquad \partial_{i+d} k(y, y') = \frac{\partial k(y, y')}{\partial y'_i}.$$

- Reproducing property in the vector-valued space $\mathcal{H}$: For any $f \in \mathcal{H}_\mathcal{Y}$ and any $T \in \mathcal{H}$ we have the following:

$$\langle T_x, f\rangle_{\mathcal{H}_\mathcal{Y}} = \langle T, \Gamma_x f\rangle_\mathcal{H}$$

  In particular for every $y \in \mathcal{Y}$ we get:

$$\langle T_x, k(y, \cdot)\rangle_{\mathcal{H}_\mathcal{Y}} = \langle T, \Gamma_x k(y, \cdot)\rangle_\mathcal{H}$$

  Using now the reproducing property in $\mathcal{H}_\mathcal{Y}$ we get:

$$T(x, y) := T_x(y) = \langle T, \Gamma_x k(y, \cdot)\rangle_\mathcal{H}$$

A .3 The conditional infinite dimensional exponential family

Let q_0 be a base density function of a probability distribution over $\mathcal{Y}$ and π a probability distribution over $\mathcal{X}$. π and q_0 are fixed and are assumed to be supported in the whole spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively.

We introduce the following functions $Z : \mathcal{H}_\mathcal{Y} \to \mathbb{R}_+^*$, such that for every $f \in \mathcal{H}_\mathcal{Y}$ we have

$$Z(f) := \int_\mathcal{Y} \exp\left(\langle f, k(y,.)\rangle_{\mathcal{H}_\mathcal{Y}}\right) q_0(\mathrm{d}y).$$

We consider now the following family of operators

$$\mathcal{T} = \{T \in \mathcal{H} : Z(T_x) < \infty, \forall x \in \mathcal{X}\}.$$

This allows to introduce the Kernel Conditional Exponential Family as the family of conditional distributions satisfying

$$\mathcal{P} = \left\{ p_T(x|y) = q_0(y) \frac{e^{(\langle T, \Gamma_x k(y,\cdot)\rangle_\mathcal{H})}}{Z(T_x)} \middle| T \in \mathcal{T} \right\}.$$

Given samples $(X_i, Y_i)_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$ following a joint distribution p_0 the goal is to approximate the conditional density function $p_0(y|x)$ in the case where $p_0(y|x) \in \mathcal{P}$ (i.e. $\exists T_0 \in \mathcal{T}$ such that $p_0(y|x) = p_{T_0}(y|x)$). To this end, we introduce the expected conditional score function between two conditional distributions $p(.|x)$ and $q(.|x)$ under π,

$$J(p||q) = \frac{1}{2} \int_x \int_y \sum_{i=1}^d \left[\partial_i \log p(y|x) - \partial_i \log q(y|x)\right]^2 p(\mathrm{d}y|x)\pi(\mathrm{d}x).$$

This function has the nice property that $J(p||q) \geq 0$ and that $J(p||q) = 0 \Leftrightarrow q = p$, which makes it a good candidate as a loss function.

The marginal distribtion $p_0(x)$ doesn't have to match $\pi(x)$ in general as long as they have the same support. For purpose of simplicity we will assume that $p_0(x) = \pi(x)$.

A .4 Assumptions

We make the following assumptions:

(A) (well specified) The true conditional density $p_0(y|x) = p_{T_0}(y|x) \in \mathcal{P}$ for some T_0 in $\mathcal{T}$.

(B) $\mathcal{Y}$ is a non-empty open subset of of the form $\mathbb{R}^d$ with a piecewise smooth boundary $\partial\mathcal{Y} := \overline{\mathcal{Y}} \setminus \mathcal{Y}$, where $\overline{\mathcal{Y}}$ denotes the closure of $\mathcal{Y}$.

(C) k is twice continuously differentiable on $\mathcal{Y} \times \mathcal{Y}$ and $\partial^{\alpha,\alpha}k$ is continuously extensible to $\overline{\mathcal{Y}} \times \overline{\mathcal{Y}}$ for all $|\alpha| \leq 2$.

(D) For all $x \in \mathcal{X}$ and all $i \in [d]$, as y approaches $\partial\mathcal{Y}$: $\|\partial_i k(y, \cdot)\|_{\mathcal{Y}} p_0(y|x) = o(\|y\|^{1-d})$

(E) The operator Γ is continuous in x and is uniformly bounded for the operator norm $\|\Gamma_x\|_{Op} \leq \kappa$ for all $x \in \mathcal{X}$.

(F) (Integrability) for some $\epsilon \geq 1$ and all $i \in [d]$:

$$\|\partial_i k(y, \cdot)\|_{\mathcal{Y}} \in L^{2\epsilon}(\mathcal{Y}, p_0),$$
$$\|\partial_i^2 k(y, \cdot)\|_{\mathcal{Y}} \in L^{\epsilon}(\mathcal{Y}, p_0),$$
$$\|\partial_i k(y, \cdot)\|_{\mathcal{Y}} \partial_i \log q_0(y) \in L^{\epsilon}(\mathcal{Y}, p_0).$$

Remark 1. *The continuity of the kernel k on the separable set $\mathcal{Y}$ ensures that $\mathcal{H}_{\mathcal{Y}}$ is also separable by [Steinwart and Christmann, 2008, Lemma 4.33]. Moreover, since $x \mapsto \Gamma_x$ is continuous it follows that the vector valued RKHS $\mathcal{H}$ consists of continuous functions with values in $\mathcal{H}_{\mathcal{Y}}$. Again, the separability of $\mathcal{X}$ and $\mathcal{H}_{\mathcal{Y}}$ ensures that $\mathcal{H}$ is also separable by [Carmeli et al., 2006, Corollary 5.2.].*

B Proofs

In this section, we prove the main theorems of the document, by extending the proofs of Sriperumbudur et al. [2017] to the case of the vector-valued RKHS. We provide complete steps for all the proofs, including those that carry over from the

earlier work, to make the presentation self-contained; the reader may compare with [Sriperumbudur et al., 2017, Section 8] to see the changes needed in the conditional setting.

B .1 Score Matching

Theorem 3 (Score Matching). *Under Assumptions (A) to (F), the following holds:*

1. $J(p_{T_0}||p_T) < +\infty$ *for all* $T \in \mathcal{T}$

2. *For all* $T \in \mathcal{H}$ *define*

$$J(T) = \frac{1}{2}\langle T - T_0, C(T - T_0)\rangle_{\mathcal{H}}, \tag{3.6}$$

where

$$C := \int_{\mathcal{X}\times\mathcal{Y}} \underbrace{\sum_{i=1}^{d} \left[\Gamma_x \partial_i k(y,\cdot) \otimes \Gamma_x \partial_i k(y,\cdot)\right]}_{C_{x,y}} p_0(dx,dy) = \mathbb{E}_{p_0}[C_{X,Y}]. \tag{3.7}$$

then C *a trace-class positive operator on* $\mathcal{H}$ *and for all* $T \in \mathcal{T}$ $J(T) = J(p_{T_0}||P_T)$.

3. *Alternatively,*

$$J(T) = \frac{1}{2}\langle T, CT\rangle_{\mathcal{H}} + \langle T, \Xi\rangle_{\mathcal{H}} + J(p_{T_0}||q_0).$$

where

$$\begin{aligned}\mathcal{H} \ni \Xi :&= \int_{\mathcal{X}\times\mathcal{Y}} \underbrace{\sum_{i=1}^{d} \Gamma_x \left[\partial_i \log q_0(y) \partial_i k(y,\cdot) + \partial_i^2 k(y,\cdot)\right]}_{\Xi_{x,y}} p_0(dx,dy) \\ &= \mathbb{E}_{p_0}[\Xi_{X,Y}]\end{aligned}$$

Moreover, T_0 *satisfies* $CT_0 = -\Xi$

4. *For any* $\lambda > 0$, *a unique minimizer* T_λ *of* $J_\lambda(T) := J(T) + \frac{\lambda}{2}\|T\|^2_{\mathcal{H}}$ *over* $\mathcal{H}$

exists and is given by:

$$T_\lambda = -(C + \lambda I)^{-1}\Xi = (C + \lambda I)^{-1}CT_0.$$

Proof. We prove the results in the same order as stated in the theorem:

Proof of (1). By the reproducing property of the real valued space $\mathcal{H}_\mathcal{Y}$ we have: $T(x, y) = \langle T_x, k(y, \cdot)\rangle_{\mathcal{H}_\mathcal{Y}}$. Using the reproducing property for the derivatives of real valued functions in an RKHS in Lemma 11, we get

$$\partial_i T(x, y) = \partial_i \langle T_x, k(y, \cdot)\rangle_{\mathcal{H}_\mathcal{Y}} = \langle T_x, \partial_i k(y, \cdot)\rangle_{\mathcal{H}_\mathcal{Y}} \qquad \forall i \in [d].$$

Finally, using the reproducing property in the vector-valued space $\mathcal{H}$,

$$\partial_i T(x, y) = \langle T, \Gamma_x \partial_i k(y, \cdot)\rangle_{\mathcal{H}}, \qquad \forall i \in [d].$$

it is easy to see that

$$J(p_{T_0}||p_T) = \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Y}} \sum_{i=1}^{d} \langle T_0 - T, \Gamma_x \partial_i k(y, .)\rangle^2_{\mathcal{H}} p_0(\mathrm{d}x, \mathrm{d}y). \tag{3.8}$$

By Assumptions **(E)** and **(F)**,

$$\|\Gamma_x \partial_i k(y, \cdot)\|_{\mathcal{H}} \leq \|\Gamma_x\|_{Op} \|\partial_i k(y, \cdot)\|_{\mathcal{H}_\mathcal{Y}} \leq \kappa \sqrt{\partial_i \partial_{i+d} k(y, y)} \in L^2(p_0),$$

and therefore by Cauchy-Schwarz inequality,

$$\begin{aligned} J(T) &= J(p_{T_0}||p_T) \\ &\leq \frac{1}{2}\|T_0 - T\|^2_{\mathcal{H}} \int_{\mathcal{X}\times\mathcal{Y}} \sum_{i=1}^{d} \|\Gamma_x \partial_i k(y, \cdot)\|^2_{\mathcal{H}} p_0(\mathrm{d}x, \mathrm{d}y) < +\infty. \end{aligned}$$

which means that $J(T) < \infty$ for all $T \in \mathcal{T}$.

Proof of (2). Starting from (3.8), it is easy to see that:

$$J(T) = \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Y}} \sum_{i=1}^{d} \langle T_0 - T, \Gamma_x \partial_i k(y,\cdot) \otimes \Gamma_x \partial_i k(y,\cdot)(T_0 - T)\rangle_{\mathcal{H}} p_0(\mathrm{d}x, \mathrm{d}y)$$
$$= \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Y}} \langle T_0 - T, C_{x,y}(T_0 - T)\rangle_{\mathcal{H}} p_0(\mathrm{d}x, \mathrm{d}y)$$

.

In the first line, we used the fact that $\langle a, b\rangle_{\mathcal{H}}^2 = \langle a, b\rangle_{\mathcal{H}}\langle a, b\rangle_{\mathcal{H}} = \langle a, b \otimes ba\rangle_{\mathcal{H}}$ for any a and b in a Hilbert space $\mathcal{H}$. By further observing that $C_{x,y}$ and $(T_0 - T) \otimes (T_0 - T)$ are Hilbert-Schmidt operators as $\|C_{x,y}\|_{HS} \leq \kappa^2 \sum_{i=1}^{d} \|\partial_i k(y,\cdot)\| < \infty$ by Lemma 7 and $\|(T_0 - T) \otimes (T_0 - T)\|_{HS} = \|(T_0 - T)\|_{\mathcal{H}}^2 < \infty$ we get that:

$$J(T) = \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Y}} \langle (T_0 - T) \otimes (T_0 - T), C_{x,y}\rangle_{HS} p_0(\mathrm{d}x, \mathrm{d}y)$$

Using Assumption **(F)** we have by Lemma 10 that $C_{x,y}$ is p_0-integrable in the Bochner sense [Retherford, 1978, Definition 1]) and that the inner product and integration may be interchanged:

$$J(T) = \frac{1}{2}\left\langle (T_0 - T) \otimes (T_0 - T), \int_{\mathcal{X}}\int_{\mathcal{Y}} C_{x,y} p_0(\mathrm{d}x, \mathrm{d}y) \right\rangle_{HS}$$
$$= \frac{1}{2}\langle T_0 - T, C(T_0 - T)\rangle_{\mathcal{H}}$$

Proof of (3). From (3.6) we have $J(T) = \frac{1}{2}\langle T, CT\rangle_{\mathcal{H}} - \langle T, CT_0\rangle_{\mathcal{H}} + \frac{1}{2}\langle T_0, CT_0\rangle_{\mathcal{H}}$. Recalling that: $\partial_i T(x,y) = \langle T, \Gamma_x \partial_i k(y,\cdot)\rangle_{\mathcal{H}}$ for all $i \in [d]$, and using $\partial_i T_0(x,y) =$

$\partial_i \log p_0(y|x) - \partial_i \log q_0(y|x)$ one gets:

$$\begin{aligned}
\langle T, CT_0\rangle_{\mathcal{H}} &= \int_{\mathcal{X}\times\mathcal{Y}} \Big[\sum_{i=1}^{d} \partial_i T(x,y)\partial_i T_0(x,y)\Big] p_0(\mathrm{d}x,\mathrm{d}y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \Big[\sum_{i=1}^{d} \partial_i T(x,y)\partial_i \log p_0(y|x)\Big] p_0(\mathrm{d}x)\mathrm{d}y \\
&\quad - \int_{\mathcal{X}\times\mathcal{Y}} \Big[\sum_{i=1}^{d} \partial_i T(x,y)\partial_i \log q_0(y|x)\Big] p_0(\mathrm{d}x,\mathrm{d}y) \\
&\overset{(a)}{=} \int_{\mathcal{X}} p_0(\mathrm{d}x) \int_{\partial\mathcal{Y}} p_0(y|x)\nabla_y T(x,y).\vec{dS} \\
&\quad - \int_{\mathcal{X}\times\mathcal{Y}} \Big[\sum_{i=1}^{d} \partial_i^2 T(x,y) + \partial_i T(x,y)\partial_i \log q_0(y|x)\Big] p_0(\mathrm{d}x,\mathrm{d}y).
\end{aligned}$$

(a) is obtained using the first Green's identity, where $\partial\mathcal{Y}$ is the boundary of $\mathcal{Y}$ and $\vec{dS}$ is the oriented surface element. The first term $\int_{\mathcal{X}} \pi(\mathrm{d}x) \int_{\partial\mathcal{Y}} p_0(y|x)\nabla_y T(x,y).\vec{dS}$ vanishes by Lemma 8, which relies on Assumption **(D)**. The second term can be written as: $\int_{\mathcal{X}\times\mathcal{Y}} \langle T, \Xi_{x,y}\rangle_{\mathcal{H}} p_0(\mathrm{d}x,\mathrm{d}y)$.

By Assumptions **(E)** and **(F)** $\Xi_{x,y}$ is Bochner p_0-integrable, therefore:

$$\int_{\mathcal{X}\times\mathcal{Y}} \langle T, \Xi_{x,y}\rangle_{\mathcal{H}} p_0(\mathrm{d}x,\mathrm{d}y) = \Big\langle T, \int_{\mathcal{X}\times\mathcal{Y}} \Xi_{x,y} p_0(\mathrm{d}x,\mathrm{d}y)\Big\rangle_{\mathcal{H}} = \langle T, \Xi\rangle_{\mathcal{H}}.$$

Hence $\langle T, CT_0\rangle_{\mathcal{H}} = -\langle T, \Xi\rangle_{\mathcal{H}}$ and $\Xi = -CT_0$. Moreover, one can clearly see that:

$$\langle T_0, CT_0\rangle_{\mathcal{H}} = \int_{\mathcal{X}\times\mathcal{Y}} \sum_{i=1}^{d} (\partial_i T_0(x,y))^2 p_0(\mathrm{d}x,\mathrm{d}y) = J(p_{T_0}||q_0).$$

And the result follows.

Proof of (4). For $\lambda > 0$, $(C + \lambda I)$ is invertible as C is a symmetric trace-class operator. Moreover, $(C + \lambda I)^{\frac{1}{2}}$ is well defined and one can easily see that:

$$J_\lambda(T) = \frac{1}{2}\|(C+\lambda I)^{\frac{1}{2}}T + (C+\lambda I)^{-\frac{1}{2}}\Xi\|_{\mathcal{H}}^2 - \frac{1}{2}\langle \Xi, (C+\lambda I)^{-1}\Xi\rangle_{\mathcal{H}} + c_0$$

with $c_0 = J(p_{T_0}||q_0)$. $J_\lambda(T)$ is minimized if and only if $(C+\lambda I)^{\frac{1}{2}}T = (C+\lambda I)^{-\frac{1}{2}}\Xi$ and therefore $T = (C+\lambda I)^{-1}\Xi$ is the unique minimizer of $J_\lambda(T)$. □

B .2 Estimator of T_0

Given samples $(X_a, Y_a)_{a=1}^n$ drawn i.i.d. from p_0 and $\lambda > 0$, we define the empirical score function as

$$\hat{J}(T) := \frac{1}{2}\langle T, \hat{C}T\rangle_{\mathcal{H}} + \langle T, \hat{\Xi}\rangle_{\mathcal{H}} + J(p_{T_0}||q_0).$$

where:

$$\hat{C} := \frac{1}{n}\sum_{a=1}^{n}\sum_{i=1}^{d} \Gamma_{X_a}\partial_i k(Y_a, \cdot) \otimes \Gamma_{X_a}\partial_i k(Y_a, \cdot)$$
$$\hat{\Xi} : = \frac{1}{n}\sum_{a=1}^{n}\sum_{i=1}^{d} \Gamma_{X_a}\big[\partial_i \log q_0(Y_a)\partial_i k(Y_a, \cdot), +\partial_i^2 k(Y_a, \cdot)\big].$$

are the empirical estimators of C and Ξ respectively.

Theorem 4 (Estimator of T_0). *For and any $\lambda > 0$, we have the following:*

1. *The unique minimizer $T_{\lambda,n}$ of $\hat{J}_\lambda(T) := \hat{J}(T) + \frac{\lambda}{2}\|T\|_{\mathcal{H}}^2$ over $\mathcal{H}$ exists and is given by*

$$T_{\lambda,n} = -(\hat{C} + \lambda I)^{-1}\hat{\Xi}.$$

2. *Moreover, $T_{\lambda,n}$ is of the form*

$$T_{\lambda,n} = -\frac{1}{\lambda}\hat{\Xi} + \sum_{b=1}^{n}\sum_{i=1}^{d} \beta_{(b-1)d+i}\Gamma_{X_b}\partial_i k(Y_b, \cdot),$$

where (β_b) are obtained by solving the following linear system:

$$(G + n\lambda I)\beta = \frac{h}{\lambda}$$

with:

$$(G)_{(a-1)d+i,(b-1)d+j} = \langle \Gamma_{X_a}\partial_i k(Y_a, .), \Gamma_{X_b}\partial_j k(Y_b, .)\,\rangle_{\mathcal{H}}.$$

and:

$$(h)_{(a-1)d+i} = \langle \hat{\Xi}, \Gamma_{X_a} \partial_i k(Y_a, .) \rangle_{\mathcal{H}}.$$

Proof. **Proof of (1).** The same proof as in Theorem 3 holds with C and Ξ replaced by $\hat{C}$ and $\hat{\Xi}$.

Proof of (2). We will use the general representer theorem stated in Lemma 13. We have that:

$$\begin{aligned} T_{\lambda,n} &= \operatorname*{arginf}_{T \in \mathcal{H}} \frac{1}{2} \langle T\hat{C}T \rangle_{\mathcal{H}} + \langle T, \hat{\Xi} \rangle_{\mathcal{H}} + \frac{\lambda}{2} \|T\|^2_{\mathcal{H}} \\ &= \operatorname*{arginf}_{T \in \mathcal{H}} \frac{1}{2} \sum_{a=1}^{n} \sum_{i=1}^{d} \langle T, \Gamma_{X_a} \partial_i k(Y_a, .) \rangle^2_{\mathcal{H}} + \langle T, \hat{\Xi} \rangle_{\mathcal{H}} + \frac{\lambda}{2} \|T\|^2_{\mathcal{H}} \\ &= \operatorname*{arginf}_{T \in \mathcal{H}} V(\langle T, \phi_1 \rangle_{\mathcal{H}}, ..., \langle T, \phi_{nd+1} \rangle_{\mathcal{H}}) + \frac{\lambda}{2} \|T\|^2_{\mathcal{H}}. \end{aligned}$$

Where $V(\theta_1, ..., \theta_{nd+1}) := \frac{1}{2n} \sum_{a=1}^{n} \sum_{i=1}^{d} \theta^2_{(a-1)d+i} + \theta_{nd+1}$ is a convex differentiable function and $\phi_{(a-1)d+i} := \Gamma_{X_a} \partial_i k(Y_a, .)$ where $a \in [n], i \in [d]$ and $\phi_{nd+1} = \hat{\Xi}$. Therefore, it follows from Lemma 13 that:

$$T_{\lambda,n} = \delta \hat{\Xi} + \sum_{a=1}^{n} \sum_{i=1}^{d} \beta_{(a-1)d+i} \phi_{(a-1)d+i}.$$

where δ and β satisfy:

$$\lambda(\beta, \delta) + \nabla V(K(\beta, \delta)) = 0$$

with $K = \begin{pmatrix} G & h \\ h^T & \|\hat{\Xi}\|^2_{\mathcal{H}} \end{pmatrix}$.

The gradient ∇V of V is given by $\nabla V(z, t) = (\frac{1}{n} z, 1)$. The above equation reduces then to $\lambda\delta + 1 = 0$ and $\lambda\beta + \frac{1}{n} G\beta + \frac{\delta}{n} h = 0$ which yields $\delta = -\frac{1}{\lambda}$ and $(\frac{1}{n} G + \lambda I)\beta = \frac{1}{n\lambda} h$. □

B .3 Consistency and convergence

Theorem 5 (Consistency and convergence rates for $T_{\lambda,n}$). *Let $\gamma > 0$ be a positive number and define $\alpha = \max(\frac{1}{2(\gamma+1)}, \frac{1}{4}) \in (\frac{1}{4}, \frac{1}{2})$, under Assumptions (A) to (F):*

1. *if $T_0 \in \overline{\mathcal{R}(C)}$ then $\|T_{n,\lambda} - T_0\| \to 0$ when $\lambda\sqrt{n} \to \infty$, $\lambda \to 0$ and $n \to \infty$.*

2. *if $T_0 \in \mathcal{R}(C^\gamma)$ for some $\gamma > 0$ then $\|T_{n,\lambda} - T_0\| = \mathcal{O}_{p_0}(n^{-\frac{1}{2}+\alpha})$ for $\lambda = n^{-\alpha}$*

Proof. Recalling that $T_{\lambda,n} = -(\hat{C} + \lambda I)^{-1}\hat{\Xi}$ We consider the following decomposition:

$$\begin{aligned}
T_{\lambda,n} - T_\lambda =& -(\hat{C} + \lambda I)^{-1}(\hat{\Xi} + (\hat{C} + \lambda I)T_\lambda) \\
\overset{(*)}{=}& -(\hat{C} + \lambda I)^{-1}(\hat{\Xi} + \hat{C}T_\lambda + C(T_0 - T_\lambda)) \\
=&(\hat{C} + \lambda I)^{-1}(C - \hat{C})(T_\lambda - T_0) - (\hat{C} + \lambda I)^{-1}(\hat{\Xi} + \hat{C}T_0) \\
=&(\hat{C} + \lambda I)^{-1}(C - \hat{C})(T_\lambda - T_0) - (\hat{C} + \lambda I)^{-1}(\hat{\Xi} - \Xi) \\
&+ (\hat{C} + \lambda I)^{-1}(C - \hat{C})T_0.
\end{aligned}$$

We used the fact that $\lambda T_\lambda = C(T_0 - T_\lambda)$ in $(*)$. Define now

$$\begin{aligned}
S_1 &:= \|(\hat{C} + \lambda I)^{-1}(C - \hat{C})(T_\lambda - T_0)\|_{\mathcal{H}} \\
S_2 &:= \|(\hat{C} + \lambda I)^{-1}(\hat{\Xi} - \Xi)\|_{\mathcal{H}} \\
S_3 &:= \|(\hat{C} + \lambda I)^{-1}(C - \hat{C})T_0\|_{\mathcal{H}} \\
\mathcal{A}_0(\lambda) &:= \|T_{\lambda,n} - T_0\|_{\mathcal{H}}.
\end{aligned}$$

it comes then:

$$\begin{aligned}
\|T_\lambda - T_0\|_{\mathcal{H}} &\leq \|T_{\lambda,n} - T_\lambda\|_{\mathcal{H}} + \|T_\lambda - T_0\|_{\mathcal{H}} \\
&\leq S_1 + S_2 + S_2 + \mathcal{A}_0(\lambda),
\end{aligned}$$

Using Lemma 16 we can bound S_1, S_2 and S_3. Note that $C_{x,y}$ as defined in

(3.7) is a positive, self-adjoint trace-class operator by Lemma 7 , we therefore have:

$$\begin{aligned}\|C_{x,y}\|_{HS}^2 &= \sum_{i,j=1}^{d} \langle \Gamma_x \partial_i k(y,\cdot), \Gamma_x \partial_j k(y,\cdot)\rangle_{\mathcal{H}}^2 \leq \sum_{i,j=1}^{d} \|\Gamma_x \partial_i k(y,\cdot)\|_{\mathcal{H}}^2 \|\Gamma_x \partial_j k(y,\cdot)\|_{\mathcal{H}}^2 \\ &\leq (\sum_{i=1}^{d} \|\Gamma_x \partial_i k(y,\cdot)\|_{\mathcal{H}}^2)^2 \leq d \sum_{i=1}^{d} \|\Gamma_x \partial_i k(y,\cdot)\|_{\mathcal{H}}^4 \leq d\kappa^4 \sum_{i=1}^{d} \|\partial_i k(y,\cdot)\|_{\mathcal{H}_{\mathcal{Y}}}^4.\end{aligned}$$

The last inequality is obtained using Assumption **(E)**. Using now Assumption **(F)** for $\epsilon = 2$ one can get:

$$\int_{\mathcal{X}\times\mathcal{Y}} \|C_{x,y}\|_{HS}^2 p_0(\mathrm{d}x,\mathrm{d}y) \leq d\kappa^4 \sum_{i=1}^{d} \int_{\mathcal{X}\times\mathcal{Y}} \|\partial_i k(y,\cdot)\|_{\mathcal{H}_{\mathcal{Y}}}^4 p_0(\mathrm{d}x,\mathrm{d}y) < +\infty.$$

Lemma 16 can then be applied to get the following inequalities:

$$\begin{aligned}S_1 \leq& \|(\hat{C}+\lambda I)^{-1}\| \|(C-\hat{C})(T_\lambda - T_0)\|_{\mathcal{H}} = \mathcal{O}_{p_0}(\frac{\mathcal{A}(\lambda)}{\lambda\sqrt{n}}) \\ S_3 \leq& \|(\hat{C}+\lambda I)^{-1}\| \|(C-\hat{C})T_0\| = \mathcal{O}_{p_0}(\frac{1}{\lambda\sqrt{n}})\|_{\mathcal{H}} \\ \|(C+\lambda I)^{-1}\| \leq& \frac{1}{\lambda}\end{aligned}$$

To bound S_2 we need to show that $\|\hat{\Xi} - \Xi\|_{\mathcal{H}} = \mathcal{O}_{p_0}(n^{-\frac{1}{2}})$. The same argument as in Sriperumbudur et al. [2017] holds:

$$\begin{aligned}\mathbb{E}_{p_0}\|\hat{\Xi} - \Xi\|_{\mathcal{H}}^2 &= \frac{1}{n}\left(\int_{\mathcal{X}\times\mathcal{Y}} \|\Xi_{x,y}\|_{\mathcal{H}}^2 p_0(\mathrm{d}x,\mathrm{d}y) - \|\Xi\|^2\right) \\ &\leq \frac{1}{n}\int_{\mathcal{X}\times\mathcal{Y}} \|\Xi_{x,y}\|_{\mathcal{H}}^2 p_0(\mathrm{d}x,\mathrm{d}y)\end{aligned}$$

By Assumption **(F)** for $\epsilon = 2$ we have that $\int_{\mathcal{X}\times\mathcal{Y}} \|\Xi_{x,y}\|_{\mathcal{H}}^2 p_0(\mathrm{d}x,\mathrm{d}y) < \infty$. One can therefore apply Chebychev inequality to get the results. It comes that:

$$S_2 \leq \|(\hat{C}+\lambda I)^{-1}\| \|\hat{\Xi} - \Xi\|_{\mathcal{H}} = \mathcal{O}_{p_0}(\frac{1}{\lambda\sqrt{n}})$$

Using the bounds on S_1, S_2 and S_3 we get:

$$\|T_{\lambda,n} - T_0\|_{\mathcal{H}} = \mathcal{O}_{p_0}(\frac{1}{\lambda\sqrt{n}} + \frac{\mathcal{A}_0(\lambda)}{\lambda\sqrt{n}}) + \mathcal{A}_0(\lambda) \tag{3.9}$$

1. By Lemma 15 we have $\mathcal{A}_0(\lambda) \to 0$ as $\lambda \to 0$ if $T_0 \in \overline{\mathcal{R}(C)}$. Therefore it follows from (3.9) that $\|T_{\lambda,n} - T_0\| \to 0$ as $\lambda \to 0$, $\lambda\sqrt{n} \to \infty$ and $n \to \infty$.

2. We have by Lemma 15 that if $T_0 \in \mathcal{R}(C^\gamma)$ for $\gamma > 0$ then:

$$\mathcal{A}_0(\lambda) \leq \max\{1, \|C\|^{\gamma-1}\}\|C^{-\gamma}T_0\|_{\mathcal{H}}\lambda^{\min\{1,\gamma\}}.$$

The result follows by choosing $\lambda = n^{-\max\{\frac{1}{4}, \frac{1}{2(\gamma+1)}\}} = n^{-\alpha}$.

□

We denote by $KL(p_{T_0}||p_T)$ the expected KL divergence between p_{T_0} and p_T under the marginal $p_0(x)$.

Theorem 6 (Consistency and convergence rates for $p_{T_{\lambda,n}}$). *Assuming Assumptions (A) to (F), and $\|k\|_\infty := \sup_{y\in\mathcal{Y}} k(y,y) < \infty$ and that $p_{T_0}(y|x)$ is supported on $\mathcal{Y}$ for all $x \in \mathcal{X}$ then the following holds:*

1. *$KL(p_{T_0}||p_{T_{\lambda,n}}) \to 0$ as $\lambda\sqrt{n} \to \infty$, $\lambda \to 0$ and $n \to \infty$.*

2. *If $T_0 \in \mathcal{R}(C^\gamma)$ for some $\gamma > 0$ then by defining $\alpha = \max(\frac{1}{2(\gamma+1)}, \frac{1}{4}) \in (\frac{1}{4}, \frac{1}{2})$, and choosing $\lambda = n^{-\alpha}$ we have that $KL(p_0||p_{T_{n,\lambda}}) = \mathcal{O}_{p_0}(n^{-1+2\alpha})$*

Proof. By Lemma 9, we have that $\mathcal{T} = \mathcal{H}$ and we can assume without loss of generality that $T_0 \in \overline{\mathcal{R}(C)}$. Using Lemma 14 (also see [van der Vaart and van Zanten, 2008, Lemma 3.1]), one can see that for a given x :

$$\begin{aligned} &KL(p_{T_0}(Y|x)||p_{T_{\lambda,n}}(Y|x)) \\ \leq &\|T_0(x) - T_{\lambda,n}(x)\|_\infty^2 \exp\|T_0(x) - T_{\lambda,n}(x)\|_\infty(1 + \|T_0(x) - T_{\lambda,n}(x)\|_\infty) \end{aligned} \tag{3.10}$$

Moreover, using Assumption (E) and the fact that $\|k\|_\infty < \infty$ one can see that

$$|T_0(x,y) - T_{\lambda,n}(x,y)|_{\mathcal{H}_\mathcal{Y}} = \langle T_0 - T_{\lambda,n}, \Gamma_x k(y,\cdot)\rangle_\mathcal{H}$$
$$\leq \|T_0 - T_{\lambda,n}\|_\mathcal{H} \|\Gamma_x k(y,\cdot)\|_\mathcal{H}$$

which gives after taking the supremum:

$$\|T_0(x) - T_{\lambda,n}(x)\|_\infty \leq \kappa \|k\|_\infty \|T_0 - T_{\lambda,n}\|_\mathcal{H} \tag{3.11}$$

for all $x \in \mathcal{X}$. Using (3.11) in (3.10) and taking the expectation with respect to x, one can conclude using Theorem 5.

□

Lemma 7. *Under Assumptions (C), (E) and (F) we have that:*

1. *$C_{x,y}$ is a trace-class positive and symmetric operator for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$*
2. *$C_{x,y}$ is Bochner-integrable for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$*
3. *C is a trace-class positive and symmetric operator*

Proof. Recall that $C = \int_{\mathcal{X}\times\mathcal{Y}} C_{x,y} p_0(\mathrm{d}x, \mathrm{d}y)$ where $C_{x,y} = \sum_{i=1}^d \Gamma_x \partial_i k(y,\cdot) \otimes \Gamma_x \partial_i k(y,\cdot)$ is a positive self-adjoint operator. Recalling that $\mathcal{H}$ is a separable Hilbert space, the trace norm of $C_{x,y}$ satisfies:

$$\|C_{x,y}\|_1 \leq \sum_{i=1}^d \|\Gamma_x \partial_i k(y,\cdot) \otimes \Gamma_x \partial_i k(y,\cdot)\|_1$$
$$\overset{(a)}{=} \sum_{i=1}^d \|\Gamma_x \partial_i k(y,\cdot)\|_\mathcal{H}^2 \leq \sum_{i=1}^d \|\Gamma_x\|_{Op}^2 \|\partial_i k(y,\cdot)\|_{\mathcal{H}_\mathcal{Y}}^2$$
$$\overset{(b)}{\leq} \kappa^2 \sum_{i=1}^d \|\partial_i k(y,\cdot)\|_{\mathcal{H}_\mathcal{Y}}^2 < \infty.$$

(a) follows from the definition of the trace-norm of the outer product of a vector in $\mathcal{H}$ with itself. (b) comes from Assumption (E). This implies that $C_{x,y}$ is trace-class. Moreover, by Assumption (F) for $\epsilon = 1$: $\|\partial_i k(y,\cdot)\|_{\mathcal{H}_\mathcal{Y}} \in L^{2\epsilon}(\mathcal{Y}, p_0)$ which leads

to:

$$\int_{\mathcal{X}\times\mathcal{Y}} \|C_{x,y}\|_1 p_0(\mathrm{d}x, \mathrm{d}y) < \infty.$$

This means that $C_{x,y}$ is p_0-integrable in the Bochner sense [Retherford, 1978, Definition 1 and Theorem 2] and its integral C is trace-class with:

$$\|C\|_1 = \Big\| \int_{\mathcal{X}\times\mathcal{Y}} C_{x,y} p_0(\mathrm{d}x, \mathrm{d}y) \Big\|_1 \leq \int_{\mathcal{X}\times\mathcal{Y}} \|C_{x,y}\|_1 p_0(\mathrm{d}x, \mathrm{d}y) < \infty.$$

□

Lemma 8. *Under Assumptions (B) to (D) we have the following:*

$$\int_{\mathcal{X}} \pi(\mathrm{d}x) \int_{\partial\mathcal{Y}} p_0(y|x) \nabla_y T(x,y).\vec{dS} = 0 \qquad \forall T \in \mathcal{T}$$

where $\partial\mathcal{Y}$ is the boundary of $\mathcal{Y}$ and $\vec{dS}$ is an oriented surface element of $\partial\mathcal{Y}$.

Proof. First let's prove that $\|\nabla_y T(x,y)\| p_0(y|x) = o(\|y\|^{1-d})$ for all $x \in \mathcal{X}$. Where the norm used is the euclidian norm in $\mathbb{R}^d$. Using the reproducing property and Cauchy-Schwarz inequality one can see that:

$$\begin{aligned}\|\nabla_y T(x,y)\|^2 = \sum_{i=1}^{d} (\partial_i T(x,y))^2 &= \sum_{i=1}^{d} \langle T_x, \partial_i k(y,.)\rangle^2 \\ &\leq \|T_x\|^2 \Big(\sum_{i=1}^{d} \|\partial_i k(y,.)\|^2 \Big)\end{aligned}$$

By Assumption (D), one can see that $\sqrt{\sum_{i=1}^{d} \|\partial_i k(y,.)\|^2} p_0(y|x) = o(\|x\|^{1-d})$, therefore it comes that $\|\nabla_y T(x,y)\| p_0(y|x) = o(\|y\|^{1-d})$. Using Lemma 12 one gets that $\int_{\partial\mathcal{Y}} p_0(y|x) \nabla_y T(x,y).\vec{dS} = 0$ for all $x \in \mathcal{X}$ which leads to the result.

□

Lemma 9 (Similar to Lemma 14 in Sriperumbudur et al. [2017])**.** *Suppose $\sup_{y\in\mathcal{Y}} k(y,y) < \infty$ and $supp(q_0) = \mathcal{Y}$. Then $\mathcal{T} = \mathcal{H}$ and for any T_0 there*

exists $\widetilde{T}_O \in \overline{\mathcal{R}(C)}$ *such that* $p_{\widetilde{T}_0} = p_0$.

Proof. The proof follows the same approach as in [Sriperumbudur et al., 2017, Lemma 14]. Since $\|k\|_\infty < \infty$ then $Z(T_x) \leq \exp \|T_x\| \|k\|_\infty < \infty$ for all $T \in \mathcal{H}$, therefore $\mathcal{T} = \mathcal{H}$. Moreover, since $\operatorname{supp}(p_{T_0})(y|x) = \mathcal{Y}$ for all x in $\mathcal{X}$, this implies that the null space of C $\mathcal{N}(C)$ can either be the set of functions $T(x, y) = m(x)$ or $\{0\}$. Indeed, for $T \in \mathcal{N}(C)$ we have $\langle T, CT \rangle = 0$ which leads to $\int_{\mathcal{X}\times\mathcal{Y}} \|\nabla_y T\|_2^2 p_0(\mathrm{d}x, \mathrm{d}y) = 0$ which means that p_0-almost surely, $T_x(y) = m(x)$ a constant function of y if the set of constant functions belong to $\mathcal{H}_\mathcal{Y}$, or $T_x(y) = 0$ otherwise. Let $\widetilde{T}_0$ be the orthogonal projection of T_0 onto $\overline{\mathcal{R}(C)} = \mathcal{N}(C)^\perp$ then T_0 can be written in the form $T_0(x, y) = m(x) + \widetilde{T}_0(x, y)$. It comes that $\int_\mathcal{Y} \exp T_0(x, y) q_0(\mathrm{d}y) = \exp m(x) \int_\mathcal{Y} \exp \widetilde{T}_0(x, y) q_0(\mathrm{d}y)$ almost surely in x. And we finally get p_0-almost surely:

$$p_{T_0}(y|x) = \frac{\exp T_0(x, y)}{Z(T_0(x))} = \frac{\exp T_0(x, y) + m(x)}{\exp m(x) Z(T_0(x))} = p_{T_0}(y|x)$$

□

C Known results

Lemma 10. *Let $\mathcal{X}$ be a topological space endowed with a probability distribution $\mathbb{P}$. Let B be a separable Banach space. Define R to be an B-valued measurable function on $\mathcal{X}$ in the Bochner sense (Retherford [1978] Definition 1), satisfying $\int_\mathcal{X} \|R(x)\|_B d\mathbb{P}(x) < \infty$, then R is $\mathbb{P}$-integrable in the Bochner sense (Retherford [1978] Definition 1, Theorem 6) and for any continuous linear operator T from B to another Banach space A, then TR is also $\mathbb{P}$-integrable in the Bochner sense and:*

$$\int_\mathcal{X} TR(x) d\mathbb{P}(x) = T \int_\mathcal{X} R(x) d\mathbb{P}(x)$$

For a proof of this result see Retherford [1978], Definition 1, Therorem 6 and 7.

Lemma 11 (RKHS of differentiable kernels (Steinwart and Christmann [2008] Chap 4.4, Corollary 4.36)). *Let $\mathcal{X} \in \mathbb{R}^d$ be an open subset, $m \geq 0$, and k be an*

m-times continuously differentiable kernel on $\mathcal{X}$ with RKHS $\mathcal{H}$. Then every function $f \in \mathcal{H}$ is m-times continuously differentiable, and for $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq m$ we have:

$$|\partial^\alpha f(x)| \leq \|f\|_{\mathcal{H}}^2 (\partial^{\alpha,\alpha} k(x,x))^{\frac{1}{2}}$$
$$\partial^\alpha f(x) = \langle f, \partial^\alpha k(x,\cdot)\rangle_{\mathcal{H}}$$

A proof of this result can be found in Steinwart and Christmann [2008] (Chap 4.4, Corollary 4.36)

Lemma 12. *Let Ω be an open set in $\mathbb{R}^d$ with piece-wise smooth boundary $\partial\Omega$. Let u be a real valued function defined over Ω and $v : \mathbb{R}^d \to \mathbb{R}^d$ a vector valued function. We assume that u and v are measurable and that $\|v(x)\||u(x)| = o(\|x\|^{1-d})$. Then the following surface integral is null:*

$$\int_{\partial\Omega} u(x)v(x).\mathrm{d}\vec{S} = 0$$

where $\mathrm{d}\vec{S}$ is an element of the surface $\partial\Omega$.

Lemma 13 (Generalized representer theorem)**.** *Let $\mathcal{H}$ be a vector-valued Hilbert space and let $(\phi_i)_{i=1}^m \in \mathcal{H}^m$. Suppose $J : \mathcal{H} \to \mathbb{R}$ is such that $J(T) = V(\langle T, \phi_1\rangle_{\mathcal{H}}, ..., \langle T, \phi_m\rangle_{\mathcal{H}})$ for $T \in \mathcal{H}$, where $V : \mathbb{R}^m \to \mathbb{R}$ is a convex and gâteaux-differentiable function. Define:*

$$T_\lambda = \operatorname*{arginf}_{T\in\mathcal{H}} J(T) + \frac{\lambda}{2}\|T\|_{\mathcal{H}}^2$$

where $\lambda > 0$. Then there exists $(\alpha_i)_{i=1}^m \in \mathbb{R}^m$ such that $T_\lambda = \sum_{i=1}^m \alpha_i\phi_i$ where $\alpha := (\alpha_1, ..., \alpha_m)$ satisfies the following equation:

$$(\lambda I + (\nabla V) \circ K)\alpha = 0,$$

with $(K)_{i,j} = \langle\phi_i, \phi_j\rangle_{\mathcal{H}}, i \in [m], j \in [m]$

Proof. Define $A : \mathcal{H} \to \mathbb{R}^m, T \mapsto (\langle T, \phi_i\rangle_{\mathcal{H}})_{i=1}^m$. Then $T_\lambda = \operatorname{arginf}_{T\in\mathcal{H}} V(AT) +$

$\frac{\lambda}{2}\|T\|_{\mathcal{H}}^2$. Taking the gâteaux-differential at T, the optimality condition yields:

$$\begin{aligned} 0 = A^*\nabla V(AT_\lambda) + \lambda T_\lambda &\Leftrightarrow A^*\Big(-\frac{1}{\lambda}\nabla V(AT_\lambda)\Big) = T_\lambda \\ &\Leftrightarrow (\exists \alpha \in \mathbb{R}^m) T_\lambda = A^*\alpha, \alpha = -\frac{1}{\lambda}\nabla V(AT_\lambda) \\ &\Leftrightarrow (\exists \alpha \in \mathbb{R}^m) T_\lambda = A^*\alpha, \alpha = -\frac{1}{\lambda}\nabla V(AA^*\alpha) \end{aligned}$$

where $A^* : \mathbb{R}^m \to \mathcal{H}$ is the adjoint of A which can be obtained as follows. Note that:

$$(\forall T \in \mathcal{H})\ (\forall \alpha \in \mathbb{R}^m) \quad \langle AT, \alpha\rangle = \sum_{i=1}^m \alpha_i \langle T, \phi\rangle_{\mathcal{H}} = \Big\langle T, \sum_{i=1}^m \alpha_i \phi_i \Big\rangle_{\mathcal{H}}$$

thus $A^*\alpha = \sum_{i=1}^m \alpha_i \phi_i$. Therefore $AA^*\alpha = \sum_{i=1}^m \alpha_j A\phi_j = \sum_{j=1}^m \alpha_j(\langle \phi_j, \phi_i\rangle_{\mathcal{H}})$ and hence $AA^* = K$. □

Lemma 14 (Bound on KL divergence between p_f and p_g (van der Vaart and van Zanten [2008] Lemma 3.1)). *Assume that $\|k\|_\infty < \infty$ and let f and g in $\mathcal{H}_\mathcal{Y}$ such that $Z(f)$ and $Z(g)$ are finite, then: $KL(p_f||q_g) \leq \|f-g\|_\infty^2 \exp\|f-g\|_\infty(1 + \|f-g\|_\infty)$*

Lemma 15 (Proposition A.3 in Sriperumbudur et al. [2017]). *Let C be a bounded, positive self-adjoint compact operator on a separable Hilbert space $\mathcal{H}$. For $\lambda > 0$ and $T \in \mathcal{H}$, define $T_\lambda := (C+\lambda I)^{-1}CT$ and $\mathcal{A}_\theta(\lambda) := \|C^\theta(T_\lambda - T)\|_{\mathcal{H}}$ for $\theta \geq 0$. Then the following hold.*

1. *For any $\theta > 0$, $\mathcal{A}_\theta(\lambda) \to 0$ as $\lambda \to 0$ and if $T \in \overline{\mathcal{R}(C)}$, then $\mathcal{A}_0(\lambda) \to 0$ as $\lambda \to 0$.*
2. *If $T \in \mathcal{R}(C^\beta)$ for $\beta \geq 0$ and $\beta + \theta > 0$, then*

$$\mathcal{A}_\theta(\lambda) \leq \max\{1, \|C\|^{\beta+\theta-1}\}\lambda^{\min\{1,\beta+\theta\}}\|C^{-\beta}T\|_{\mathcal{H}}$$

Lemma 16 (Proposition A.4 in Sriperumbudur et al. [2017]). *Let $\mathcal{X}$ be a topological space, $\mathcal{H}$ be a separable Hilbert space and $\mathcal{L}_2^+(\mathcal{H})$ be the space of posi-*

tive, self-adjoint Hilbert-Schmidt operators on $\mathcal{H}$. Define $R := \int_{\mathcal{X}} r(x)d\mathbb{P}(x)$ and $\hat{R} := \frac{1}{n}\sum_{a=1}^{m} r(X_a)$ where $\mathbb{P} \in M_+^1(\mathcal{X})$ is a positive measure with finite mean, $(X_a)_{a=1}^m \sim \mathbb{P}$ and r is an $\mathcal{L}_2^+(\mathcal{H})$-valued measurable function on $\mathcal{X}$ satisfying $\int_{\mathcal{X}} \|r(x)\|_{HS}^2 d\mathbb{P}(x) < \infty$. Define $g_\lambda := (R+\lambda I)^{-1}Rg$ for $g \in \mathcal{H}$, $\lambda > 0$ and $\mathcal{A}_0(\lambda) := \|g_\lambda - g\|_{\mathcal{H}}$. Let $\alpha \geq 0$ and $\theta \geq 0$. Then the following hold:

1. $\|(\hat{R}-R)(g_\lambda - g)\|_{\mathcal{H}} = O_{\mathbb{P}}(\frac{\mathcal{A}_0(\lambda)}{\sqrt{m}})$
2. $\|R^\alpha (R+\lambda I)^{-\theta}\| \leq \lambda^{\alpha-\theta}$.
3. $\|\hat{R}^\alpha (\hat{R}+\lambda I)^{-\theta}\| \leq \lambda^{\alpha-\theta}$.
4. $\|(R+\lambda I)^{-\theta}(\hat{R}-R)\| = O_{\mathbb{P}}(\frac{1}{\sqrt{m\lambda^{2\theta}}})$.

Part II

Structuring and regularizing implicit generative models

Chapter 4

Optimized MMD under gradient constraints

We propose a principled method for gradient-based regularization of the critic of GAN-like models trained by adversarially optimizing the kernel of a Maximum Mean Discrepancy (MMD). We show that controlling the gradient of the critic is vital to having a sensible loss function, and devise a method to enforce exact, analytical gradient constraints at no additional cost compared to existing approximate techniques based on additive regularizers. The new loss function is provably continuous, and experiments show that it stabilizes and accelerates training, giving image generation models that outperform state-of-the art methods on 160×160 CelebA and 64×64 unconditional ImageNet.

1 Introduction

There has been an explosion of interest in *implicit generative models* (IGMs) over the last few years, especially after the introduction of generative adversarial networks (GANs) Goodfellow et al. [2014]. These models allow approximate samples from a complex high-dimensional target distribution $\mathbb{P}$, using a model distribution $\mathbb{Q}_\theta$, where estimation of likelihoods, exact inference, and so on are not tractable. GAN-type IGMs have yielded very impressive empirical results, particularly for image generation, far beyond the quality of samples seen from most earlier generative models [e.g. Karras et al., 2018, Radford et al., 2016, Gulrajani et al., 2017, Huang

et al., 2018a, Jin et al., 2017].

These excellent results, however, have depended on adding a variety of methods of regularization and other tricks to stabilize the notoriously difficult optimization problem of GANs [Salimans et al., 2016, Radford et al., 2016]. Some of this difficulty is perhaps because when a GAN is viewed as minimizing a discrepancy $\mathcal{D}_{\mathrm{GAN}}(\mathbb{P}, \mathbb{Q}_\theta)$, its gradient $\nabla_\theta \mathcal{D}_{\mathrm{GAN}}(\mathbb{P}, \mathbb{Q}_\theta)$ does not provide useful signal to the generator if the target and model distributions are not absolutely continuous, as is nearly always the case Arjovsky and Bottou [2017].

An alternative set of losses are the integral probability metrics (IPMs) [Müller, 1997], which can give credit to models $\mathbb{Q}_\theta$ "near" to the target distribution $\mathbb{P}$ [Arjovsky et al., 2017, Gneiting and Raftery, 2007] and [Bottou et al., 2018, Section 4]. IPMs are defined in terms of a *critic function*: a "well behaved" function with large amplitude where $\mathbb{P}$ and $\mathbb{Q}_\theta$ differ most. The IPM is the difference in the expected critic under $\mathbb{P}$ and $\mathbb{Q}_\theta$, and is zero when the distributions agree. The Wasserstein IPMs, whose critics are made smooth via a Lipschitz constraint, have been particularly successful in IGMs [Arjovsky et al., 2017, Gulrajani et al., 2017, Genevay et al., 2018]. But the Lipschitz constraint must hold uniformly, which can be hard to enforce. A popular approximation has been to apply a gradient constraint only in expectation [Gulrajani et al., 2017]: the critic's gradient norm is constrained to be small on points chosen uniformly between $\mathbb{P}$ and $\mathbb{Q}$.

Another class of IPMs used as IGM losses are the Maximum Mean Discrepancies (MMDs) [Gretton et al., 2012], as in [Li et al., 2015, Dziugaite et al., 2015]. Here the critic function is a member of a reproducing kernel Hilbert space (except in Unterthiner et al. [2018], who learn a deep approximation to an RKHS critic). Better performance can be obtained, however, when the MMD kernel is not based directly on image pixels, but on learned features of images. Wasserstein-inspired gradient regularization approaches can be used on the MMD critic when learning these features: [Li et al., 2017] uses weight clipping [Arjovsky et al., 2017], and [Bińkowski* et al., 2018, Bellemare et al., 2017] use a gradient penalty [Gulrajani et al., 2017].

The recent Sobolev GAN [Mroueh et al., 2018] uses a similar constraint on the expected gradient norm, but phrases it as estimating a Sobolev IPM rather than loosely approximating Wasserstein. This expectation can be taken over the same distribution as [Gulrajani et al., 2017], but other measures are also proposed, such as $(\mathbb{P} + \mathbb{Q}_\theta)/2$. A second recent approach, the spectrally normalized GAN [Miyato et al., 2018], controls the Lipschitz constant of the critic by enforcing the spectral norms of the weight matrices to be 1. Gradient penalties also benefit GANs based on f-divergences [Nowozin et al., 2016]: for instance, the spectral normalization technique of [Miyato et al., 2018] can be applied to the critic network of an f-GAN. Alternatively, a gradient penalty can be defined to approximate the effect of blurring $\mathbb{P}$ and $\mathbb{Q}_\theta$ with noise [Roth et al., 2017], which addresses the problem of non-overlapping support Arjovsky and Bottou [2017]. This approach has recently been shown to yield locally convergent optimization in some cases with non-continuous distributions, where the original GAN does not Mescheder et al. [2018].

In this work, we introduce a novel regularization for the MMD GAN critic of [Bellemare et al., 2017, Li et al., 2017, Bińkowski* et al., 2018], which *directly targets generator performance*, rather than adopting regularization methods intended to approximate Wasserstein distances Arjovsky et al. [2017], Gulrajani et al. [2017]. The new MMD regularizer derives from an approach widely used in semi-supervised learning [Bousquet et al., 2004, Section 2], where the aim is to define a classification function f which is positive on $\mathbb{P}$ (the positive class) and negative on $\mathbb{Q}_\theta$ (negative class), in the absence of labels on many of the samples. The decision boundary between the classes is assumed to be in a region of low density for both $\mathbb{P}$ and $\mathbb{Q}_\theta$: f should therefore be flat where $\mathbb{P}$ and $\mathbb{Q}_\theta$ have support (areas with constant label), and have a larger slope in regions of low density. Bousquet et al. [2004] propose as their regularizer on f a sum of the variance and a density-weighted gradient norm.

We adopt a related penalty on the MMD critic, with the difference that we only apply the penalty on $\mathbb{P}$: thus, the critic is flatter where $\mathbb{P}$ has high mass, but does not vanish on the generator samples from $\mathbb{Q}_\theta$ (which we optimize). In excluding $\mathbb{Q}_\theta$ from the critic function constraint, we also avoid the concern raised by [Miyato

et al., 2018] that a critic depending on $\mathbb{Q}_\theta$ will change with the current minibatch – potentially leading to less stable learning. The resulting discrepancy is no longer an integral probability metric: it is asymmetric, and the critic function class depends on the target $\mathbb{P}$ being approximated.

We first discuss in Section 2 how MMD-based losses can be used to learn implicit generative models, and how a naive approach could fail. This motivates our new discrepancies, introduced in Section 3 . Section 4 demonstrates that these losses outperform state-of-the-art models for image generation.

2 Learning implicit generative models with MMD-based losses

An IGM is a model $\mathbb{Q}_\theta$ which aims to approximate a target distribution $\mathbb{P}$ over a space $\mathcal{X} \subseteq \mathbb{R}^d$. We will define $\mathbb{Q}_\theta$ by a *generator* function $G_\theta : \mathcal{Z} \to \mathcal{X}$, implemented as a deep network with parameters θ, where $\mathcal{Z}$ is a space of latent codes, say $\mathbb{R}^{128}$. We assume a fixed distribution on $\mathcal{Z}$, say $Z \sim \text{Uniform}\left([-1, 1]^{128}\right)$, and call $\mathbb{Q}_\theta$ the distribution of $G_\theta(Z)$. We will consider learning by minimizing a discrepancy $\mathcal{D}$ between distributions, with $\mathcal{D}(\mathbb{P}, \mathbb{Q}_\theta) \geq 0$ and $\mathcal{D}(\mathbb{P}, \mathbb{P}) = 0$, which we call our *loss*. We aim to minimize $\mathcal{D}(\mathbb{P}, \mathbb{Q}_\theta)$ with stochastic gradient descent on an estimator of $\mathcal{D}$.

In the present work, we will build losses $\mathcal{D}$ based on the Maximum Mean Discrepancy,

$$\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = \sup_{f\,:\,\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)], \tag{4.1}$$

an integral probability metric where the critic class is the unit ball within $\mathcal{H}_k$, the reproducing kernel Hilbert space with a kernel k. The optimization in (4.1) admits a simple closed-form optimal critic, $f^*(t) \propto \mathbb{E}_{X\sim\mathbb{P}}[k(X, t)] - \mathbb{E}_{Y\sim\mathbb{Q}}[k(Y, t)]$. There is also an unbiased, closed-form estimator of MMD_k^2 with appealing statistical properties Gretton et al. [2012] – in particular, its sample complexity is *independent* of the dimension of $\mathcal{X}$, compared to the exponential dependence Weed et al. [2019]

of the Wasserstein distance

$$\mathcal{W}(\mathbb{P}, \mathbb{Q}) = \sup_{f\,:\,\|f\|_{\mathrm{Lip}} \leq 1} \mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)]. \tag{4.2}$$

The MMD is *continuous in the weak topology* for any bounded kernel with Lipschitz embeddings [Sriperumbudur, 2016, Theorem 3.2(b)], meaning that if $\mathbb{P}_n$ converges in distribution to $\mathbb{P}$, $\mathbb{P}_n \xrightarrow{D} \mathbb{P}$, then $\mathrm{MMD}(\mathbb{P}_n, \mathbb{P}) \to 0$. ($\mathcal{W}$ is continuous in the slightly stronger Wasserstein topology [Villani, 2009, Definition 6.9]; $\mathbb{P}_n \xrightarrow{\mathcal{W}} \mathbb{P}$ implies $\mathbb{P}_n \xrightarrow{D} \mathbb{P}$, and the two notions coincide if $\mathcal{X}$ is bounded.) Continuity means the loss can provide better signal to the generator as $\mathbb{Q}_\theta$ approaches $\mathbb{P}$, as opposed to e.g. Jensen-Shannon where the loss could be constant until suddenly jumping to 0 [e.g. Arjovsky et al., 2017, Example 1]. The MMD is also *strict*, meaning it is zero iff $\mathbb{P} = \mathbb{Q}_\theta$, for *characteristic* kernels Sriperumbudur et al. [2011]. The Gaussian kernel yields an MMD both continuous in the weak topology and strict. Thus in principle, one need not conduct any alternating optimization in an IGM at all, but merely choose generator parameters θ to minimize MMD_k.

Despite these appealing properties, using simple pixel-level kernels leads to poor generator samples Dziugaite et al. [2015], Li et al. [2015], Sutherland et al. [2017], Bottou et al. [2018]. More recent MMD GANs Li et al. [2017], Bellemare et al. [2017], Bińkowski* et al. [2018] achieve better results by using a parameterized *family* of kernels, $\{k_\psi\}_{\psi\in\Psi}$, in the Optimized MMD loss previously studied by Sriperumbudur et al. [2009], Sriperumbudur [2016]:

$$\mathcal{D}^{\Psi}_{\mathrm{MMD}}(\mathbb{P}, \mathbb{Q}) := \sup_{\psi\in\Psi} \mathrm{MMD}_{k_\psi}(\mathbb{P}, \mathbb{Q}). \tag{4.3}$$

We primarily consider kernels defined by some fixed kernel K on top of a learned low-dimensional representation $\phi_\psi : \mathcal{X} \to \mathbb{R}^s$, i.e. $k_\psi(x, y) = K(\phi_\psi(x), \phi_\psi(y))$, denoted $k_\psi = K \circ \phi_\psi$. In practice, K is a simple characteristic kernel, e.g. Gaussian, and ϕ_ψ is usually a deep network with output dimension say $s = 16$ [Bińkowski* et al., 2018] or even $s = 1$ (in our experiments). If ϕ_ψ is powerful enough, this choice is sufficient as soon as ϕ_ψ is optimized and we need

not try to ensure each k_ψ is characteristic, as did Li et al. [2017].

Proposition 17. *Suppose $k = K \circ \phi_\psi$, with K characteristic and $\{\phi_\psi\}$ rich enough that for any $\mathbb{P} \neq \mathbb{Q}$, there is a $\psi \in \Psi$ for which $\phi_\psi \# \mathbb{P} \neq \phi_\psi \# \mathbb{Q}$.*[1] *Then if $\mathbb{P} \neq \mathbb{Q}$, $\mathcal{D}^\Psi_{\mathrm{MMD}}(\mathbb{P}, \mathbb{Q}) > 0$.*

Proof. Let $\hat{\psi} \in \Psi$ be such that $\phi_{\hat{\psi}}(\mathbb{P}) \neq \phi_{\hat{\psi}}(\mathbb{Q})$. Then, since K is characteristic,

$$\mathcal{D}^\Psi_{\mathrm{MMD}}(\mathbb{P}, \mathbb{Q}) = \sup_{\psi \in \Psi} \mathrm{MMD}_K(\phi_\psi \# \mathbb{P}, \phi_\psi \# \mathbb{Q}) \geq \mathrm{MMD}_K(\phi_{\hat{\psi}} \# \mathbb{P}, \phi_{\hat{\psi}} \# \mathbb{Q}) > 0.$$

To estimate $\mathcal{D}^\Psi_{\mathrm{MMD}}$, one can conduct alternating optimization to estimate a $\hat{\psi}$ and then update the generator according to $\mathrm{MMD}_{k_{\hat{\psi}}}$, similar to the scheme used in GANs and WGANs. (This form of estimator is justified by an envelope theorem [Milgrom and Segal, 2002], although it is invariably biased Bińkowski* et al. [2018].) Unlike $\mathcal{D}_{\mathrm{GAN}}$ or $\mathcal{W}$, fixing a $\hat{\psi}$ and optimizing the generator still yields a sensible distance $\mathrm{MMD}_{k_{\hat{\psi}}}$.

Early attempts at minimizing $\mathcal{D}^\Psi_{\mathrm{MMD}}$ in an IGM, though, were unsuccessful [Sutherland et al., 2017, footnote 7]. This could be because for some kernel classes, $\mathcal{D}^\Psi_{\mathrm{MMD}}$ is stronger than Wasserstein or MMD.

Example 1 (DiracGAN [Mescheder et al., 2018])**.** *We wish to model a point mass at the origin of $\mathbb{R}$, $\mathbb{P} = \delta_0$, with any possible point mass, $\mathbb{Q}_\theta = \delta_\theta$ for $\theta \in \mathbb{R}$. We use a Gaussian kernel of any bandwidth, which can be written as $k_\psi = K \circ \phi_\psi$ with $\phi_\psi(x) = \psi x$ for $\psi \in \Psi = \mathbb{R}$ and $K(a, b) = \exp\left(-\frac{1}{2}(a-b)^2\right)$. Then*

$$\mathrm{MMD}^2_{k_\psi}(\delta_0, \delta_\theta) = 2\left(1 - \exp\left(-\tfrac{1}{2}\psi^2\theta^2\right)\right), \qquad \mathcal{D}^\Psi_{\mathrm{MMD}}(\delta_0, \delta_\theta) = \begin{cases} \sqrt{2} & \theta \neq 0 \\ 0 & \theta = 0 \end{cases}.$$

Considering $\mathcal{D}^\Psi_{\mathrm{MMD}}(\delta_0, \delta_{1/n}) = \sqrt{2} \not\to 0$, even though $\delta_{1/n} \xrightarrow{\mathcal{W}} \delta_0$, shows that the Optimized MMD distance is not continuous in the weak or Wasserstein topologies.

[1] $f\#\mathbb{P}$ denotes the *pushforward* of a distribution: if $X \sim \mathbb{P}$, then $f(X) \sim f\#\mathbb{P}$.

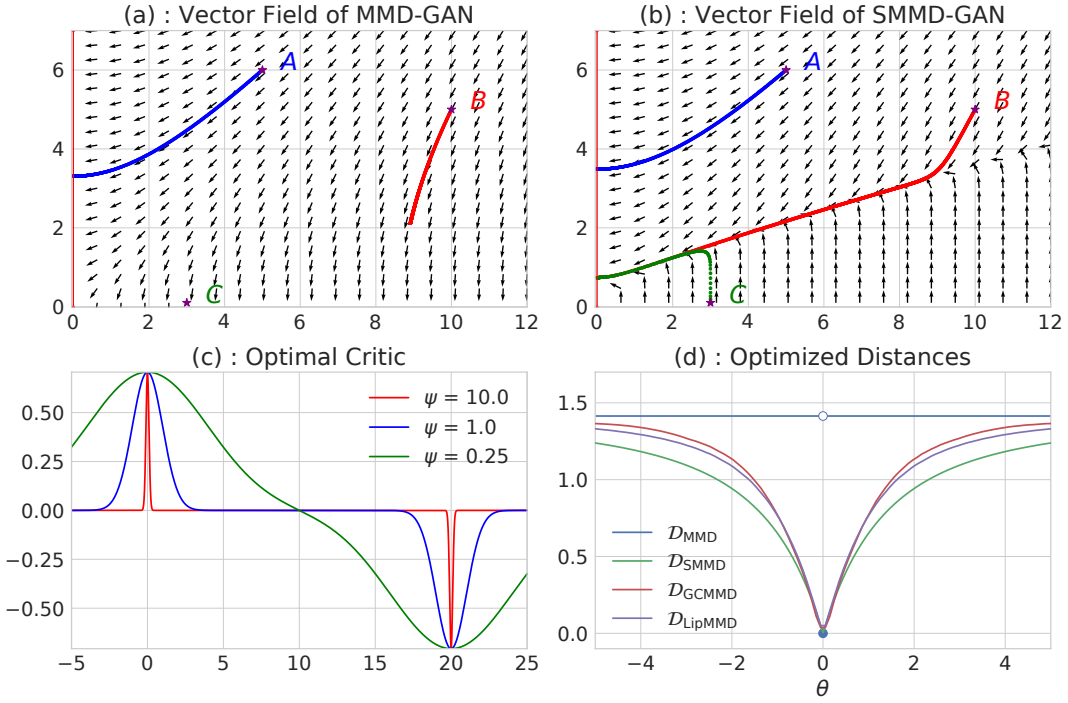


Figure 4.1: The setting of Example 1. (a, b): parameter-space gradient fields for the MMD and the SMMD (Section 3 .3); the horizontal axis is θ, and the vertical $1/\psi$. (c): optimal MMD critics for $\theta = 20$ with different kernels. (d): the MMD and the distances of Section 3 optimized over ψ.

This also causes optimization issues. Figure 4.1 (a) shows gradient vector fields in parameter space, $v(\theta, \psi) \propto \big(-\nabla_\theta \operatorname{MMD}^2_{k_\psi}(\delta_0, \delta_\theta), \nabla_\psi \operatorname{MMD}^2_{k_\psi}(\delta_0, \delta_\theta)\big)$. Some sequences following v (e.g. A) converge to an optimal solution $(0, \psi)$, but some (B) move in the wrong direction, and others (C) are stuck because there is essentially no gradient. Figure 4.1 (c, red) shows that the optimal $\mathcal{D}^{\Psi}_{\mathrm{MMD}}$ critic is very sharp near $\mathbb{P}$ and $\mathbb{Q}$; this is less true for cases where the algorithm converged.

We can avoid these issues if we ensure a bounded Lipschitz critic:[2]

Proposition 18. *Define the critic function* $f_\psi(x)$ *as:*

$$f_\psi(x) = (\mathbb{E}_{X\sim\mathbb{P}} k_\psi(X, x) - \mathbb{E}_{Y\sim\mathbb{Q}} k_\psi(Y, x))/\operatorname{MMD}_{k_\psi}(\mathbb{P}, \mathbb{Q}).$$

Assume the critics $f_\psi(x)$ *are uniformly bounded and have a common Lipschitz*

[2] [Li et al., 2017, Theorem 4] makes a similar claim to Proposition 18, but its proof was incorrect: it tries to uniformly bound $\operatorname{MMD}_{k_\psi} \leq \mathcal{W}^2$, but the bound used is for a Wasserstein in terms of $\|k_\psi(x,\cdot) - k_\psi(y,\cdot)\|_{\mathcal{H}_{k_\psi}}$.

constant: $\sup_{x\in\mathcal{X},\psi\in\Psi}|f_\psi(x)| < \infty$ *and* $\sup_{\psi\in\Psi}\|f_\psi\|_{\mathrm{Lip}} < \infty$. *In particular, this holds when* $k_\psi = K \circ \phi_\psi$ *and*

$$\sup_{a\in\mathbb{R}^s} K(a,a) < \infty, \qquad \|K(a,\cdot) - K(b,\cdot)\|_{\mathcal{H}_K} \le L_K\|a-b\|_{\mathbb{R}^s},$$

$$\sup_{\psi\in\Psi}\|\phi_\psi\|_{\mathrm{Lip}} \le L_\phi < \infty.$$

Then $\mathcal{D}^\Psi_{\mathrm{MMD}}$ *is continuous in the weak topology: if* $\mathbb{P}_n \xrightarrow{D} \mathbb{P}$, *then* $\mathcal{D}^\Psi_{\mathrm{MMD}}(\mathbb{P}_n,\mathbb{P}) \to 0$.

Proof. The main result is [Dudley, 2002, Corollary 11.3.4]. To show the claim for $k_\psi = K\circ\phi_\psi$, note that $|f_\psi(x) - f_\psi(y)| \le \|f_\psi\|_{\mathcal{H}_{k_\psi}}\|k_\psi(x,\cdot) - k_\psi(y,\cdot)\|_{\mathcal{H}_{k_\psi}}$, which since $\|f_\psi\|_{\mathcal{H}_{k_\psi}} = 1$ is

$$\|K(\phi_\psi(x),\cdot) - K(\phi_\psi(y),\cdot)\|_{\mathcal{H}_K} \le L_K\|\phi_\psi(x) - \phi_\psi(y)\|_{\mathbb{R}^s} \le L_K L_\phi\|x-y\|_{\mathbb{R}^d}.$$

Indeed, if we put a box constraint on ψ Li et al. [2017] or regularize the gradient of the critic function Bińkowski* et al. [2018], the resulting MMD GAN generally matches or outperforms WGAN-based models. Unfortunately, though, an additive gradient penalty doesn't substantially change the vector field of Figure 4.1 (a), as shown in Figure 4.10 (Section D .1). We will propose distances with much better convergence behavior.

3 New discrepancies for learning implicit generative models

Our aim here is to introduce a discrepancy that can provide useful gradient information when used as an IGM loss. Proofs of results in this section are deferred to Section A .

3 .1 Lipschitz Maximum Mean Discrepancy

Proposition 18 shows that an MMD-like discrepancy can be continuous under the weak topology even when optimizing over kernels, if we directly restrict the critic

functions to be Lipschitz. We can easily define such a distance, which we call the Lipschitz MMD: for some $\lambda > 0$,

$$\mathrm{LipMMD}_{k,\lambda}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{H}_k\,:\,\|f\|^2_{\mathrm{Lip}}+\lambda\|f\|^2_{\mathcal{H}_k}\leq 1} \mathbb{E}_{X\sim\mathbb{P}}\left[f(X)\right] - \mathbb{E}_{Y\sim\mathbb{Q}}\left[f(Y)\right]. \quad (4.4)$$

For a universal kernel k, we conjecture that $\lim_{\lambda\to 0}\mathrm{LipMMD}_{k,\lambda}(\mathbb{P},\mathbb{Q}) \to \mathcal{W}(\mathbb{P},\mathbb{Q})$. But for any k and λ, LipMMD is upper-bounded by $\mathcal{W}$, as (4.4) optimizes over a smaller set of functions than (4.2). Thus $\mathcal{D}^{\Psi,\lambda}_{\mathrm{LipMMD}}(\mathbb{P},\mathbb{Q}) := \sup_{\psi\in\Psi}\mathrm{LipMMD}_{k_\psi,\lambda}(\mathbb{P},\mathbb{Q})$ is also upper-bounded by $\mathcal{W}$, and hence is continuous in the Wasserstein topology. It also shows excellent empirical behavior on Example 1 (Figure 4.1 (d), and Figure 4.10 in Section D .1). But estimating $\mathrm{LipMMD}_{k,\lambda}$, let alone $\mathcal{D}^{\Psi,\lambda}_{\mathrm{LipMMD}}$, is in general extremely difficult (Section B), as finding $\|f\|_{\mathrm{Lip}}$ requires optimization in the input space. Constraining the *mean* gradient rather than the *maximum*, as we will do next, is far more tractable.

3 .2 Gradient-Constrained Maximum Mean Discrepancy

We define the Gradient-Constrained MMD for $\lambda > 0$ and using some measure μ as

$$\mathrm{GCMMD}_{\mu,k,\lambda}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{H}_k\,:\,\|f\|_{S(\mu),k,\lambda}\leq 1} \mathbb{E}_{X\sim\mathbb{P}}\left[f(X)\right] - \mathbb{E}_{Y\sim\mathbb{Q}}\left[f(Y)\right], \quad (4.5)$$

$$\text{where } \|f\|^2_{S(\mu),k,\lambda} := \|f\|^2_{L^2(\mu)} + \|\nabla f\|^2_{L^2(\mu)} + \lambda\|f\|^2_{\mathcal{H}_k}. \quad (4.6)$$

$\|\cdot\|^2_{L^2(\mu)} = \int\|\cdot\|^2\,\mu(\mathrm{d}x)$ denotes the squared L^2 norm. Rather than directly constraining the Lipschitz constant, the second term $\|\nabla f\|^2_{L^2(\mu)}$ encourages the function f to be flat where μ has mass. In experiments we use $\mu = \mathbb{P}$, flattening the critic near the target sample. We add the first term following Bousquet et al. [2004]: in one dimension and with μ uniform, $\|\cdot\|_{S(\mu),\cdot,0}$ is then an RKHS norm with the kernel $\kappa(x,y) = \exp(-\|x-y\|)$, which is also a Sobolev space. The correspondence to a Sobolev norm is lost in higher dimensions [Wendland, 2005, Ch. 10], but we also found the first term to be beneficial in practice.

We can exploit some properties of $\mathcal{H}_k$ to compute (4.5) analytically. Call the difference in kernel mean embeddings $\eta := \mathbb{E}_{X\sim\mathbb{P}}[k(X,\cdot)] - \mathbb{E}_{Y\sim\mathbb{Q}}[k(Y,\cdot)] \in \mathcal{H}_k$;

recall $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\eta\|_{\mathcal{H}_k}$.

Proposition 19. *Let $\hat{\mu} = \sum_{m=1}^{M} \delta_{X_m}$ be an empirical measure of M points. Let $\eta(X) \in \mathbb{R}^M$ have mth entry $\eta(X_m)$, and $\nabla\eta(X) \in \mathbb{R}^{Md}$ have (m,i)th entry* [3] *$\partial_i\eta(X_m)$. Then under Assumptions (A) to (D) in Section A .1, the Gradient-Constrained MMD is*

$$\mathrm{GCMMD}^2_{\hat{\mu},k,\lambda}(\mathbb{P}, \mathbb{Q}) = \frac{1}{\lambda}\left(\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) - \bar{P}(\eta)\right)$$
$$\bar{P}(\eta) = \begin{bmatrix} \eta(X) \\ \nabla\eta(X) \end{bmatrix}^\mathsf{T} \left(\begin{bmatrix} K & G^\mathsf{T} \\ G & H \end{bmatrix} + M\lambda I_{M+Md} \right)^{-1} \begin{bmatrix} \eta(X) \\ \nabla\eta(X) \end{bmatrix},$$

where K is the kernel matrix $K_{m,m'} = k(X_m, X_{m'})$, G is the matrix of left derivatives $G_{(m,i),m'} = \partial_i k(X_m, X_{m'})$, and H that of derivatives of both arguments $H_{(m,i),(m',j)} = \partial_i\partial_{j+d}k(X_m, X_{m'})$.

As long as $\mathbb{P}$ and $\mathbb{Q}$ have integrable first moments, and μ has second moments, Assumptions (A) to (D) are satisfied e.g. by a Gaussian or linear kernel on top of a differentiable ϕ_ψ. We can thus estimate the GCMMD based on samples from $\mathbb{P}$, $\mathbb{Q}$, and μ by using the empirical mean $\hat{\eta}$ for η.

This discrepancy indeed works well in practice: Section 4 .1.2 shows that optimizing our estimate of $\mathcal{D}^{\mu,\Psi,\lambda}_{\mathrm{GCMMD}} = \sup_{\psi\in\Psi} \mathrm{GCMMD}_{\mu,k_\psi,\lambda}$ yields a good generative model on MNIST. But the linear system of size $M + Md$ is impractical: even on 28×28 images and using a low-rank approximation, the model took days to converge. We therefore design a less expensive discrepancy in the next section.

The GCMMD is related to some discrepancies previously used in IGM training. The Fisher GAN [Mroueh and Sercu, 2017] uses only the variance constraint $\|f\|^2_{L^2(\mu)} \leq 1$. The Sobolev GAN [Mroueh et al., 2018] constrains $\|\nabla f\|^2_{L^2(\mu)} \leq 1$, along with a vanishing boundary condition on f to ensure a well-defined solution (although this was not used in the implementation, and can cause very unintuitive critic behavior; see Section D .2). The authors considered several choices of μ,

[3] We use (m,i) to denote $(m-1)d + i$; thus $\nabla\eta(X)$ stacks $\nabla\eta(X_1), \ldots, \nabla\eta(X_M)$ into one vector.

including the WGAN-GP measure [Gulrajani et al., 2017] and mixtures $(\mathbb{P} + \mathbb{Q}_\theta)/2$. Rather than enforcing the constraints in closed form as we do, though, these models used additive regularization. We will compare to the Sobolev GAN in experiments.

3 .3 Scaled Maximum Mean Discrepancy

We will now derive a lower bound on the Gradient-Constrained MMD which retains many of its attractive qualities but can be estimated in time linear in the dimension d.

Proposition 20. *Make Assumptions (A) to (D). For any* $f \in \mathcal{H}_k$, $\|f\|_{S(\mu),k,\lambda} \leq \sigma_{\mu,k,\lambda}^{-1}\|f\|_{\mathcal{H}_k}$, *where*

$$\sigma_{\mu,k,\lambda} := 1 \Big/ \sqrt{\lambda + \int k(x,x)\mu(\mathrm{d}x) + \sum_{i=1}^{d} \int \frac{\partial^2 k(y,z)}{\partial y_i \partial z_i}\Big|_{(y,z)=(x,x)} \mu(\mathrm{d}x)}.$$

Depending on the choice of the kernel, $\sigma_{\mu,k,\lambda}$ can have a simple expression. For instance, if $k_\psi = K \circ \phi_\psi$ and $K(a,b) = g(-\|a-b\|^2)$, then

$$\sigma_{k,\mu,\lambda}^{-2} = \lambda + g(0) + 2|g'(0)|\mathbb{E}_\mu\left[\|\nabla\phi_\psi(X)\|_F^2\right].$$

Estimating these terms based on samples from μ is straightforward, giving a natural estimator for $\sigma_{\mu,k,\lambda}$. We then define the Scaled Maximum Mean Discrepancy based on Proposition 20:

$$\begin{aligned}\mathrm{SMMD}_{\mu,k,\lambda}(\mathbb{P},\mathbb{Q}) :&= \sup_{f\,:\,\sigma_{\mu,k,\lambda}^{-1}\|f\|_{\mathcal{H}}\leq 1} \mathbb{E}_{X\sim\mathbb{P}}\left[f(X)\right] - \mathbb{E}_{Y\sim\mathbb{Q}}\left[f(Y)\right] \\ &= \sigma_{\mu,k,\lambda}\,\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}). \end{aligned} \tag{4.7}$$

While the GCMMD is obtained by constraining the witness function to be within an ellipsoid of the RKHS to control its smoothness, The SMMD is obtained by imposing a stronger constrain on the witness function compared to the GCMMD. Indeed, the latter controls the smoothness of the witness function by constraining in an ellipsoid of the RKHS. However, such constraint requires solving a linear system (Proposition 19) which can be computationally demanding. Instead, the SMMD

further constrains the witness function to be in a ball inside the ellipsoid defined by GCMMD, thus controlling the smoothness of the witness function while having a simpler expression.

Because the constraint in the optimization of (4.7) is more restrictive than in that of (4.5), we have that $\mathrm{SMMD}_{\mu,k,\lambda}(\mathbb{P},\mathbb{Q}) \leq \mathrm{GCMMD}_{\mu,k,\lambda}(\mathbb{P},\mathbb{Q})$. The Sobolev norm $\|f\|_{S(\mu),\lambda}$, and a fortiori the gradient norm under μ, is thus also controlled for the SMMD critic. Of course, if μ and k are fixed, the SMMD is simply a constant times the MMD, and so behaves in essentially the same way as the MMD. But optimizing the SMMD over a kernel family Ψ gives a distance $\mathcal{D}^{\mu,\Psi,\lambda}_{\mathrm{SMMD}}(\mathbb{P},\mathbb{Q})$ that is very different from $\mathcal{D}^{\Psi}_{\mathrm{MMD}}$ defined in (4.3).

$$\mathcal{D}^{\mu,\Psi,\lambda}_{\mathrm{SMMD}}(\mathbb{P},\mathbb{Q}) := \sup_{\psi\in\Psi} \mathrm{SMMD}_{\mu,k_\psi,\lambda}(\mathbb{P},\mathbb{Q}).$$

Figure 4.1 (b) shows the vector field for the Optimized SMMD loss in Example 1, using the WGAN-GP measure $\mu = \mathrm{Uniform}(0,\theta)$. The optimization surface is far more amenable: in particular the location C, which formerly had an extremely small gradient that made learning effectively impossible, now converges very quickly by first reducing the critic gradient until some signal is available. Figure 4.1 (d) demonstrates that $\mathcal{D}^{\mu,\Psi,\lambda}_{\mathrm{SMMD}}$, like $\mathcal{D}^{\mu,\Psi,\lambda}_{\mathrm{GCMMD}}$ and $\mathcal{D}^{\Psi,\lambda}_{\mathrm{LipMMD}}$ but in sharp contrast to $\mathcal{D}^{\Psi}_{\mathrm{MMD}}$, is continuous with respect to the location θ and provides a strong gradient towards 0.

Comparison of Gradient-Constrained MMD to Scaled MMD. Figure 4.2 shows the behavior of the MMD, the Gradient-Constrained SMMD, and the Scaled MMD when comparing Gaussian distributions. We can see that $\mathrm{MMD} \propto \mathrm{SMMD}$ and the Gradient-Constrained MMD behave similarly in this case, and that optimizing the SMMD and the Gradient-Constrained MMD is also similar. Optimizing the MMD would yield an essentially constant distance.

Continuity of $\mathcal{D}^{\mu,\Psi,\lambda}_{\mathrm{SMMD}}$ in the Wasserstein topology. We can establish that $\mathcal{D}^{\mu,\Psi,\lambda}_{\mathrm{SMMD}}$ is continuous in the Wasserstein topology under some conditions:

Theorem 21. *Let $k_\psi = K \circ \phi_\psi$, with $\phi_\psi : \mathcal{X} \to \mathbb{R}^s$ a fully-connected L-layer*

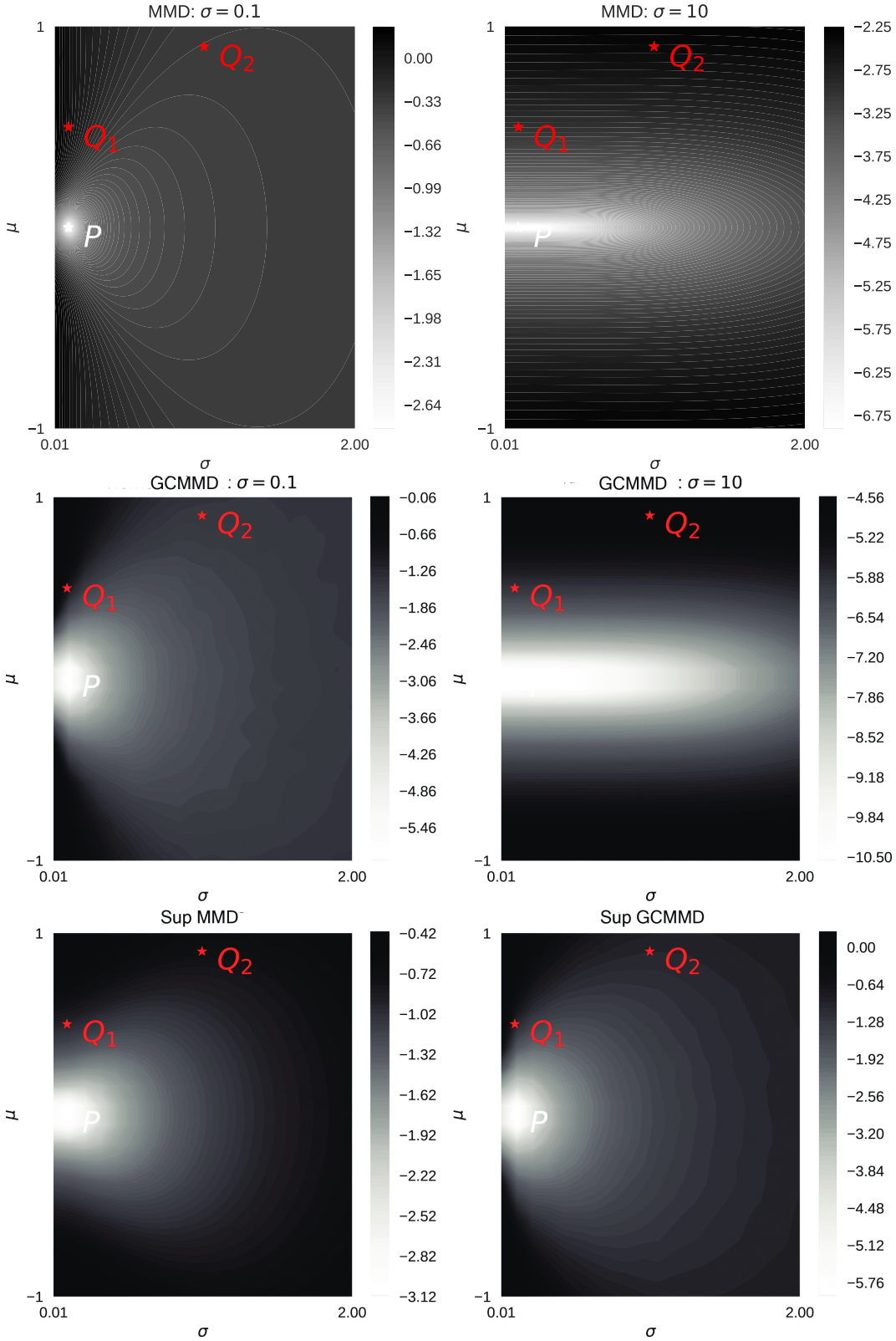


Figure 4.2: Plots of various distances between one dimensional Gaussians, where $P = \mathcal{N}(0, 0.1^2)$, and the colors show $\log \mathcal{D}(P, \mathcal{N}(\mu, \sigma^2))$. All distances use $\lambda = 1$. Top left: MMD with a Gaussian kernel of bandwidth $\psi = 0.1$. Top right: MMD with bandwidth $\psi = 10$. Middle left: Gradient-Constrained MMD with bandwidth $\psi = 0.1$. Middle right: Gradient-Constrained MMD with bandwidth $\psi = 10$. Bottom left: Optimized SMMD, allowing any $\psi \in \mathbb{R}$. Bottom right: Optimized Gradient-Constrained MMD.

*network with Leaky-ReLU$_\alpha$ activations whose layers do not increase in width, and K satisfying mild smoothness conditions $Q_K < \infty$ (Assumptions **(II)** to **(V)** in Section A .2). Let Ψ^κ be the set of parameters where each layer's weight matrices have condition number* $\mathrm{cond}(W^l) = \|W^l\| / \sigma_{\min}(W^l) \leq \kappa < \infty$. *If μ has a density (Assumption **(I)**), then*

$$\mathcal{D}_{\mathrm{SMMD}}^{\mu,\Psi^\kappa,\lambda}(\mathbb{P},\mathbb{Q}) \leq \frac{Q_K \kappa^{L/2}}{\sqrt{d_L}\alpha^{L/2}} \mathcal{W}(\mathbb{P},\mathbb{Q}).$$

Thus if $\mathbb{P}_n \xrightarrow{\mathcal{W}} \mathbb{P}$, then $\mathcal{D}_{\mathrm{SMMD}}^{\mu,\Psi^\kappa,\lambda}(\mathbb{P}_n,\mathbb{P}) \to 0$, even when μ depends on $\mathbb{P}$ and $\mathbb{Q}$.

The assumptions are discussed in more detail in Section A .2; in particular, they require that the network layers never increase in width, as well as requiring μ to have a density on $\mathcal{X}$. The proof gives some insight about the choice of critic family, highlighting the importance of controlling the condition number of the weight matrices per layer for the critic, and requiring that the layers be decreasing in width: $d_l \leq d_{l-1}$. It also requires Leaky-ReLU activations rather than just ReLU in the critic, as previously suggested empirically [Radford et al., 2016]. Section A .2.1 shows counterexamples when the condition number is unbounded or with networks that get wider.

Uniform bounds vs bounds in expectation Controlling the squared $L^2(\mu)$ norm of $\nabla f_\psi(X)$, i.e.: $\|\nabla f_\psi\|^2_{L^2(\mu)} := \mathbb{E}_\mu \|\nabla f_\psi(X)\|^2$ does not necessarily imply a bound on $\|f\|_{\mathrm{Lip}} \geq \sup_{x\in\mathcal{X}} \|\nabla f_\psi(X)\|$, and so does not in general give continuity via Proposition 18. However, Theorem 21 implies that when the network's weights are well-conditioned, it is sufficient to only control $\|\nabla f_\psi\|^2_{L^2(\mu)}$, which is far easier in practice than controlling $\|f\|_{\mathrm{Lip}}$.

If we instead tried to directly controlled $\|f\|_{\mathrm{Lip}}$ with e.g. spectral normalization (SN) [Miyato et al., 2018], we could significantly reduce the expressiveness of the parametric family. In Example 1, constraining $\|\phi_\psi\|_{\mathrm{Lip}} = 1$ limits us to only $\Psi = \{1\}$. Thus $\mathcal{D}_{\mathrm{MMD}}^{\{1\}}$ is simply the MMD with an RBF kernel of bandwidth 1, which has poor gradients when θ is far from 0 (Figure 4.1 (c), blue). The Cauchy-Schwartz bound of Proposition 20 allows jointly adjusting the smoothness of k_ψ and

the critic f, while SN must control the two independently. Relatedly, limiting $\|\phi\|_{\text{Lip}}$ by limiting the Lipschitz norm of each layer could substantially reduce capacity, while $\|\nabla f_\psi\|_{L^2(\mu)}$ need not be decomposed by layer. Another advantage is that μ provides a data-dependent measure of complexity as in Bousquet et al. [2004]: we do not needlessly prevent ourselves from using critics that behave poorly only far from the data.

Spectral parametrization When the generator is near a local optimum, the critic might identify only one direction on which $\mathbb{Q}_\theta$ and $\mathbb{P}$ differ. If the generator parameterization is such that there is no local way for the generator to correct it, the critic may begin to single-mindedly focus on this difference, choosing redundant convolutional filters and causing the condition number of the weights to diverge. If this occurs, the generator will be motivated to fix this single direction while ignoring all other aspects of the distributions, after which it may become stuck. We can help avoid this collapse by using a critic parameterization that encourages diverse filters with higher-rank weight matrices. Miyato et al. [2018] propose to parameterize the weight matrices as $W = \gamma \bar{W}/\|\bar{W}\|_{\text{op}}$, where $\|\bar{W}\|_{\text{op}}$ is the spectral norm of $\bar{W}$. This parametrization works particularly well with $\mathcal{D}^{\mu,\Psi,\lambda}_{\text{SMMD}}$; Figure 4.3 (b) shows the singular values of the second layer of a critic's network (and Figure 4.5, in Section 4 .1.1, shows more layers), while Figure 4.3 (d) shows the evolution of the condition number during training. The conditioning of the weight matrix remains stable throughout training with spectral parametrization, while it worsens through training in the default case.

4 Experiments

4 .1 Image generation

We evaluated unsupervised image generation on three datasets: CIFAR-10 Krizhevsky [2009] (60 000 images, 32×32), CelebA Liu et al. [2015a] (202 599 face images, resized and cropped to 160×160 as in Bińkowski* et al. [2018]), and the more challenging ILSVRC2012 (ImageNet) dataset Russakovsky et al. [2014] (1 281 167 images, resized to 64×64).

Losses All models are based on a scalar-output critic network $\phi_\psi : \mathcal{X} \to \mathbb{R}$, except MMDGAN-GP where $\phi_\psi : \mathcal{X} \to \mathbb{R}^{16}$ as in Bińkowski* et al. [2018]. We didn't notice major differences by choosing different output dimensions for ϕ_ψ. The WGAN and Sobolev GAN use a critic $f = \phi_\psi$, while the GAN uses a discriminator $D_\psi(x) = 1/(1 + \exp(-\phi_\psi(x)))$. The MMD-based methods use a kernel $k_\psi(x, y) = \exp(-(\phi_\psi(x) - \phi_\psi(y))^2/2)$, except for MMDGAN-GP which uses a mixture of RQ kernels as in Bińkowski* et al. [2018]. Increasing the output dimension of the critic or using a different kernel didn't substantially change the performance of our proposed method. We also consider SMMD with a linear top-level kernel, $k(x, y) = \phi_\psi(x)\phi_\psi(y)$; because this becomes essentially identical to a WGAN (Section C), we refer to this method as SWGAN. SMMD and SWGAN use $\mu = \mathbb{P}$; Sobolev GAN uses $\mu = (\mathbb{P} + \mathbb{Q})/2$ as in Mroueh et al. [2018]. We choose λ and an overall scaling to obtain the losses:

$$\text{SMMD: } \frac{\widehat{\text{MMD}}^2_{k_\psi}(\mathbb{P}, \mathbb{Q}_\theta)}{1 + 10\mathbb{E}_{\hat{\mathbb{P}}}\left[\|\nabla\phi_\psi(X)\|_F^2\right]},$$
$$\text{SWGAN: } \frac{\mathbb{E}_{\hat{\mathbb{P}}}\left[\phi_\psi(X)\right] - \mathbb{E}_{\hat{\mathbb{Q}}_\theta}\left[\phi_\psi(X)\right]}{\sqrt{1 + 10\left(\mathbb{E}_{\hat{\mathbb{P}}}\left[|\phi_\psi(X)|^2\right] + \mathbb{E}_{\hat{\mathbb{P}}}\left[\|\nabla\phi_\psi(X)\|_F^2\right]\right)}}.$$

Architecture For CIFAR-10, we used the CNN architecture proposed by Miyato et al. [2018] with a 7-layer critic and a 4-layer generator. For CelebA, we used a 5-layer DCGAN discriminator and a 10-layer ResNet generator as in Bińkowski* et al. [2018]. For ImageNet, we used a 10-layer ResNet for both the generator and discriminator. In all experiments we used 64 filters for the smallest convolutional layer, and double it at each layer (CelebA/ImageNet) or every other layer (CIFAR-10). The input codes for the generator are drawn from $\text{Uniform}\left([-1, 1]^{128}\right)$ as commonly used in the prior works. Note that the dimension of the latent is typically orders of magnitude smaller than the dimension of the image $32 * 32$ and thus the model is assuming a small intrinsic dimension of the data. We consider two parameterizations for each critic: a standard one where the parameters can take any real value, and a spectral parametrization (denoted SN-) as above Miyato et al. [2018]. Models without explicit gradient control (SN-GAN, SN-MMDGAN,

SN-MMGAN-L2, SN-WGAN) fix $\gamma = 1$, for spectral normalization; others learn γ, using a spectral parameterization.

Training All models were trained for $150\,000$ generator updates on a single GPU, except for ImageNet where the model was trained on 3 GPUs simultaneously. To limit communication overhead we averaged the MMD estimate on each GPU, giving the block MMD estimator Zaremba et al. [2013]. We always used 64 samples per GPU from each of $\mathbb{P}$ and $\mathbb{Q}$, and 5 critic updates per generator step. We used initial learning rates of 0.0001 for CIFAR-10 and CelebA, 0.0002 for ImageNet, and decayed these rates using the KID adaptive scheme of Bińkowski* et al. [2018]: every $2\,000$ steps, generator samples are compared to those from $20\,000$ steps ago, and if the relative KID test Bounliphone et al. [2016] fails to show an improvement three consecutive times, the learning rate is decayed by 0.8. We used the Adam optimizer Kingma and Ba [2015] with $\beta_1 = 0.5$, $\beta_2 = 0.9$.

Evaluation To compare the sample quality of different models, we considered three different scores based on the Inception network Szegedy et al. [2016] trained for ImageNet classification, all using default parameters in the implementation of Bińkowski* et al. [2018]. The *Inception Score (IS)* Salimans et al. [2016] is based on the entropy of predicted labels; higher values are better. Though standard, this metric has many issues, particularly on datasets other than ImageNet [Barratt and Sharma, 2018, Heusel et al., 2017, Bińkowski* et al., 2018]. The *FID* Heusel et al. [2017] instead measures the similarity of samples from the generator and the target as the Wasserstein-2 distance between Gaussians fit to their intermediate representations. It is more sensible than the IS and becoming standard, but its estimator is strongly biased Bińkowski* et al. [2018]. The *KID* Bińkowski* et al. [2018] is similar to FID, but by using a polynomial-kernel MMD its estimates enjoy better statistical properties and are easier to compare. (A similar score was recommended by Huang et al. [2018b].)

Results Table 4.1a presents the scores for models trained on both CIFAR-10 and CelebA datasets. On CIFAR-10, SN-SWGAN and SN-SMMDGAN performed comparably to SN-GAN. But on CelebA, SN-SWGAN and SN-SMMDGAN dramat-

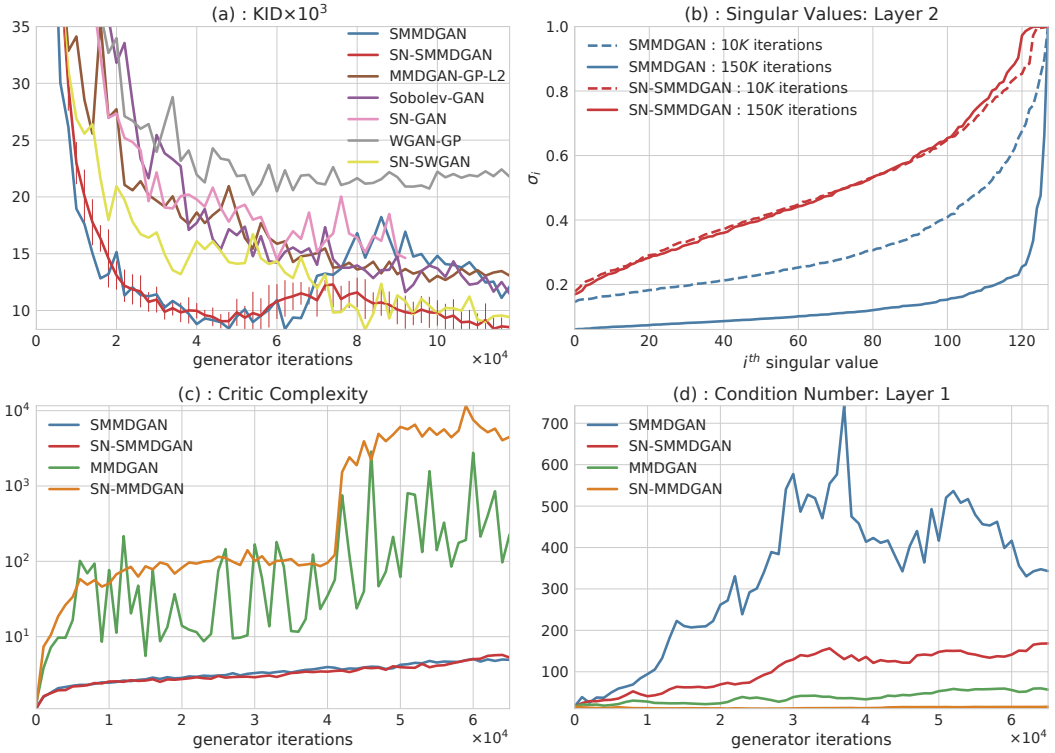


Figure 4.3: The training process on CelebA. (a) KID scores. We report a final score for SN-GAN slightly before its sudden failure mode; MMDGAN and SN-MMDGAN were unstable and had scores around 100. (b) Singular values of the second layer, both early (dashed) and late (solid) in training. (c) $\sigma^{-2}_{\mu,k,\lambda}$ for several MMD-based methods. (d) The condition number in the first layer through training. SN alone does not control $\sigma_{\mu,k,\lambda}$, and SMMD alone does not control the condition number.

ically outperformed the other methods with the same architecture in all three metrics. It also trained faster, and consistently outperformed other methods over multiple initializations (Figure 4.3 (a)). It is worth noting that SN-SWGAN far outperformed WGAN-GP on both datasets. Table 4.1b presents the scores for SMMDGAN and SN-SMMDGAN trained on ImageNet, and the scores of pre-trained models using BGAN [Berthelot et al., 2017] and SN-GAN [Miyato et al., 2018].[4] The proposed methods substantially outperformed both methods in FID and KID scores. Figure 4.4 shows samples on ImageNet and CelebA. **Spectrally normalized WGANs / MMDGANs** To control for the contribution of the spectral parametrization to the

[4]These models are courtesy of the respective authors and also trained at 64×64 resolution. SN-GAN used the same architecture as our model, but trained for 250 000 generator iterations; BS-GAN used a similar 5-layer ResNet architecture and trained for 74 epochs, comparable to SN-GAN.

Table 4.1: Mean (standard deviation) of score estimates, based on 50 000 samples from each model.

(a) CIFAR-10 and CelebA.

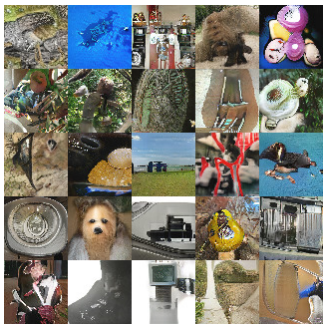
Method	CIFAR-10 IS	CIFAR-10 FID	CIFAR-10 KID$\times 10^3$	CelebA IS	CelebA FID	CelebA KID$\times 10^3$
WGAN-GP	6.9±0.2	31.1±0.2	22.2±1.1	2.7±0.0	29.2±0.2	22.0±1.0
MMDGAN-GP-L2	6.9±0.1	31.4±0.3	23.3±1.1	2.6±0.0	20.5±0.2	13.0±1.0
Sobolev-GAN	7.0±0.1	30.3±0.3	22.3±1.2	**2.9±0.0**	16.4±0.1	10.6±0.5
SMMDGAN	7.0±0.1	31.5±0.4	22.2±1.1	2.7±0.0	18.4±0.2	11.5±0.8
SN-GAN	**7.2±0.1**	26.7±0.2	**16.1±0.9**	2.7±0.0	22.6±0.1	14.6±1.1
SN-SWGAN	**7.2±0.1**	28.5±0.2	**17.6±1.1**	2.8±0.0	14.1±0.2	7.7±0.5
SN-SMMDGAN	**7.3±0.1**	**25.0±0.3**	**16.6±2.0**	2.8±0.0	**12.4±0.2**	**6.1±0.4**

(b) ImageNet.

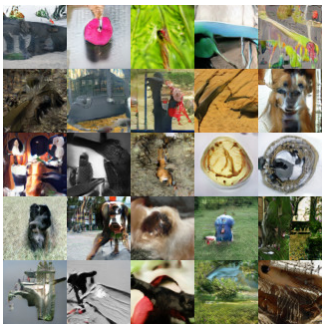
Method	IS	FID	KID$\times 10^3$
BGAN	10.7±0.4	43.9±0.3	47.0±1.1
SN-GAN	**11.2±0.1**	47.5±0.1	44.4±2.2
SMMDGAN	10.7±0.2	38.4±0.3	39.3±2.5
SN-SMMDGAN	10.9±0.1	**36.6±0.2**	**34.6±1.6**

performance, we evaluated variants of MMDGANs, WGANs and Sobolev-GAN using spectral normalization (in Table 4.2, Section 4 .1.1). WGAN and Sobolev-GAN led to unstable training and didn't converge at all (Figure 4.7) despite many attempts. MMDGAN converged on CIFAR-10 (Figure 4.7) but was unstable on CelebA (Figure 4.6). The gradient control due to SN is thus probably too loose for these methods. This is reinforced by Figure 4.3 (c), which shows that the expected gradient of the critic network is much better-controlled by SMMD, even when SN is used. We also considered variants of these models with a learned γ while also adding a gradient penalty and an L_2 penalty on critic activations [Bińkowski* et al., 2018, footnote 19]. These generally behaved similarly to MMDGAN, and didn't lead to substantial improvements. We ran the same experiments on CelebA, but aborted the runs early when it became clear that training was not successful.

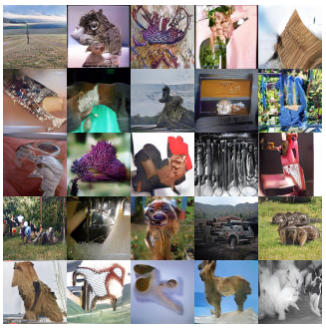
Rank collapse We occasionally observed the failure mode for SMMD where the critic becomes low-rank, discussed in Section 3 .3, especially on CelebA; this failure was obvious even in the training objective. Figure 4.3 (b) is one of these examples. Spectral parametrization seemed to prevent this behavior. We also found



(a) Scaled MMD GAN with SN



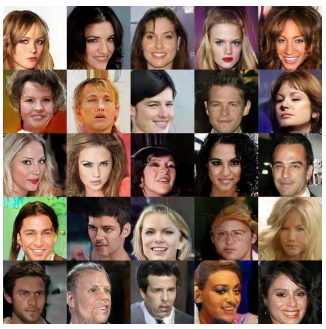
(b) SN-GAN



(c) Boundary Seeking GAN

(d) Scaled MMD GAN with SN

(e) Scaled WGAN with SN



(f) MMD GAN with GP+L2

Figure 4.4: Samples from various models. Top: 64×64 ImageNet; bottom: 160×160 CelebA.

one could avoid collapse by reverting to an earlier checkpoint and increasing the RKHS regularization parameter λ, but did not do this for any of the experiments here.

4 .1.1 Spectral normalization and Scaled MMD

Figure 4.5 shows the distribution of critic weight singular values, like Figure 4.3, at more layers. Figure 4.7 and Table 4.2 show results for the spectral normalization variants considered in the experiments. MMDGAN, with neither spectral normalization nor a gradient penalty, did surprisingly well in this case, though it fails badly in other situations.

Figure 4.5 compares the decay of singular values for layer of the critic's network at both early and later stages of training in two cases: with or without the spectral parametrization. The model was trained on CelebA using SMMD. Figure 4.7 shows the evolution per iteration of Inception score, FID and KID for Sobolev-GAN,

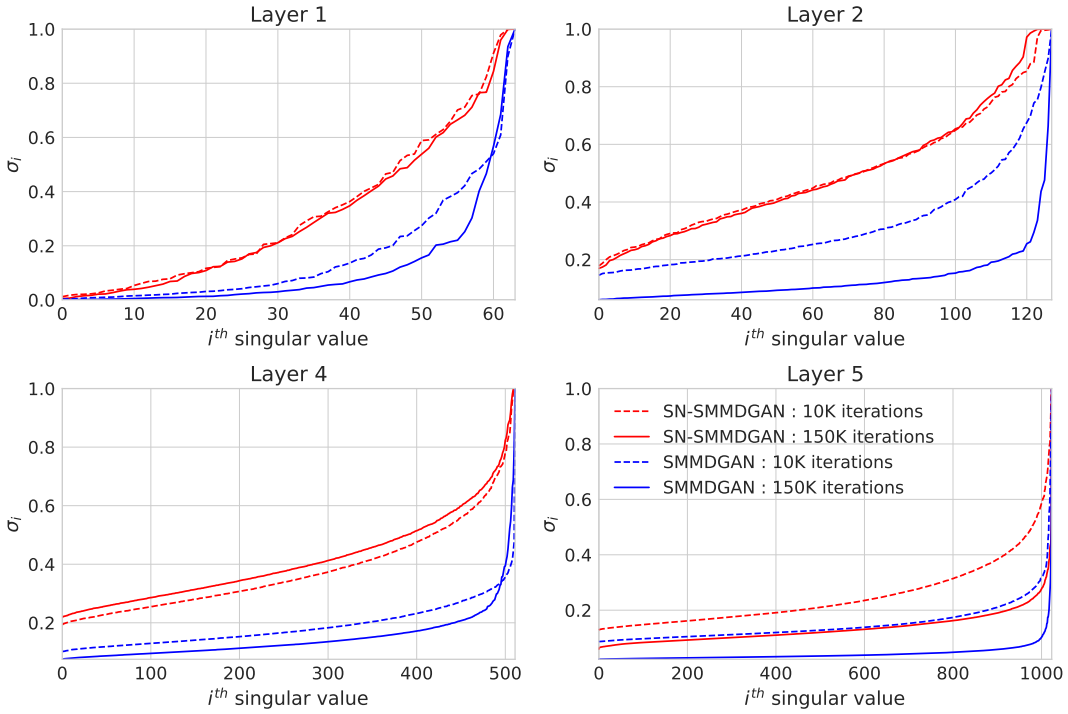


Figure 4.5: Singular values at different layers, for the same setup as Figure 4.3.

Table 4.2: Mean (standard deviation) of score evaluations on CIFAR-10 for different methods using Spectral Normalization.

Method	IS	FID	KID$\times 10^3$
MMDGAN	5.5±0.0	73.9±0.1	39.4±1.5
SN-WGAN	2.2±0.0	208.5±0.2	178.9±1.5
SN-WGAN-GP	2.5±0.0	154.3±0.2	125.3±0.9
SN-Sobolev-GAN	2.9±0.0	140.2±0.2	130.0±1.9
SN-MMDGAN-GP	4.6±0.1	96.8±0.4	59.5±1.4
SN-MMDGAN-L2	7.1±0.1	31.9±0.2	21.7±0.9
SN-MMDGAN	6.9±0.1	31.5±0.2	21.7±1.0
SN-MMDGAN-GP-L2	6.9±0.2	32.3±0.3	20.9±1.1
SN-SMMDGAN	**7.3±0.1**	**25.0±0.3**	**16.6±2.0**

MMDGAN and variants of MMDGAN and WGAN using spectral normalization. It is often the case that this parametrization alone is not enough to achieve good results.

4 .1.2 IGMs with Optimized Gradient-Constrained MMD loss

We implemented the estimator of Proposition 19 using the empirical mean estimator of η, and sharing samples for $\mu = \mathbb{P}$. To handle the large but approximately low-rank matrix system, we used an incomplete Cholesky decomposition [Shawe-Taylor and

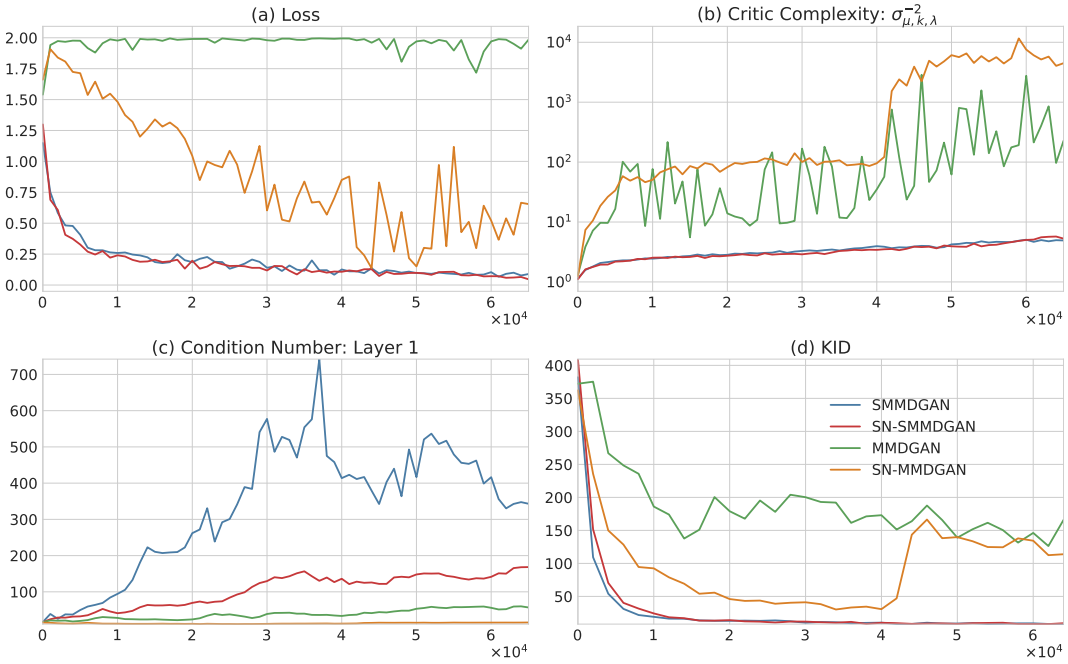


Figure 4.6: Evolution of various quantities per generator iteration on CelebA during training. 4 models are considered: (SMMDGAN, SN-SMMDGAN, MMDGAN, SN-MMDGAN). (a) Loss: $\mathrm{SMMD}^2 = \sigma^2_{\mu,k,\lambda}\,\mathrm{MMD}^2_k$ for SMMDGAN and SN-SMMDGAN, and MMD^2_k for MMDGAN and SN-MMDGAN. The loss saturates for MMDGAN (green); spectral normalization allows some improvement in loss, but training is still unstable (orange). SMMDGAN and SN-SMMDGAN both lead to stable, fast training (blue and red). (b) SMMD controls the critic complexity well, as expected (blue and red); SN has little effect on the complexity (orange). (c) Ratio of the highest singular value to the smallest for the first layer of the critic network: $\sigma_{\max}/\sigma_{\min}$. SMMD tends to increase the condition number of the weights during training (blue), while SN helps controlling it (red). (d) KID score during training: Only variants using SMMD lead to stable training in this case.

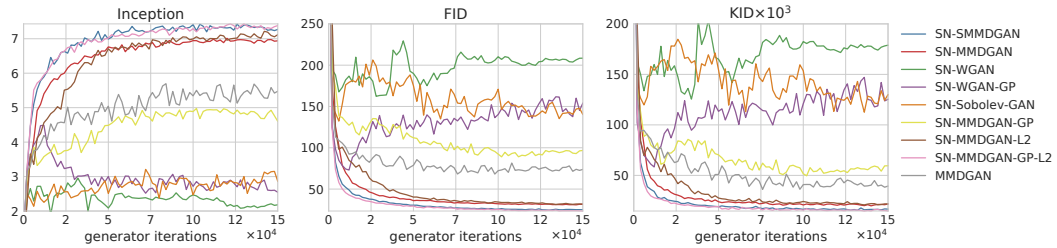


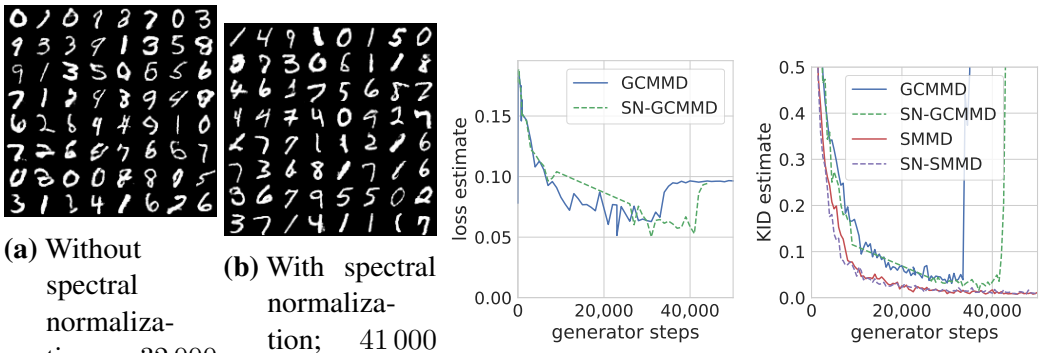
Figure 4.7: Evolution per iteration of different scores for variants of methods, mostly using spectral normalization, on CIFAR-10.

Cristianini, 2004, Algorithm 5.12] to obtain $R \in \mathbb{R}^{\ell \times M(1+d)}$ such that $\begin{bmatrix} K & G^\mathsf{T} \\ G & H \end{bmatrix} \approx R^\mathsf{T} R$. Then the Woodbury matrix identity allows an efficient evaluation:

$$\left(R^\mathsf{T} R + M\lambda I\right)^{-1} = \frac{1}{M\lambda}\left(I - R(RR^\mathsf{T} + M\lambda I)^{-1} R\right).$$

Even though only a small ℓ is required for a good approximation, and the full matrices K, G, and H need never be constructed, backpropagation through this procedure is slow and not especially GPU-friendly; training on CPU was faster. Thus we were only able to run the estimator on MNIST, and even that took days to conduct the optimization on powerful workstations.

The learned models, however, were reasonable. Using a DCGAN architecture, batches of size 64, and a procedure that otherwise agreed with the setup of Section 4 , samples with and without spectral normalization are shown in Figures 4.8a and 4.8b. After the points in training shown, however, the same rank collapse as discussed in Section 4 occurred. Here it seems that spectral normalization may have delayed the collapse, but not prevented it. Figure 4.8c shows generator loss estimates through training, including the obvious peak at collapse; Figure 4.8d shows KID scores based on the MNIST-trained convnet representation [Bińkowski* et al., 2018], including comparable SMMD models for context. The fact that SMMD models converged somewhat faster than Gradient-Constrained MMD models here may be more related to properties of the estimator of Proposition 19 rather than the distances; more work would be needed to fully compare the behavior of the two distances.



(a) Without spectral normalization; 32 000 generator iterations.

(b) With spectral normalization; 41 000 generator iterations.

(c) Generator losses.

(d) KID scores.

Figure 4.8: The MNIST models with Optimized Gradient-Constrained MMD loss.

Supplementary

A Proofs

We first review some basic properties of Reproducing Kernel Hilbert Spaces. We consider here a separable RKHS $\mathcal{H}$ with basis $(e_i)_{i\in I}$, where I is either finite if $\mathcal{H}$ is finite-dimensional, or $I = \mathbb{N}$ otherwise. We also assume that the reproducing kernel k is continuously twice differentiable.

We use a slightly nonstandard notation for derivatives: $\partial_i f(x)$ denotes the ith partial derivative of f evaluated at x, and $\partial_i\partial_{j+d}k(x,y)$ denotes $\frac{\partial^2 k(a,b)}{\partial a_i \partial b_j}|_{(a,b)=(x,y)}$. Then the following reproducing properties hold for any given function f in $\mathcal{H}$ [Steinwart and Christmann, 2008, Lemma 4.34]:

$$f(x) = \langle f, k(x,.)\rangle_{\mathcal{H}} \tag{4.8}$$

$$\partial_i f(x) = \langle f, \partial_i k(x,.)\rangle_{\mathcal{H}}. \tag{4.9}$$

We say that an operator $A : \mathcal{H} \mapsto \mathcal{H}$ is Hilbert-Schmidt if $\|A\|_{HS}^2 = \sum_{i\in I}\|Ae_i\|_{\mathcal{H}}^2$ is finite. $\|A\|_{\mathrm{HS}}$ is called the Hilbert-Schmidt norm of A. The space of Hilbert-Schmidt operators itself a Hilbert space with the inner product $\langle A, B\rangle_{HS} = \sum_{i\in I}\langle Ae_i, Be_i\rangle_{\mathcal{H}}$. Moreover, we say that an operator A is trace-class if its trace norm is finite, i.e. $\|A\|_1 = \sum_{i\in I}\langle e_i, (A^*A)^{\frac{1}{2}}e_i\rangle_{\mathcal{H}} < \infty$. The outer product $f \otimes g$ for $f, g \in \mathcal{H}$ gives an $\mathcal{H} \to \mathcal{H}$ operator such that $(f \otimes g)v = \langle g, v\rangle_{\mathcal{H}} f$ for all v in $\mathcal{H}$.

Given two vectors f and g in $\mathcal{H}$ and a Hilbert-Schmidt operator A we have the following properties:

(i) The outer product $f \otimes g$ is a Hilbert-Schmidt operator with Hilbert-Schmidt norm given by: $\|f \otimes g\|_{\mathrm{HS}} = \|f\|_{\mathcal{H}}\|g\|_{\mathcal{H}}$.

(ii) The inner product between two rank-one operators $f \otimes g$ and $u \otimes v$ is $\langle f \otimes g, u \otimes v\rangle_{\mathrm{HS}} = \langle f, u\rangle_{\mathcal{H}}\langle g, v\rangle_{\mathcal{H}}$.

(iii) The following identity holds: $\langle f, Ag\rangle_{\mathcal{H}} = \langle f \otimes g, A\rangle_{\mathrm{HS}}$.

Define the following covariance-type operators:

$$D_x = k(x,\cdot) \otimes k(x,\cdot) + \sum_{i=1}^{d} \partial_i k(x,\cdot) \otimes \partial_i k(x,\cdot)$$
$$D_\mu = \mathbb{E}_{X\sim\mu} D_X \quad D_{\mu,\lambda} = D_\mu + \lambda I; \tag{4.10}$$

these are useful in that, using (4.8) and (4.9) it follows:

$$\langle f, D_x g\rangle = f(x)g(x) + \sum_{i=1}^{d} \partial_i f(x)\, \partial_i g(x).$$

A .1 Definitions and estimators of the new distances

We will need the following assumptions about the distributions $\mathbb{P}$ and $\mathbb{Q}$, the measure μ, and the kernel k:

(A) $\mathbb{P}$ and $\mathbb{Q}$ have integrable first moments.

(B) $\sqrt{k(x,x)}$ grows at most linearly in x: for all x in $\mathcal{X}$, $\sqrt{k(x,x)} \leq C(\|x\| + 1)$ for some constant C.

(C) The kernel k is twice continuously differentiable.

(D) The functions $x \mapsto k(x,x)$ and $x \mapsto \partial_i\partial_{i+d}k(x,x)$ for $1 \leq i \leq d$ are μ-integrable.

When $k = K \circ \phi_\psi$, Assumption **(B)** is automatically satisfied by a K such as the Gaussian; when K is linear, it is true for a quite general class of networks ϕ_ψ [Bińkowski* et al., 2018, Lemma 1].

We will first give a form for the Gradient-Constrained MMD (4.5) in terms of the operator (4.10):

Proposition 22. *Under Assumptions (A) to (D), the Gradient-Constrained MMD is given by*

$$\mathrm{GCMMD}_{\mu,k,\lambda}(\mathbb{P},\mathbb{Q}) = \sqrt{\langle \eta, D_{\mu,\lambda}^{-1}\eta\rangle_{\mathcal{H}}}.$$

Proof of Proposition 22. Let f be a function in $\mathcal{H}$. We will first express the squared λ-regularized Sobolev norm of f (4.6) as a quadratic form in $\mathcal{H}$. Recalling the reproducing properties of (4.8) and (4.9), we have:

$$\|f\|^2_{S(\mu),k,\lambda} = \int \langle f, k(x,\cdot)\rangle^2_{\mathcal{H}}\ \mu(\mathrm{d}x) + \sum_{i=1}^{d}\int \langle f, \partial_i k(x,\cdot)\rangle^2_{\mathcal{H}}\ \mu(\mathrm{d}x) + \lambda\|f\|^2_{\mathcal{H}}.$$

Using Property (ii) and the operator (4.10), one further gets

$$\|f\|^2_{S(\mu),k,\lambda} = \int \langle f\otimes f, D_x\rangle_{\mathrm{HS}}\ \mu(\mathrm{d}x) + \lambda\|f\|^2_{\mathcal{H}}.$$

Under Assumption (D), and using Lemma 24, one can take the integral inside the inner product, which leads to $\|f\|^2_{S(\mu),k,\lambda} = \langle f\otimes f, D_\mu\rangle_{\mathrm{HS}} + \lambda\|f\|^2_{\mathcal{H}}$. Finally, using Property (iii) it follows that

$$\|f\|^2_{S(\mu),k,\lambda} = \langle f, D_{\mu,\lambda} f\rangle_{\mathcal{H}}.$$

Under Assumptions (A) and (B), Lemma 24 applies, and it follows that $k(x,\cdot)$ is also Bochner integrable under $\mathbb{P}$ and $\mathbb{Q}$. Thus

$$\mathbb{E}_{\mathbb{P}}\left[\langle f, k(x,\cdot)\rangle_{\mathcal{H}}\right] - \mathbb{E}_{\mathbb{Q}}\left[\langle f, k(x,\cdot)\rangle_{\mathcal{H}}\right] = \langle f, \mathbb{E}_{\mathbb{P}}\left[k(x,\cdot)\right] - \mathbb{E}_{\mathbb{P}}\left[k(x,\cdot)\right]\rangle_{\mathcal{H}} = \langle f,\eta\rangle_{\mathcal{H}},$$

where η is defined as this difference in mean embeddings.

Since $D_{\mu,\lambda}$ is symmetric positive definite, its square-root $D^{\frac{1}{2}}_{\mu,\lambda}$ is well-defined and is also invertible. For any $f\in\mathcal{H}$, let $g = D^{\frac{1}{2}}_{\mu,\lambda} f$, so that $\langle f, D_{\mu,\lambda} f\rangle_{\mathcal{H}} = \|g\|^2_{\mathcal{H}}$. Note that for any $g\in\mathcal{H}$, there is a corresponding $f = D^{-\frac{1}{2}}_{\mu,\lambda} g$. Thus we can re-express

the maximization problem in (4.5) in terms of g:

$$\begin{aligned}\mathrm{GCMMD}_{\mu,k,\lambda}(\mathbb{P},\mathbb{Q}) &:= \sup_{\substack{f\in\mathcal{H}\\ \langle f, D_{\mu,\lambda}f\rangle_{\mathcal{H}}\le 1}} \langle f,\eta\rangle_{\mathcal{H}} = \sup_{\substack{g\in\mathcal{H}\\ \|g\|_{\mathcal{H}}\le 1}} \langle D_{\mu,\lambda}^{-\frac{1}{2}} g,\eta\rangle_{\mathcal{H}} \\ &= \sup_{\substack{g\in\mathcal{H}\\ \|g\|_{\mathcal{H}}\le 1}} \langle g, D_{\mu,\lambda}^{-\frac{1}{2}}\eta\rangle_{\mathcal{H}} = \|D_{\mu,\lambda}^{-\frac{1}{2}}\eta\|_{\mathcal{H}} = \sqrt{\langle \eta, D_{\mu,\lambda}^{-1}\eta\rangle_{\mathcal{H}}}. \quad \square\end{aligned}$$

Proposition 22, though, involves inverting the infinite-dimensional operator $D_{\mu,\lambda}$ and thus doesn't directly give us a computable estimator. Proposition 19 solves this problem in the case where μ is a discrete measure:

Before proving Proposition 19, we note the following interesting alternate form. Let $\bar{e}_i$ be the ith standard basis vector for $\mathbb{R}^{M+Md}$, and define $T : \mathcal{H} \to \mathbb{R}^{M+Md}$ as the linear operator

$$T = \sum_{m=1}^{M} \bar{e}_m \otimes k(X_m,\cdot) + \sum_{m=1}^{M}\sum_{i=1}^{d} \bar{e}_{m+(m,i)} \otimes \partial_i k(X_m,\cdot).$$

Then $\begin{bmatrix}\eta(X)\\ \nabla\eta(X)\end{bmatrix} = T\eta$, and $\begin{bmatrix}K & G^{\mathsf{T}}\\ G & H\end{bmatrix} = TT^*$. Thus we can write

$$\mathrm{GCMMD}^2_{\hat{\mu},k,\lambda} = \frac{1}{\lambda}\left\langle \eta, \left(I - T^*(TT^* + M\lambda I)^{-1}T\right)\eta\right\rangle_{\mathcal{H}}.$$

Proof of Proposition 19. Let $g \in \mathcal{H}$ be the solution to the regression problem $D_{\mu,\lambda}g = \eta$:

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}\left[g(X_m)k(X_m,\cdot) + \sum_{i=1}^{d}\partial_i g(X_m)\partial_i k(X_m,\cdot)\right] + \lambda g &= \eta \\ g = \frac{1}{\lambda}\eta - \frac{1}{\lambda M}\sum_{m=1}^{M}\left[g(X_m)k(X_m,\cdot) + \sum_{i=1}^{d}\partial_i g(X_m)\partial_i k(X_m,\cdot)\right]&. \qquad (4.11)\end{aligned}$$

Taking the inner product of both sides of (4.11) with $k(X_{m'},\cdot)$ for each $1 \le m' \le M$

yields the following M aligns:

$$g(X_{m'}) = \frac{1}{\lambda}\eta(X_{m'}) - \frac{1}{\lambda M}\sum_{m=1}^{M}\left[g(X_m)K_{m,m'} + \sum_{i=1}^{d}\partial_i g(X_m)\, G_{(m,i),m'}\right]. \tag{4.12}$$

Doing the same with $\partial_j k(X_{m'}, \cdot)$ gives Md aligns:

$$\partial_j g(X_{m'}) = \frac{1}{\lambda}\partial_j\eta(X_{m'}) - \frac{1}{\lambda M}\sum_{m=1}^{M}\left[g(X_m)G_{(m',j),m} + \sum_{i=1}^{d}\partial_i g(X_m)H_{(m,i),(m',j)}\right]. \tag{4.13}$$

From (4.11), it is clear that g is a linear combination of the form:

$$g(x) = \frac{1}{\lambda}\eta(x) - \frac{1}{\lambda M}\sum_{m=1}^{M}\left[\alpha_m k(X_m, x) + \sum_{i=1}^{d}\beta_{m,i}\partial_i k(X_m, x)\right],$$

where the coefficients $\alpha := (\alpha_m = g(X_m))_{1\le m\le M}$ and $\beta := (\beta_{m,i} = \partial_i g(X_m))_{\substack{1\le m\le M\\ 1\le i\le d}}$ satisfy the system of aligns (4.12) and (4.13). We can rewrite this system as

$$\begin{bmatrix} K + M\lambda I_M & G^\intercal \\ G & H + M\lambda I_{Md}\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = M\begin{bmatrix}\eta(X)\\ \nabla\eta(X)\end{bmatrix},$$

where I_M, I_{Md} are the identity matrices of dimension M, Md. Since K and H must be positive semidefinite, an inverse exists. We conclude by noticing that

$$\mathrm{GCMMD}_{\hat{\mu},k,\lambda}(\mathbb{P},\mathbb{Q})^2 = \langle\eta, g\rangle_{\mathcal{H}} = \frac{1}{\lambda}\|\eta\|^2_{\mathcal{H}} - \frac{1}{\lambda M}\sum_{m=1}^{M}\left[\alpha_m\eta(X_m) + \sum_{i=1}^{d}\beta_{m,i}\partial_i\eta(X_m)\right].$$

The following result was key to our definition of the SMMD in Section 3 .3.

Proposition 23. *Under Assumptions (A) to (D), we have for all $f \in \mathcal{H}$ that*

$$\|f\|_{S(\mu),k,\lambda} \le \sigma^{-1}_{\mu,k,\lambda}\|f\|_{\mathcal{H}_k},$$

where $\sigma_{k,\mu,\lambda} := 1/\sqrt{\lambda + \int k(x,x)\mu(\mathrm{d}x) + \sum_{i=1}^{d}\int \partial_i\partial_{i+d}k(x,x)\mu(\mathrm{d}x)}$.

Proof of Proposition 23. The key idea here is to use the Cauchy-Schwarz inequality for the Hilbert-Schmidt inner product. Letting $f \in \mathcal{H}$, $\|f\|^2_{S(\mu),k,\lambda}$ is

$$\begin{aligned}
&\int f(x)^2 \,\mu(\mathrm{d}x) + \int \|\nabla f(x)\|^2 \,\mu(\mathrm{d}x) + \lambda\|f\|^2_{\mathcal{H}} \\
&\overset{(a)}{=} \int \langle f, k(x,\cdot)\otimes k(x,\cdot) f\rangle_{\mathcal{H}} \,\mu(\mathrm{d}x) \\
&\quad + \sum_{i=1}^{d} \int \langle f, \partial_i k(x,\cdot)\otimes \partial_i k(x,\cdot) f\rangle_{\mathcal{H}} \,\mu(\mathrm{d}x) + \lambda\|f\|^2_{\mathcal{H}} \\
&\overset{(b)}{=} \int \langle f\otimes f, k(x,\cdot)\otimes k(x,\cdot)\rangle_{\mathrm{HS}} \;\mu(\mathrm{d}x) \\
&\quad + \sum_{i=1}^{d} \int \langle f\otimes f, \partial_i k(x,\cdot)\otimes \partial_i k(x,\cdot)\rangle_{\mathrm{HS}} \,\mu(\mathrm{d}x) + \lambda\|f\|^2_{\mathcal{H}} \\
&\overset{(c)}{\leq} \|f\|^2_{\mathcal{H}} \left[\int k(x,x)\,\mu(\mathrm{d}x) + \sum_{i=1}^{d}\int \partial_i\partial_{i+d}k(x,x)\,\mu(\mathrm{d}x) + \lambda\right].
\end{aligned}$$

(a) follows from the reproducing properties (4.8) and (4.9) and Property (ii). (b) is obtained using Property (iii), while (c) follows from the Cauchy-Schwarz inequality and Property (i). □

Lemma 24. *Under Assumption (D), D_x is Bochner integrable and its integral D_μ is a trace-class symmetric positive semi-definite operator with $D_{\mu,\lambda} = D + \lambda I$ invertible for any positive λ. Moreover, for any Hilbert-Schmidt operator A we have: $\langle A, D_\mu\rangle_{HS} = \int \langle A, D_x\rangle_{HS}\,\mu(\mathrm{d}x)$.*

Under Assumptions (A) and (B), $k(x,\cdot)$ is Bochner integrable with respect to any probability distribution $\mathbb{P}$ with finite first moment and the following relation holds: $\langle f, \mathbb{E}_{\mathbb{P}}[k(x,\cdot)]\rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}[\langle f, k(x,\cdot)\rangle_{\mathcal{H}}]$ for all f in $\mathcal{H}$.

Proof. The operator D_x is positive self-adjoint. It is also trace-class, as by the triangle inequality

$$\begin{aligned}
\|D_x\|_1 &\leq \|k(x,\cdot)\otimes k(x,\cdot)\|_1 + \sum_{i=1}^{d}\|\partial_i k(x,\cdot)\otimes\partial_i k(x,\cdot)\|_1 \\
&= \|k(x,\cdot)\|^2_{\mathcal{H}} + \sum_{i=1}^{d}\|\partial_i k(x,\cdot)\|^2_{\mathcal{H}} < \infty.
\end{aligned}$$

By Assumption (D), we have that $\int \|D_x\|_1 \, \mu(\mathrm{d}x) < \infty$ which implies that D_x is μ-integrable in the Bochner sense [Retherford, 1978, Definition 1 and Theorem 2]. Its integral D_μ is trace-class and satisfies $\|D_\mu\|_1 \leq \int \|D_x\|_1 \, \mu(\mathrm{d}x)$. This allows to have $\langle A, D_\mu \rangle_{HS} = \int \langle A, D_x \rangle_{HS} \, \mu(\mathrm{d}x)$ for all Hilbert-Schmidt operators A. Moreover, the integral preserves the symmetry and positivity. It follows that $\mathcal{D}_{\mu,\lambda}$ is invertible.

The Bochner integrability of $k(x, \cdot)$ under a distribution $\mathbb{P}$ with finite moment follows directly from Assumptions (A) and (B), since $\int \|k(x,\cdot)\| \, \mathbb{P}(\mathrm{d}x) \leq C \int (\|x\| + 1) \, \mathbb{P}(\mathrm{d}x) < \infty$. This allows us to write $\langle f, \mathbb{E}_\mathbb{P}[k(x,\cdot)] \rangle_\mathcal{H} = \mathbb{E}_\mathbb{P}[\langle f, k(x,\cdot) \rangle_\mathcal{H}]$. □

A .2 Continuity of the Optimized Scaled MMD in the Wasserstein topology

To prove Theorem 25, we we will first need some new notation. We assume the kernel is $k = K \circ \phi_\psi$, i.e. $k_\psi(x, y) = K(\phi_\psi(x), \phi_\psi(y))$, where the representation function ϕ_ψ is a network $\phi_\psi(X) : \mathbb{R}^d \to \mathbb{R}^{d_L}$ consisting of L fully-connected layers:

$$
\begin{aligned}
h^0_\psi(X) &= X \\
h^l_\psi(X) &= W^l \sigma_{l-1}(h^{l-1}_\psi(X)) + b^l \qquad \text{for } 1 \leq l \leq L \\
\phi_\psi(X) &= h^L_\psi(X).
\end{aligned}
$$

The intermediate representations $h^l_\psi(X)$ are of dimension d_l, the weights W^l are matrices in $\mathbb{R}^{d_l \times d_{l-1}}$, and biases b^l are vectors in $\mathbb{R}^{d_l}$. The elementwise activation function σ is given by $\sigma_0(x) = x$, and for $l > 0$ the activation σ_l is a leaky ReLU with leak coefficient $0 < \alpha < 1$:

$$
\sigma_l(x) = \sigma(x) = \begin{cases} x & x > 0 \\ \alpha x & x \leq 0 \end{cases} \qquad \text{for } l > 0. \tag{4.14}
$$

The parameter ψ is the concatenation of all the layer parameters:

$$
\psi = \left((W^L, b^L), (W^{L-1}, b^{L-1}), \ldots, (W^1, b^1) \right).
$$

We denote by Ψ the set of all such possible parameters, i.e. $\Psi = \mathbb{R}^{d_L \times d_{L-1}} \times \mathbb{R}^{d_L} \times \cdots \times \mathbb{R}^{d_1 \times d} \times \mathbb{R}^{d_1}$. Define the following restrictions of Ψ:

$$\begin{aligned} \Psi^\kappa &:= \left\{ \psi \in \Psi \mid \forall 1 \leq l \leq L,\ \mathrm{cond}(W^l) \leq \kappa \right\} \\ \Psi_1^\kappa &:= \left\{ \psi \in \Psi^\kappa \mid \forall 1 \leq l \leq L,\ \|W^l\| = 1 \right\}. \end{aligned} \tag{4.15}$$

Ψ^κ is the set of those parameters such that W^l have a small condition number, $\mathrm{cond}(W) = \sigma_{\max}(W)/\sigma_{\min}(W)$. Ψ_1^κ is the set of per-layer normalized parameters with a condition number bounded by κ.

Recall the definition of Scaled MMD, (4.7), where $\lambda > 0$ and μ is a probability measure:

$$\begin{aligned} \mathrm{SMMD}_{\mu,k,\lambda}(\mathbb{P}, \mathbb{Q}) &:= \sigma_{\mu,k,\lambda}\, \mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) \\ \sigma_{k,\mu,\lambda} &:= 1/\sqrt{\lambda + \int k(x,x)\, \mu(\mathrm{d}x) + \sum_{i=1}^{d} \int \partial_i \partial_{i+d} k(x,x)\, \mu(\mathrm{d}x)}. \end{aligned}$$

The Optimized SMMD over the restricted set Ψ^κ is given by:

$$\mathcal{D}^{\mu,\Psi^\kappa,\lambda}_{\mathrm{SMMD}}(\mathbb{P}, \mathbb{Q}) := \sup_{\psi \in \Psi^\kappa} \mathrm{SMMD}_{\mu,k_\psi,\lambda}\,.$$

The constraint to $\psi \in \Psi^\kappa$ is critical to the proof. In practice, using a spectral parametrization helps enforce this assumption, as shown in Figures 4.3 and 4.5. Other regularization methods, like orthogonal normalization Brock et al. [2017], are also possible.

We will use the following assumptions:

(I) μ is a probability distribution absolutely continuous with respect to the Lebesgue measure.

(II) The dimensions of the weights are decreasing per layer: $d_{l+1} \leq d_l$ for all $0 \leq l \leq L-1$.

(III) The non-linearity used is Leaky-ReLU, (4.14), with leak coefficient $\alpha \in (0, 1)$.

(**IV**) The top-level kernel K is globally Lipschitz in the RKHS norm: there exists a positive constant $L_K > 0$ such that $\|K(a, .) - K(b, .)\| \leq L_K \|a - b\|$ for all a and b in $\mathbb{R}^{d_L}$.

(**V**) There is some $\gamma_K > 0$ for which K satisfies

$$\nabla_b \nabla_c K(b, c)\big|_{(b,c)=(a,a)} \succeq \gamma^2 I \qquad \text{for all } a \in \mathbb{R}^{d_L}.$$

Assumption (I) ensures that the points where $\phi_\psi(X)$ is not differentiable are reached with probability 0 under μ. This assumption can be easily satisfied e.g. if we define μ by adding Gaussian noise to $\mathbb{P}$.

Assumption (II) helps ensure that the span of W^l is never contained in the null space of W^{l+1}. Using Leaky-ReLU as a non-linearity, Assumption (III), further ensures that the network ϕ_ψ is locally full-rank almost everywhere; this might not be true with ReLU activations, where it could be always 0. Assumptions (II) and (III) can be easily satisfied by design of the network.

Assumptions (IV) and (V) only depend on the top-level kernel K and are easy to satisfy in practice. In particular, they always hold for a smooth translation-invariant kernel, such as the Gaussian, as well as the linear kernel.

We are now ready to prove Theorem 25.

Theorem 25. *Under Assumptions (I) to (V),*

$$\mathcal{D}^{\mu,\Psi^\kappa,\lambda}_{\mathrm{SMMD}}(\mathbb{P}, \mathbb{Q}) \leq \frac{L_K \, \kappa^{L/2}}{\gamma \sqrt{d_L} \, \alpha^{L/2}} \, \mathcal{W}(\mathbb{P}, \mathbb{Q}),$$

which implies that if $\mathbb{P}_n \xrightarrow{\mathcal{W}} \mathbb{P}$, *then* $\mathcal{D}^{\mu,\Psi^\kappa,\lambda}_{\mathrm{SMMD}}(\mathbb{P}_n, \mathbb{P}) \to 0$.

Proof. Define the pseudo-distance corresponding to the kernel k_ψ

$$d_\psi(x, y) = \|k_\psi(x, \cdot) - k_\psi(y, \cdot)\|_{\mathcal{H}_\psi} = \sqrt{k_\psi(x, x) + k_\psi(y, y) - 2k_\psi(x, y)}.$$

Denote by $\mathcal{W}_{d_\psi}(\mathbb{P}, \mathbb{Q})$ the optimal transport metric between $\mathbb{P}$ and $\mathbb{Q}$ using the cost

d_ψ, given by

$$\mathcal{W}_{d_\psi}(\mathbb{P},\mathbb{Q}) = \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(X,Y)\sim\pi}\left[d_\psi(X,Y)\right].$$

where Π is the set of couplings with marginals $\mathbb{P}$ and $\mathbb{Q}$. By Lemma 26,

$$\mathrm{MMD}_\psi(\mathbb{P},\mathbb{Q}) \le \mathcal{W}_{d_\psi}(\mathbb{P},\mathbb{Q}).$$

Recall that ϕ_ψ is Lipschitz, $\|\phi_\psi\|_{\mathrm{Lip}} < \infty$, so along with Assumption **(IV)** we have that

$$d_\psi(x,y) \le L_K\|\phi_\psi(x) - \phi_\psi(y)\| \le L_K\|\phi_\psi\|_{\mathrm{Lip}}\|x-y\|.$$

Thus

$$\mathcal{W}_{d_\psi}(\mathbb{P},\mathbb{Q}) \le \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(X,Y)\sim\pi}\left[L_K\|\phi_\psi\|_{\mathrm{Lip}}\|X-Y\|\right] = L_K\|\phi_\psi\|_{\mathrm{Lip}}\,\mathcal{W}(\mathbb{P},\mathbb{Q}),$$

where $\mathcal{W}$ is the standard Wasserstein distance (4.2), and so

$$\mathrm{MMD}_\psi(\mathbb{P},\mathbb{Q}) \le L_k\|\phi_\psi\|_{\mathrm{Lip}}\,\mathcal{W}(\mathbb{P},\mathbb{Q}).$$

We have that:

$$\partial_i\partial_{i+d}k(x,y) = \left[\partial_i\phi_\psi(x)\right]^\mathsf{T}\left[\nabla_a\nabla_b K(a,b)\big|_{(a,b)=(\phi_\psi(x),\phi_\psi(y))}\right]\left[\partial_i\phi_\psi(y)\right],$$

where the middle term is a $d_L \times d_L$ matrix and the outer terms are vectors of length d_L. Thus Assumption **(V)** implies that $\partial_i\partial_{i+d}k(x,x) \ge \gamma_K^2\|\partial_i\phi_\psi(x)\|^2$, and hence

$$\sigma_{\mu,k,\lambda}^{-2} \ge \gamma_K^2\mathbb{E}[\|\nabla\phi_\psi(X)\|_F^2]$$

so that

$$\mathrm{SMMD}_\psi^2(\mathbb{P},\mathbb{Q}) = \sigma_{\mu,k,\lambda}^2\,\mathrm{MMD}_\psi^2(\mathbb{P},\mathbb{Q}) \le \frac{L_K^2\|\phi_\psi\|_{\mathrm{Lip}}^2}{\gamma_K^2\,\mathbb{E}\left[\|\nabla\phi_\psi(X)\|_F^2\right]}\,\mathcal{W}^2(\mathbb{P},\mathbb{Q}).$$

Using Lemma 27, we can write $\phi_\psi(X) = \alpha(\psi)\phi_{\bar\psi}(X)$ with $\bar\psi \in \Psi_1^\kappa$. Then we

have

$$\frac{\|\phi_\psi\|_{\mathrm{Lip}}^2}{\mathbb{E}_\mu\left[\|\nabla\phi_\psi(X)\|_F^2\right]} = \frac{\alpha(\psi)^2\|\phi_{\bar\psi}\|_{\mathrm{Lip}}^2}{\alpha(\psi)^2\mathbb{E}_\mu\left[\left\|\nabla\phi_{\bar\psi}(X)\right\|_F^2\right]} \leq \frac{1}{\mathbb{E}_\mu\left[\left\|\nabla\phi_{\bar\psi}(X)\right\|_F^2\right]},$$

where we used $\|\phi_{\bar\psi}\|_{\mathrm{Lip}} \leq \prod_{l=1}^L \|\bar W^l\| = 1$. But by Lemma 28, for Lebesgue-almost all X, $\|\nabla\phi_{\bar\psi}(X)\|_F^2 \geq d_L(\alpha/\kappa)^L$. Using Assumption (I), this implies that

$$\frac{\|\phi_\psi\|_{\mathrm{Lip}}^2}{\mathbb{E}_\mu\left[\|\nabla\phi_\psi(X)\|_F^2\right]} \leq \frac{1}{\mathbb{E}_\mu\left[\|\nabla\phi_{\bar\psi}(X)\|_F^2\right]} \leq \frac{\kappa^L}{d_L\alpha^L}.$$

Thus for any $\psi \in \Psi^\kappa$,

$$\mathrm{SMMD}_\psi(\mathbb{P},\mathbb{Q}) \leq \frac{L_K\,\kappa^{L/2}}{\gamma_K\sqrt{d_L}\,\alpha^{L/2}}\,\mathcal{W}(\mathbb{P},\mathbb{Q}).$$

The desired bound on $\mathcal{D}_{\mathrm{SMMD}}^{\mu,\Psi^\kappa,\lambda}$ follows immediately. □

Lemma 26. *Let $(x,y) \mapsto k(x,y)$ be the continuous kernel of an RKHS $\mathcal{H}$ defined on a Polish space $\mathcal{X}$, and define the corresponding pseudo-distance $d_k(x,y) := \|k(x,\cdot) - k(y,\cdot)\|_{\mathcal{H}}$. Then the following inequality holds for any distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$, including when the quantities are infinite:*

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) \leq \mathcal{W}_{d_k}(\mathbb{P},\mathbb{Q}).$$

Proof. Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions, and let $\Pi(\mathbb{P},\mathbb{Q})$ be the set of couplings between them. Let $\pi^* \in \operatorname{argmin}_{(X,Y)\sim\pi}[c_k(X,Y)]$ be an optimal coupling, which is guaranteed to exist [Villani, 2009, Theorem 4.1]; by definition $\mathcal{W}_{d_k}(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{(X,Y)\sim\pi^*}[d_k(X,Y)]$. When $\mathcal{W}_{d_k}(\mathbb{P},\mathbb{Q}) = \infty$ the inequality trivially holds, so assume that $\mathcal{W}_{d_k}(\mathbb{P},\mathbb{Q}) < \infty$.

Take a sample $(X,Y) \sim \pi^\star$ and a function $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq 1$. By the Cauchy-Schwarz inequality,

$$\|f(X) - f(Y)\| \leq \|f\|_{\mathcal{H}}\|k(X,\cdot) - k(Y,\cdot)\|_{\mathcal{H}} \leq \|k(X,\cdot) - k(Y,\cdot)\|_{\mathcal{H}}.$$

Taking the expectation with respect to $\pi^\star$, we obtain

$$\mathbb{E}_{\pi^\star}[|f(X) - f(Y)|] \leq \mathbb{E}_{\pi^\star}[\|k(X,\cdot) - k(Y,\cdot)\|_{\mathcal{H}}].$$

The right-hand side is just the definition of $\mathcal{W}_{d_k}(\mathbb{P},\mathbb{Q})$. By Jensen's inequality, the left-hand side is lower-bounded by

$$|\mathbb{E}_{\pi^*}[f(X) - f(Y)]| = |\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{Y\sim\mathbb{Q}}[f(Y)]|$$

since $\pi^\star$ has marginals $\mathbb{P}$ and $\mathbb{Q}$. We have shown so far that for any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq 1$,

$$|\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(Y)]| \leq \mathcal{W}_{c_k}(\mathbb{P},\mathbb{Q});$$

the result follows by taking the supremum over f. □

Lemma 27. *Let $\psi = ((W^L, b^L), (W^{L-1}, b^{L-1}), \ldots, (W^1, b^1)) \in \Psi^\kappa$. There exists a corresponding scalar $\alpha(\psi)$ and $\bar{\psi} = ((\bar{W}^L, \bar{b}^L), (\bar{W}^{L-1}, \bar{b}^{L-1}), \ldots, (\bar{W}^1, \bar{b}^1)) \in \Psi_1^\kappa$, defined by (4.15), such that for all X,*

$$\phi_\psi(X) = \alpha(\psi)\, \phi_{\bar{\psi}}(X).$$

Proof. Set $\bar{W}^l = \frac{1}{\|W^l\|} W^l$, $\bar{b}^l = \frac{1}{\prod_{m=1}^l \|W^m\|} b^l$, and $\alpha(\psi) = \prod_{l=1}^L \|W^l\|$. Note that the condition number is unchanged, $\mathrm{cond}(\bar{W}^l) = \mathrm{cond}(W^l) \leq \kappa$, and $\|\bar{W}^l\| = 1$, so $\bar{\psi} \in \Phi_1^\kappa$. It is also easy to see from (4.14) that

$$h_{\bar{\psi}}^l(X) = \frac{1}{\prod_{m=1}^l \|W^m\|} h_\psi^l(X)$$

so that

$$\alpha(\psi) h_{\bar{\psi}}^L(X) = \frac{\prod_{l=1}^L \|W^l\|}{\prod_{l=1}^L \|W^l\|} h_{\bar{\psi}}^L(X) = \phi_\psi(X).$$

Lemma 28. *Make Assumptions **(II)** and **(III)**, and let $\psi \in \Psi_1^\kappa$. Then the set of inputs*

for which any intermediate activation is exactly zero,

$$\mathcal{N}_\psi = \bigcup_{l=1}^{L} \bigcup_{k=1}^{d_l} \left\{ X \in \mathbb{R}^d \mid \left(h_\psi^l(X)\right)_k = 0 \right\},$$

has zero Lebesgue measure. Moreover, for any $X \notin \mathcal{N}_\psi$, $\nabla_X \phi_\psi(X)$ exists and

$$\|\nabla_X \phi_\psi(X)\|_F^2 \geq \frac{d_L \alpha^L}{\kappa^L}.$$

Proof. First, note that the network representation at layer l is piecewise affine. Specifically, define $M_X^l \in \mathbb{R}^{d_l}$ by, using Assumption **(III)**,

$$(M_X^l)_k = \sigma_l'(h_k^l(X)) = \begin{cases} 1 & h_k^l(X) > 0 \\ \alpha & h_k^l(X) < 0 \end{cases};$$

it is undefined when any $h_k^l(X) = 0$, i.e. when $X \in \mathcal{N}_\psi$. Let $V_X^l := W^l \operatorname{diag}\left(M_X^{l-1}\right)$. Then

$$h_\psi^l(X) = W^l \sigma_{l-1}(h_\psi^{l-1}(X)) + b^l = V_X^l X + b^l,$$

and thus

$$h_\psi^l(X) = \underline{\mathbf{W}}_X^l X + \underline{\mathbf{b}}_X^l,$$

where $\underline{\mathbf{b}}_X^0 = 0$, $\underline{\mathbf{b}}_X^l = V_X^l \underline{\mathbf{b}}^{l-1} + b^l$, and $\underline{\mathbf{W}}_X^l = V_X^l V_X^{l-1} \cdots V_X^1$, so long as $X \notin \mathcal{N}_\psi$.

Because $\psi \in \Psi_1^\kappa$, we have $\|W^l\| = 1$ and $\sigma_{\min}(W^l) \geq 1/\kappa$; also, $\|M_X^l\| \leq 1$, $\sigma_{\min}(M_X^l) \geq \alpha$. Thus $\|\underline{\mathbf{W}}_X^l\| \leq 1$, and using Assumption **(II)** with Lemma 29 gives $\sigma_{\min}(\underline{\mathbf{W}}_X^l) \geq (\alpha/\kappa)^l$. In particular, each $\underline{\mathbf{W}}_X^l$ is full-rank.

Next, note that $\underline{\mathbf{b}}_X^l$ and $\underline{\mathbf{W}}_X^l$ each only depend on X through the activation patterns M_X^l. Letting $H_X^l = (M_X^l, M_X^{l-1}, \ldots, M_X^1)$ denote the full activation patterns up to level l, we can thus write

$$h_\psi^l(X) = \underline{\mathbf{W}}^{H_X^l} X + \underline{\mathbf{b}}^{H_X^l}.$$

There are only finitely many possible values for H_X^l; we denote the set of such values as $\mathcal{H}^l$. Then we have that

$$\mathcal{N}_\psi \subseteq \bigcup_{l=0}^{L} \bigcup_{k=1}^{d_L} \bigcup_{H^l \in \mathcal{H}^l} \left\{ X \in \mathbb{R}^d \mid \underline{\mathbf{W}}_k^{H^l} X + \underline{\mathbf{b}}_k^{H^l} = 0 \right\}.$$

Because each $\underline{\mathbf{W}}_k^{H^l}$ is of rank d_l, each set in the union is either empty or an affine subspace of dimension $d - d_l$. As each $d_l > 0$, each set in the finite union has zero Lebesgue measure, and $\mathcal{N}_\psi$ also has zero Lebesgue measure.

We will now show that the activation patterns are piecewise constant, so that $\nabla_X h_\psi^l(X) = \underline{\mathbf{W}}^{H_X^l}$ for all $X \notin \mathcal{N}_\psi$. Because $\psi \in \Psi_1^\kappa$, we have $\|h_\psi^l\|_{\text{Lip}} \leq 1$, and in particular

$$\left| \left(h_\psi^l(X)\right)_k - \left(h_\psi^l(X')\right)_k \right| \leq \|X - X'\|.$$

Thus, take some $X \notin \mathcal{N}_\psi$, and find the smallest absolute value of its activations, $\epsilon = \min_{l=1,\dots,L} \min_{k=1,\dots,d_l} \left| \left(h_\psi^l(X)\right)_k \right|$; clearly $\epsilon > 0$. For any X' with $\|X - X'\| < \epsilon$, we know that for all l and k,

$$\operatorname{sign}\left(\left(h_\psi^l(X)\right)_k \right) = \operatorname{sign}\left(\left(h_\psi^l(X')\right)_k \right),$$

implying that $H_X^l = H_{X'}^l$ as well as $X' \notin \mathcal{N}_\psi$. Thus for any point $X \notin \mathcal{N}_\psi$, $\nabla \phi_\psi(X) = \underline{\mathbf{W}}^{H_X^L}$. Finally, we obtain

$$\|\nabla \phi_\psi(X)\|_F^2 = \|\underline{\mathbf{W}}^{H_X^L}\|_F^2 \geq d_L \, \sigma_{\min}\left(\underline{\mathbf{W}}^{H_X^L}\right)^2 \geq \frac{d_L \alpha^L}{\kappa^L}.$$

Lemma 29. *Let $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, with $m \geq n \geq p$. Then $\sigma_{\min}(AB) \geq \sigma_{\min}(A)\, \sigma_{\min}(B)$.*

Proof. A more general version of this result can be found in [Güngör, 2007, Theorem 2]; we provide a proof here for completeness.

If B has a nontrivial null space, $\sigma_{\min}(B) = 0$ and the inequality holds. Other-

wise, let $\mathbb{R}^n_*$ denote $\mathbb{R}^n \setminus \{0\}$. Recall that for $C \in \mathbb{R}^{m\times n}$ with $m \geq n$,

$$\sigma_{\min}(C) = \sqrt{\lambda_{\min}(C^\mathsf{T}C)} = \sqrt{\inf_{x\in\mathbb{R}^n_*} \frac{x^\mathsf{T}C^\mathsf{T}Cx}{x^\mathsf{T}x}} = \inf_{x\in\mathbb{R}^n_*} \frac{\|Cx\|}{\|x\|}.$$

Thus, as $Bx \neq 0$ for $x \neq 0$,

$$\begin{aligned}
\sigma_{\min}(AB) &= \inf_{x\in\mathbb{R}^p_*} \frac{\|ABx\|}{\|x\|} = \inf_{x\in\mathbb{R}^p_*} \frac{\|ABx\|\|Bx\|}{\|Bx\|\|x\|} \\
&\geq \left(\inf_{x\in\mathbb{R}^p_*} \frac{\|ABx\|}{\|Bx\|}\right)\left(\inf_{x\in\mathbb{R}^p_*} \frac{\|Bx\|}{\|x\|}\right) \\
&\geq \left(\inf_{y\in\mathbb{R}^n_*} \frac{\|Ay\|}{\|y\|}\right)\left(\inf_{x\in\mathbb{R}^p_*} \frac{\|Bx\|}{\|x\|}\right) = \sigma_{\min}(A)\,\sigma_{\min}(B). \qquad \square
\end{aligned}$$

A .2.1 Necessity of some assumptions

Here we analyze through simple examples what happens when the condition number can be unbounded, and when Assumption **(II)**, about decreasing widths of the network, is violated.

Condition Number: We start by a first example where the condition number can be arbitrarily high. We consider a two-layer network on $\mathbb{R}^2$, defined by

$$\phi_\alpha(X) = \begin{bmatrix} 1 & -1 \end{bmatrix} \sigma(W_\alpha X) \qquad W_\alpha = \begin{bmatrix} 1 & 1 \\ 1 & 1+\alpha \end{bmatrix} \tag{4.16}$$

where $\alpha > 0$. As α approaches 0 the matrix W_α becomes singular which means that its condition number blows up. We are interested in analyzing the behavior of the Lipschitz constant of ϕ and the expected squared norm of its gradient under μ as α approaches 0.

One can easily compute the squared norm of the gradient of ϕ which is given by

$$\|\nabla\phi_\alpha(X)\|^2 = \begin{cases} \alpha^2 & X \in A_1 \\ \gamma^2\alpha^2 & X \in A_2 \\ (1-\gamma)^2 + (1+\alpha-\gamma)^2 & X \in A_3 \\ (1-\gamma)^2 + (\gamma\alpha+\gamma-1)^2 & X \in A_4 \end{cases}$$

Here A_1, A_2, A_3 and A_4 are defined by (4.17) and are represented in Figure 4.9:

$$\begin{aligned} A_1 &:= \{X \in \mathbb{R}^2 | X_1 + X_2 \geq 0 \quad X_1 + (1+\alpha)X_2 \geq 0\} \\ A_2 &:= \{X \in \mathbb{R}^2 | X_1 + X_2 < 0 \quad X_1 + (1+\alpha)X_2 < 0\} \\ A_3 &:= \{X \in \mathbb{R}^2 | X_1 + X_2 < 0 \quad X_1 + (1+\alpha)X_2 \geq 0\} \\ A_4 &:= \{X \in \mathbb{R}^2 | X_1 + X_2 \geq 0 \quad X_1 + (1+\alpha)X_2 < 0\} \end{aligned} \tag{4.17}$$

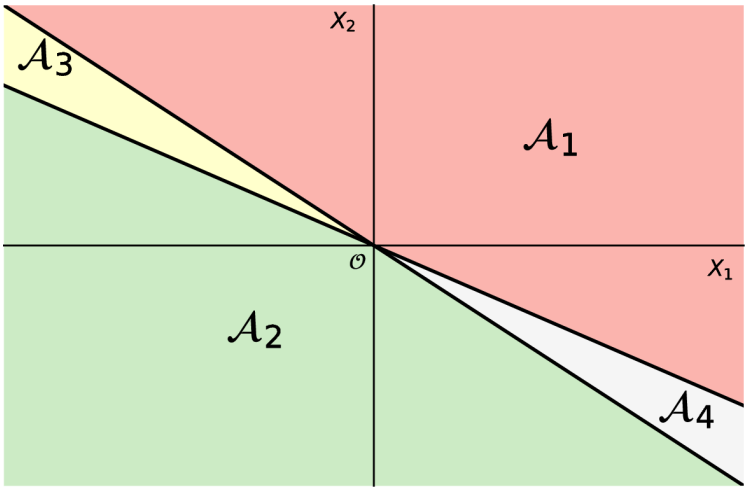


Figure 4.9: Decomposition of $\mathbb{R}^2$ into 4 regions A_1, A_2, A_3 and A_4 as defined in (4.17). As α approaches 0, the area of sets A_3 and A_4 becomes negligible.

It is easy to see that whenever μ has a density, the probability of the sets A_3 and A_4 goes to 0 are $\alpha \to 0$. Hence one can deduce that $\mathbb{E}_\mu[\|\nabla\phi_\alpha(X)\|^2] \to 0$ when $\alpha \to 0$. On the other hand, the squared Lipschitz constant of ϕ is given by $(1-\gamma)^2+(1+\alpha-\gamma)^2$ which converges to $2(1-\gamma)^2$. This shows that controlling the expectation of the gradient doesn't allow to effectively control the Lipschitz constant of ϕ.

Monotonicity of the dimensions: We would like to consider a second example where Assumption **(II)** doesn't hold. Consider the following two layer network defined by:

$$\phi(X) = \begin{bmatrix} -1 & 0 & 1 \end{bmatrix} \sigma(W_\beta X) \qquad W_\beta := \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & \beta \end{bmatrix}$$

for $\beta > 0$. Note that W_β is a full rank matrix, but Assumption **(II)** doesn't hold. Depending on the sign of the components of $W_\beta X$ one has the following expression for $\|\nabla\phi_\alpha(X)\|^2$:

$$\|\nabla\phi_\alpha(X)\|^2 = \begin{cases} \beta^2 & X \in B_1 \\ \gamma^2\beta^2 & X \in B_2 \\ \beta^2 & X \in B_3 \\ (1-\gamma)^2 + \gamma^2\beta^2 & X \in B_4 \\ (1-\gamma)^2 + \beta^2 & X \in B_5 \\ \gamma^2\beta^2 & X \in B_6 \end{cases}$$

where $(B_i)_{1\leq i\leq 6}$ are defined by (4.18)

$$\begin{aligned} B_1 &:= \{X \in \mathbb{R}^2 | X_1 \geq 0 \quad X_2 \geq 0\} \\ B_2 &:= \{X \in \mathbb{R}^2 | X_1 < 0 \quad X_2 < 0\} \\ B_3 &:= \{X \in \mathbb{R}^2 | X_1 \geq \quad X_2 < 0 \quad X_1 + \beta X_2 \geq 0\} \\ B_4 &:= \{X \in \mathbb{R}^2 | X_1 \geq \quad X_2 < 0 \quad X_1 + \beta X_2 < 0\} \\ B_5 &:= \{X \in \mathbb{R}^2 | X_1 > 0 \quad X_2 \geq 0 \quad X_1 + \beta X_2 \geq 0\} \\ B_6 &:= \{X \in \mathbb{R}^2 | X_1 > 0 \quad X_2 \geq 0 \quad X_1 + \beta X_2 < 0\} \end{aligned} \tag{4.18}$$

The squared Lipschitz constant is given by $\|\phi\|_L^2(1-\gamma)^2+\beta^2$ while the expected squared norm of the gradient of ϕ is given by:

$$\mathbb{E}_\mu[\|\phi(X)\|^2] = 3\beta^2(p(B_1\cup B_3\cup B_5)+\gamma^2 p(B_2\cup B_4\cup B_6))+(1-\gamma)^2 p(B_4\cup B_5).$$

Again the set $B_4\cup B_5$ becomes negligible as β approaches 0 which implies that $\mathbb{E}_\mu[\|\phi(X)\|^2]\to 0$. On the other hand $\|\phi\|_L^2$ converges to $(1-\gamma)^2$. Note that unlike in the first example in (4.16), the matrix W_β has a bounded condition number. In this example, the columns of W_0 are all in the null space of $\begin{bmatrix}-1 & 0 & 1\end{bmatrix}$, which implies $\nabla\phi_0(X)=0$ for all $X\in\mathbb{R}^2$, even though all matrices have full rank.

B An estimator for Lipschitz MMD

We now describe briefly how to estimate the Lipschitz MMD in low dimensions. Recall that

$$\mathrm{LipMMD}_{k,\lambda}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{H}_k\,:\,\|f\|_{\mathrm{Lip}}^2+\lambda\|f\|_{\mathcal{H}_k}^2\le 1} \mathbb{E}_{X\sim\mathbb{P}}[f(X)]-\mathbb{E}_{X\sim\mathbb{Q}}[f(Y)].$$

For $f\in\mathcal{H}_k$, it is the case that

$$\begin{aligned}\|f\|_{\mathrm{Lip}}^2 &= \sup_{x\in\mathbb{R}^d}\|\nabla f(x)\|^2 = \sup_{x\in\mathbb{R}^d}\sum_{i=1}^d \langle \partial_i k(x,\cdot), f\rangle_{\mathcal{H}_k}^2 \\ &= \sup_{x\in\mathbb{R}^d}\left\langle f, \sum_{i=1}^d [\partial_i k(x,\cdot)\otimes\partial_i k(x,\cdot)]\, f\right\rangle_{\mathcal{H}_k}.\end{aligned}$$

Thus we can approximate the constraint $\|f\|_{\mathrm{Lip}}^2+\lambda\|f\|_{\mathcal{H}_k}^2\le 1$ by enforcing the constraint on a set of m points $\{Z_i\}$ reasonably densely covering the region around the supports of $\mathbb{P}$ and $\mathbb{Q}$, rather than enforcing it at every point in $\mathcal{X}$. An estimator of the Lipschitz MMD based on $X\sim\mathbb{P}^{n_X}$ and $Y\sim\mathbb{Q}^{n_Y}$ is

$$\begin{aligned}\widehat{\mathrm{LipMMD}}_{k,\lambda}(X,Y,Z) \approx \sup_{f\in\mathcal{H}_k} &\frac{1}{n_X}\sum_{j=1}^{n_X} f(X_j)-\frac{1}{n_Y}\sum_{j=1}^{n_Y} f(Y_j) \\ &\text{s.t. } \forall j,\ \|\nabla f(Z_j)\|^2+\lambda\|f\|_{\mathcal{H}_k}^2\le 1 \qquad (4.19)\end{aligned}$$

By the generalized representer theorem, the optimal f for (4.19) will be of the form

$$f(\cdot) = \sum_{j=1}^{n_X} \alpha_j k(X_j, \cdot) + \sum_{j=1}^{n_Y} \beta_j k(Y_j, \cdot) + \sum_{i=1}^{d} \sum_{j=1}^{m} \gamma_{(i,j)} \partial_i k(Z_j, \cdot).$$

Writing $\delta = (\alpha, \beta, \gamma)$, the objective function is linear in δ,

$$\begin{bmatrix} \frac{1}{n_X} & \cdots & \frac{1}{n_X} & -\frac{1}{n_Y} & \cdots & -\frac{1}{n_Y} & 0 & \cdots & 0 \end{bmatrix} = \delta.$$

The constraints are quadratic, built from the following matrices, where the X and Y samples are concatenated together, as are the derivatives with each dimension of the Z samples:

$$K := \begin{bmatrix} k(X_1, X_1) & \cdots & k(X_1, Y_{n_Y}) \\ \vdots & \ddots & \vdots \\ k(Y_{n_X}, X_1) & \cdots & k(Y_{n_Y}, Y_{n_Y}) \end{bmatrix}$$

$$B := \begin{bmatrix} \partial_1 k(Z_1, X_1) & \cdots & \partial_1 k(Z_1, Y_{n_Y}) \\ \vdots & \ddots & \vdots \\ \partial_d k(Z_m, X_1) & \cdots & \partial_d k(Z_m, Y_{n_Y}) \end{bmatrix}$$

$$H := \begin{bmatrix} \partial_1 \partial_{1+d} k(Z_1, Z_1) & \cdots & \partial_1 \partial_{d+d} k(Z_1, Z_m) \\ \vdots & \ddots & \vdots \\ \partial_d \partial_{1+d} k(Z_m, Z_1) & \cdots & \partial_d \partial_{d+d} k(Z_m, Z_m) \end{bmatrix}.$$

Given these matrices, and letting $O_j = \sum_{i=1}^{d} e_{(i,j)} e_{(i,j)}^\mathsf{T}$ where $e_{(i,j)}$ is the (i,j)th standard basis vector in $\mathbb{R}^{md}$, we have that

$$\|f\|_{\mathcal{H}_k}^2 = \delta^\mathsf{T} \begin{bmatrix} K & B^\mathsf{T} \\ B & H \end{bmatrix} \delta$$

$$\|\nabla f(Z_j)\|^2 = \sum_{i=1}^{d} (\partial_i f(Z_j))^2 = \delta^\mathsf{T} \begin{bmatrix} B^\mathsf{T} O_j B & B^\mathsf{T} O_j H \\ H O_j B & H O_j H \end{bmatrix} \delta.$$

Thus the optimization problem (4.19) is a linear problem with convex quadratic constraints, which can be solved by standard convex optimization software. The approximation is reasonable only if we can effectively cover the region of interest with densely spaced $\{Z_i\}$; it requires a nontrivial amount of computation even for the very simple 1-dimensional toy problem of Example 1.

One advantage of this estimator, though, is that finding its derivative with respect to the input points or the kernel parameterization is almost free once we have computed the estimate, as long as our solver has computed the dual variables μ corresponding to the constraints in (4.19). We just need to exploit the envelope theorem and then differentiate the KKT conditions, as done for instance in Amos and Kolter [2017]. The differential of (4.19) ends up being, assuming the optimum of (4.19) is at $\hat{\delta} \in \mathbb{R}^{n_X+n_Y+md}$ and $\hat{\mu} \in \mathbb{R}^m$,

$$\begin{aligned}\widehat{\mathrm{dLipMMD}}_{k,\lambda}(X,Y,Z) =& \hat{\delta}^\mathsf{T} \begin{bmatrix} \mathrm{d}K \\ \mathrm{d}B \end{bmatrix} \begin{bmatrix} \frac{1}{n_X} & \cdots & \frac{1}{n_X} & -\frac{1}{n_Y} & \cdots & -\frac{1}{n_Y} \end{bmatrix}^\mathsf{T} \\ & - \sum_{j=1}^{m} \hat{\mu}_j \hat{\delta}^\mathsf{T} (\mathrm{d}P_j) \hat{\delta}\end{aligned}$$

with P_j defined as:

$$\begin{aligned}P_j :=& \begin{bmatrix} (\mathrm{d}B)^\mathsf{T} O_j B + B^\mathsf{T} O_j (\mathrm{d}H) & (\mathrm{d}B)^\mathsf{T} O_j H + B^\mathsf{T} O_j (\mathrm{d}H) \\ (\mathrm{d}H) O_j B + H O_j (\mathrm{d}B) & (\mathrm{d}H) O_j H + H O_j (\mathrm{d}H) \end{bmatrix} \\ & + \lambda \begin{bmatrix} \mathrm{d}K & \mathrm{d}B^\mathsf{T} \\ \mathrm{d}B & \mathrm{d}H \end{bmatrix}.\end{aligned}$$

C Near-equivalence of WGAN and linear-kernel MMD GANs

For an MMD GAN-GP with kernel $k(x,y) = \phi(x)\phi(y)$, we have that

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = |\mathbb{E}_\mathbb{P}\phi(x) - \mathbb{E}_\mathbb{Q}\phi(Y)|$$

and the corresponding critic function is

$$\begin{aligned}\frac{\eta(t)}{\|\eta\|_{\mathcal{H}}} &= \frac{\mathbb{E}_{X\sim\mathbb{P}}\phi(X)\phi(t) - \mathbb{E}_{Y\sim\mathbb{Q}}\phi(Y)\phi(t)}{|\mathbb{E}_{\mathbb{P}}\phi(X) - \mathbb{E}_{\mathbb{Q}}\phi(Y)|} \\ &= \operatorname{sign}\left(\mathbb{E}_{X\sim\mathbb{P}}\phi(X) - \mathbb{E}_{Y\sim\mathbb{Q}}\phi(Y)\right)\phi(t).\end{aligned}$$

Thus if we assume $\mathbb{E}_{X\sim\mathbb{P}}\phi(X) > \mathbb{E}_{Y\sim\mathbb{Q}}\phi(Y)$, as that is the goal of our critic training, we see that the MMD becomes identical to the WGAN loss, and the gradient penalty is applied to the same function.

(MMD GANs, however, would typically train on the unbiased estimator of MMD^2, giving a very slightly different loss function. Bińkowski* et al. [2018] also applied the gradient penalty to η rather than the true critic $\eta/\|\eta\|$.)

The SMMD with a linear kernel is thus analogous to applying the scaling operator to a WGAN; hence the name SWGAN.

D Experiments on synthetic data

D .1 DiracGAN vector fields for more losses

Figure 4.10 shows parameter vector fields, like those in Figure 4.1, for Example 1 and for a variety of different losses:

$$\begin{aligned}
\text{MMD:} &\quad -\mathrm{MMD}_\psi^2 \\
\text{MMD-GP:} &\quad -\mathrm{MMD}_\psi^2 + \lambda\mathbb{E}_{\mathbb{P}}[(\|\nabla f(X)\| - 1)^2] \\
\text{MMD-GP-Unif:} &\quad -\mathrm{MMD}_\psi^2 + \lambda\mathbb{E}_{\widetilde{X}\simeq\mu^*}[(\|\nabla f(\widetilde{X})\| - 1)^2] \\
\text{SN-MMD:} &\quad -2\,\mathrm{MMD}_1(\mathbb{P},\mathbb{Q})^2 \\
\text{Sobolev-MMD:} &\quad -\mathrm{MMD}_\psi^2 + \lambda(\mathbb{E}_{(\mathbb{P}+\mathbb{Q})/2}[\|\nabla f(X)\|^2] - 1)^2 \\
\text{CenteredSobolev-MMD:} &\quad -\mathrm{MMD}_\psi^2 + \lambda(\mathbb{E}_{(\mathbb{P}+\mathbb{Q})/2}[\|\nabla f(X)\|^2])^2 \\
\text{LipMMD:} &\quad -\mathrm{LipMMD}_{k_\psi,\lambda}^2 \\
\text{GC-MMD:} &\quad -\mathrm{GCMMD}_{\mathcal{N}(0,10^2),k_\psi,\lambda}^2 \\
\text{SMMD:} &\quad -\mathrm{SMMD}_{k_\psi,\mathbb{P},\lambda}^2
\end{aligned} \tag{4.20}$$

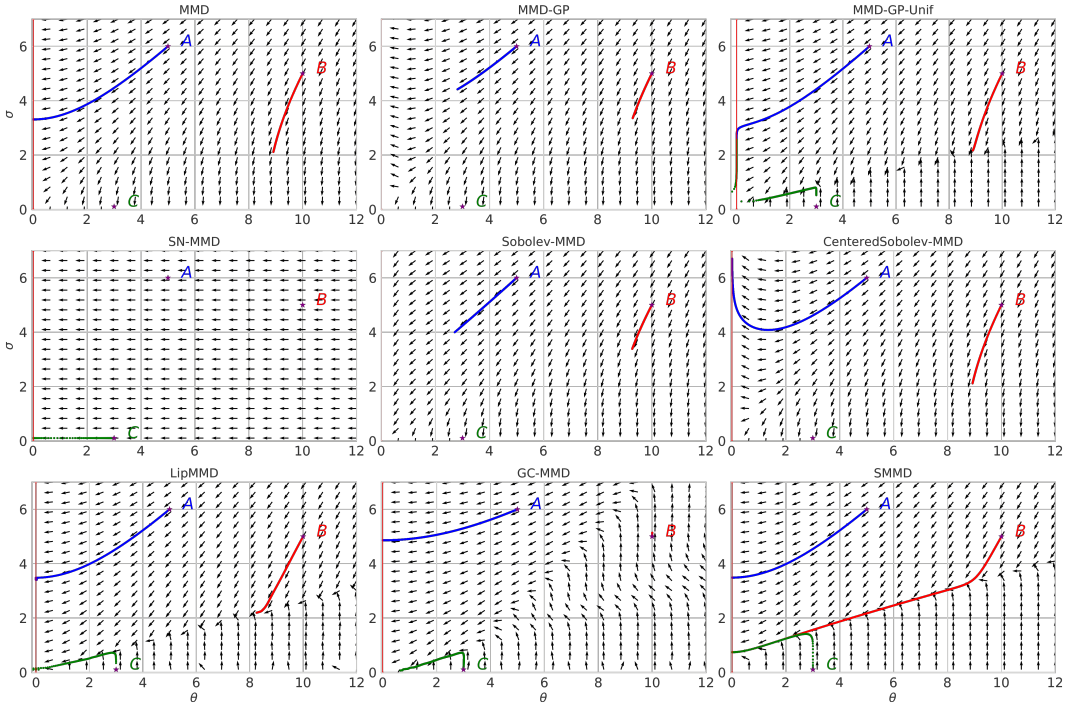
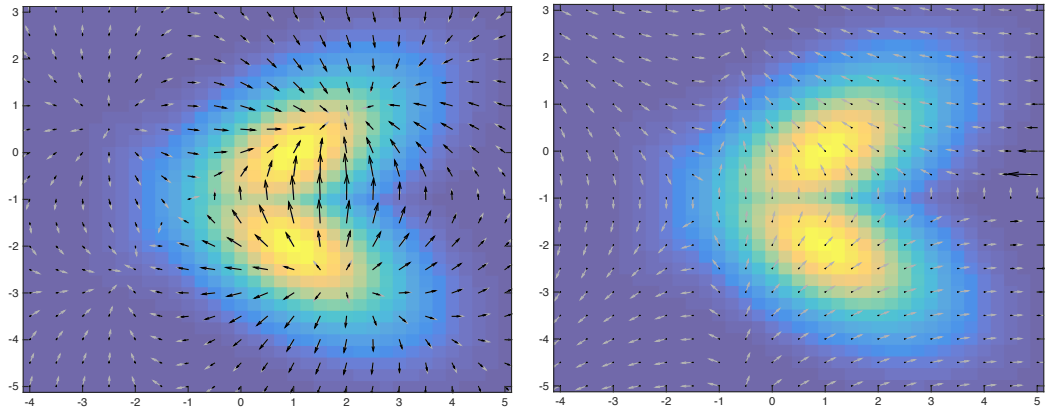


Figure 4.10: Vector fields for different losses with respect to the generator parameter θ and the feature representation parameter ψ; the losses use a Gaussian kernel, and are shown in (4.20). Following Mescheder et al. [2018], $\mathbb{P} = \delta_0$, $\mathbb{Q} = \delta_\theta$ and $\phi_\psi(x) = \psi x$. The curves show the result of taking simultaneous gradient steps in (θ, ψ) beginning from three initial parameter values.

The squared MMD between δ_0 and δ_θ under a Gaussian kernel of bandwidth $1/\psi$ and is given by $2(1 - e^{-\frac{\psi^2\theta^2}{2}})$. MMD-GP-unif uses a gradient penalty as in Bińkowski* et al. [2018] where each samples from μ^* is obtained by first sampling X and Y from $\mathbb{P}$ and $\mathbb{Q}$ and then sampling uniformly between X and Y. MMD-GP uses the same gradient penalty, but the expectation is taken under $\mathbb{P}$ rather than μ^*. SN-MMD refers to MMD with spectral normalization; here this means that $\psi = 1$. Sobolev-MMD refers to the loss used in Mroueh et al. [2018] with the quadratic penalty only. $\mathrm{GCMMD}_{\mu,k,\lambda}$ is defined by (4.5), with $\mu = \mathcal{N}(0, 10^2)$.

D .2 Vector fields of Gradient-Constrained MMD and Sobolev GAN critics

Mroueh et al. [2018] argue that *the gradient of the critic (...) defines a transportation plan for moving the distribution mass* (from generated to reference distribution) and present the solution of Sobolev PDE for 2-dimensional Gaussians. We observed that



(a) Gradient-Constrained MMD critic gradient.

(b) Sobolev IPM critic gradient.

Figure 4.11: Vector fields of critic gradients between two Gaussians. The grey arrows show normalized gradients, i.e. gradient directions, while the black ones are the actual gradients. Note that for the Sobolev critic, gradients norms are orders of magnitudes higher on the right hand side of the plot than in the areas of high density of the given distributions.

in this simple example the gradient of the Sobolev critic can be very high outside of the areas of high density, which is not the case with the Gradient-Constrained MMD. Figure 4.11 presents critic gradients in both cases, using $\mu = (\mathbb{P} + \mathbb{Q})/2$ for both. This unintuitive behavior is most likely related to the vanishing boundary condition, assummed by Sobolev GAN. Solving the actual Sobolev PDE, we found that the Sobolev critic has very high gradients close to the boundary in order to match the condition; moreover, these gradients point in opposite directions to the target distribution.

Chapter 5

Generalized energy based models

We introduce the Generalized Energy Based Model (GEBM) for generative modelling. These models combine two trained components: a base distribution (generally an implicit model), which can learn the support of data with low intrinsic dimension in a high dimensional space; and an energy function, to refine the probability mass on the learned support. Both the energy function and base jointly constitute the final model, unlike GANs, which retain only the base distribution (the "generator"). GEBMs are trained by alternating between learning the energy and the base. We show that both training stages are well-defined: the energy is learned by maximising a generalized likelihood, and the resulting energy-based loss provides informative gradients for learning the base. Samples from the posterior on the latent space of the trained model can be obtained via MCMC, thus finding regions in this space that produce better quality samples. Empirically, the GEBM samples on image-generation tasks are of much better quality than those from the learned generator alone, indicating that all else being equal, the GEBM will outperform a GAN of the same complexity. When using normalizing flows as base measures, GEBMs succeed on density modelling tasks, returning comparable performance to direct maximum likelihood of the same networks.

1 Introduction

Energy-based models (EBMs) have a long history in physics, statistics and machine learning [LeCun et al., 2006]. They belong to the class of *explicit* models, and

can be described by a family of energies E which define probability distributions with density proportional to $\exp(-E)$. Those models are often known up to a normalizing constant $Z(E)$, also called the *partition function*. The learning task consists of finding an optimal function that best describes a given system or target distribution $\mathbb{P}$. This can be achieved using maximum likelihood estimation (MLE), however the intractability of the normalizing partition function makes this learning task challenging. Thus, various methods have been proposed to address this [Hinton, 2002, Hyvärinen, 2005, Gutmann and Hyvärinen, 2012, Dai et al., 2019a,b]. All these methods estimate EBMs that are supported over the whole space. In many applications, however, $\mathbb{P}$ is believed to be supported on an unknown lower dimensional manifold. This happens in particular when there are strong dependencies between variables in the data, and suggests incorporating a low-dimensionality hypothesis in the model.

Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] are a particular way to enforce low dimensional structure in a model. They rely on an *implicit* model, the generator, to produce samples supported on a low-dimensional manifold by mapping a pre-defined latent noise to the sample space using a trained function. GANs have been very successful in generating high-quality samples on various tasks, especially for unsupervised image generation [Brock et al., 2017]. The generator is trained *adversarially* against a discriminator network whose goal is to distinguish samples produced by the generator from the target data. This has inspired further research to extend the training procedure to more general losses [Nowozin et al., 2016, Arjovsky et al., 2017, Li et al., 2017, Bińkowski* et al., 2018] and to improve its stability [Miyato et al., 2018, Gulrajani et al., 2017, Nagarajan and Kolter, 2017, Kodali et al., 2017]. While the generator of a GAN has effectively a low-dimensional support, it remains challenging to refine the distribution of mass on that support using pre-defined latent noise. For instance, as shown by Cornish et al. [2020] for normalizing flows, when the latent distribution is unimodal and the target distribution possesses multiple disconnected low-dimensional components, the generator, as a continuous map, compensates for this mismatch using steeper

slopes. In practice, this implies the need for more complicated generators.

In the present work, we propose a new class of models, called *Generalized Energy Based Models* (GEBMs), which can represent distributions supported on low-dimensional manifolds, while offering more flexibility in refining the mass on those manifolds. GEBMs combine the strength of both *implicit* and *explicit* models in two separate components: a base distribution (often chosen to be an implicit model) which learns the low-dimensional support of the data, and an energy function that can refine the probability mass on that learned support. We propose to train the GEBM by alternating between learning the energy and the base, analogous to f-GAN training [Goodfellow et al., 2014, Nowozin et al., 2016]. The energy is learned by maximizing a generalized notion of likelihood which we relate to the *Donsker-Varadhan* lower-bound [Donsker and Varadhan, 1975] and *Fenchel duality*, as in [Nguyen et al., 2010, Nowozin et al., 2016]. Although the partition function is intractable in general, we propose a method to learn it in an amortized fashion without introducing additional surrogate models, as done in variational inference [Kingma and Welling, 2014, Rezende et al., 2014] or by Dai et al. [2019a,b]. The resulting maximum likelihood estimate, the *KL Approximate Lower-bound Estimate (*KALE*)*, is then used as a loss for training the base. When the class of energies is rich and smooth enough, we show that KALE leads to a meaningful criterion for measuring weak convergence of probabilities. Following recent work by Chu et al. [2020], Sanjabi et al. [2018], we show that KALE possesses well defined gradients w.r.t. the parameters of the base, ensuring well-behaved training. We also provide convergence rates for the empirical estimator of KALE when the variational family is sufficiently well behaved, which may be of independent interest.

The main advantage of GEBMs becomes clear when sampling from these models: the posterior over the latents of the base distribution incorporates the learned energy, putting greater mass on regions in this latent space that lead to better quality samples. Sampling from the GEBM can thus be achieved by first sampling from the posterior distribution of the latents via MCMC in the low-dimensional latent space, then mapping those latents to the input space using the implicit map of the

base. This is in contrast to standard GANs, where the latents of the base have a fixed distribution. We focus on a class of samplers that exploit gradient information, and show that these samplers enjoy fast convergence properties by leveraging the recent work of Eberle et al. [2017]. While there has been recent interest in using the discriminator to improve the quality of the generator during sampling [Azadi et al., 2019, Turner et al., 2019, Neklyudov et al., 2019, Grover et al., 2019, Tanaka, 2019, Wu et al., 2019a], our approach emerges naturally from the model we consider.

We begin in Section 2 by introducing the GEBM model. In Section 3 , we describe the learning procedure using KALE, then derive a method for sampling from the learned model in Section 4 . In Section 5 we discuss related work. Finally, experimental results are presented in Section 6 .

2 Generalized Energy-Based Models

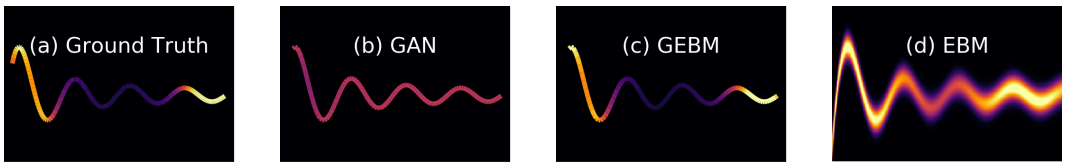


Figure 5.1: Data generating distribution supported on a line and with higher density at the extremities. Models are learned using either a GAN, GEBM, or EBM. More details are provided in Section B .1.

In this section, we introduce generalized energy based models (GEBM), that combine the strengths of both energy-based models and implicit generative models, and admit the first of these as a special case. An **energy-based model** (EBM) is defined by a set $\mathcal{E}$ of real valued functions called *energies*, where each $E \in \mathcal{E}$ specifies a probability density over the data space $\mathcal{X} \subset \mathbb{R}^d$ up to a normalizing constant,

$$\mathbb{Q}(\mathrm{d}x) = \exp\left(-E(x) - A\right)\mathrm{d}x, \qquad A = \log\left(\int \exp(-E(x))\,\mathrm{d}x\right). \tag{5.1}$$

While EBMs have been shown recently to be powerful models for representing complex high dimensional data distributions, they still unavoidably lead to a blurred model whenever data are concentrated on a lower-dimensional manifold. This is the case in Figure 5.1(a), where the ground truth distribution is supported on a 1-D

line and embedded in a 2-D space. The EBM in Figure 5.1(d) learns to give higher density to a halo surrounding the data, and thus provides a blurred representation. That is a consequence of EBM having a density defined over the whole space, and can result in blurred samples for image models.

An **implicit generative model** (IGM) is a family of probability distributions $\mathbb{G}_\theta$ parametrized by a learnable *generator* function $G : \mathcal{Z} \mapsto \mathcal{X}$ that maps latent samples z from a fixed latent distribution η to the data space $\mathcal{X}$. The latent distribution η is required to have a density over the latent space $\mathcal{Z}$ and is often easy to sample from. Thus, Sampling from $\mathbb{G}$ is simply achieved by first sampling z from η then applying G,

$$x \sim \mathbb{G} \quad \iff \quad x = G(z), \quad z \sim \eta. \tag{5.2}$$

GANs are popular instances of these models, and are trained *adversarially* [Goodfellow et al., 2014]. When the latent space $\mathcal{Z}$ has a smaller dimension than the input space $\mathcal{X}$, the IGM will be supported on a lower dimensional manifold of $\mathcal{X}$, and thus will not possess a Lebesgue density on $\mathcal{X}$ [Bottou et al., 2018]. IGMs are therefore good candidates for modelling low dimensional distributions. While GANs can accurately learn the low-dimensional support of the data, they can have limited power for representing the distribution of mass on the support. This is illustrated in Figure 5.1(b).

A **generalized energy-based model** (GEBM) $\mathbb{Q}$ is defined by a combination of a *base* $\mathbb{G}$ and an *energy* E defined over a subset $\mathcal{X}$ of $\mathbb{R}^d$. The **base** component can typically be chosen to be an IGM as in (5.2). The **generalized energy** component can refine the mass on the support defined by the *base*. It belongs to a class $\mathcal{E}$ of real valued functions defined on the input space $\mathcal{X}$, and represents the negative log-density of a sample from the GEBM with respect to the base $\mathbb{G}$,

$$\mathbb{Q}(\mathrm{d}x) = \exp\left(-E(x) - A_{\mathbb{G},E}\right) \mathbb{G}(\mathrm{d}x), \tag{5.3}$$

where $A_{\mathbb{G},E}$ is the logarithm of the normalizing constant of the model w.r.t. $\mathbb{G}$:

$$A_{\mathbb{G},E} = \log\left(\int \exp(-E(x))\mathbb{G}(\mathrm{d}x)\right).$$

Thus, a GEBM $\mathbb{Q}$ re-weights samples from the base according to the un-normalized importance weights $\exp(-E(x))$. Using the latent structure of the base $\mathbb{G}$, this importance weight can be pulled-back to the latent space to define a *posterior latent* distribution ν,

$$\nu(z) := \eta(z)\exp\left(-E\left(G(z)\right) - A_{\mathbb{G},E}\right). \tag{5.4}$$

Hence, the *posterior latent* ν can be used instead of the latent noise η for sampling from $\mathbb{Q}$, as summarized by Proposition 30:

Proposition 30. *Sampling from $\mathbb{Q}$ requires sampling a latent z from ν (5.4) then applying the map G,*

$$x \sim \mathbb{Q} \quad \Longleftrightarrow \quad x = G(z), \quad z \sim \nu. \tag{5.5}$$

In order to hold, Proposition 30 does not need the generator G to be invertible. We provide a proof in Section A .2 which relies on a characterization of probability distribution using generalized moments. We will see later in Section 4 how equation (5.5) can be used to provide practical sampling algorithms from the GEBM. Next we discuss the advantages of GEBMs.

Advantages of Generalized Energy Based Models. The GEBM defined by (5.3) can be related to exponential tilting (re-weighting) [Siegmund, 1976, Xie et al., 2016] of the base $\mathbb{G}$. The important difference over classical EBMs is that the base $\mathbb{G}$ is allowed to change its support and shape in space. By learning the base $\mathbb{G}$, GEBMs can accurately learn the low-dimensional support of data, just like IGMs do. They also benefit from the flexibility of EBMs for representing densities using an energy E to refine distribution of mass on the support defined by $\mathbb{G}$, as seen in Figure 5.1(c).

Compared to EBMs , that put mass on the whole space by construction (positive density), GEBMs have the additional flexibility to concentrate the probability mass on a low-dimensional support learned by the base $\mathbb{G}$, provided that the dimension of the latent space $\mathcal{Z}$ is smaller than the dimension of the ambient space $\mathcal{X}$: see Figure 5.1(c) vs Figure 5.1(d). In the particular case when the dimension of $\mathcal{Z}$ is equal to the ambient dimension and G is invertible, the base $\mathbb{G}$ becomes supported over the whole space $\mathcal{X}$, and GEBM recover usual EBMs. The next proposition further shows that any EBM can be viewed as a particular cases of GEBMs, as proved in Section A .2.

Proposition 31. *Any EBM with energy E (as in (5.1)) can be expressed as a GEBM with base $\mathbb{G}$ given as a normalizing flow with density $exp(-r(x))$ and a generalized energy $\tilde{E}(x) = E(x) - r(x)$. In this particular case, the dimension of the latent is necessarily equal to the data dimension, i.e. $dim(\mathcal{Z}) = dim(\mathcal{X})$.*

Compared to IGMs , that rely on a fixed pre-determined latent noise distribution η, GEBMs offer the additional flexibility of learning a richer latent noise distribution. This is particularly useful when the data is multimodal. In IGMs, such a GANs, the latent noise η is usually unimodal thus requiring a more sophisticated generator to distort a unimodal noise distribution into a distribution with multiple modes, as shown by Cornish et al. [2020]. Instead, GEBMs allow to sample from a *posterior* ν over the latent noise defined in (5.4). This posterior noise can be multimodal in latent space (by incorporating information from the energy) and thus can put more or less mass in specific regions of the manifold defined by the base $\mathbb{G}$. This allows GEBMs to capture multimodality in data, provided the support of the base is broad enough to subsume the data support Figure 5.1(c). This additional flexibility comes at no additional training cost compared to GANs. Indeed, GANs still require another model during training, the discriminator network, but do not use it for sampling. Instead, GEBMs avoid this waste since the base and energy can be trained jointly, with no other additional model, and then both are used for sampling.

3 Learning GEBMs

In this section we describe a general procedure for learning GEBMs. We decompose the learning procedure into two steps: an *energy learning* step and a *base learning* step. The overall learning procedure alternates between these two steps, as done in GAN training [Goodfellow et al., 2014].

3 .1 Learning the energy

When the base $\mathbb{G}$ is fixed, varying the energy E leads to a family of models that all admit a density $\exp(-E-A_{\mathbb{G},E})$ w.r.t. $\mathbb{G}$. When the base $\mathbb{G}$ admits a density $\exp(-r)$ defined over the whole space, it is possible to learn the energy E by maximizing the likelihood of the model $-\int(E+r)\,\mathrm{d}\mathbb{P} - A_{\mathbb{G},E}$. However, in general $\mathbb{G}$ is supported on a lower-dimensional manifold so that r is ill-defined and the usual notion of likelihood cannot be used. Instead, we introduce a generalized notion of likelihood which does not require a well defined density $\exp(-r)$ for $\mathbb{G}$:

Definition 6 (Generalized Likelihood)**.** *The expected $\mathbb{G}$-log-likelihood under a target distribution $\mathbb{P}$ of a GEBM model $\mathbb{Q}$ with base $\mathbb{G}$ and energy E is defined as*

$$\mathcal{L}_{\mathbb{P},\mathbb{G}}(E) := -\int E(x)d\mathbb{P}(x) - A_{\mathbb{G},E}. \tag{5.6}$$

To provide intuitions about the generalized likelihood in Definition 6, we start by discussing the particular case where $KL(\mathbb{P}||\mathbb{G}) < +\infty$. We then present the training method in the general case where $\mathbb{P}$ and $\mathbb{G}$ might not share the same support, i.e. $KL(\mathbb{P}||\mathbb{G}) = +\infty$.

Special case of finite $KL(\mathbb{P}||\mathbb{G})$. When the Kullback-Leibler divergence between $\mathbb{P}$ and $\mathbb{G}$ is well defined, (5.6) corresponds to the Donsker-Varadhan (DV) lower bound on the KL [Donsker and Varadhan, 1975], meaning that $\mathrm{KL}(\mathbb{P}||\mathbb{G}) \geq \mathcal{L}_{\mathbb{P},\mathbb{G}}(E)$ for all E. Moreover, the following proposition holds:

Proposition 32. *Assume that $KL(\mathbb{P}||\mathbb{G}) < +\infty$ and $0 \in \mathcal{E}$. If, in addition, $E^\star$*

maximizes (5.6), then:

$$KL(\mathbb{P}||\mathbb{Q}) \leq KL(\mathbb{P}||\mathbb{G}). \tag{5.7}$$

In addition, we have that $KL(\mathbb{P}||\mathbb{Q}) = 0$ *when* $E^\star$ *is the negative log-density ratio of* $\mathbb{P}$ *w.r.t.* $\mathbb{G}$.

We refer to Section A .2 for a proof. According to (5.7), the GEBM systematically improves over the IGM defined by $\mathbb{G}$, with no further improvement possible in the limit case when $\mathbb{G} = \mathbb{P}$. Hence as long as there is an error in mass on the common support of $\mathbb{P}$ and $\mathbb{G}$, the GEBM improves over the base $\mathbb{G}$.

Estimating the likelihood in the General setting. Definition 6 can be used to learn a maximum likelihood energy $E^\star$ by maximizing $\mathcal{L}_{\mathbb{P},\mathbb{G}}(E)$ w.r.t. E even when the $\mathrm{KL}(\mathbb{P}||\mathbb{G})$ is infinite and when $\mathbb{P}$ and $\mathbb{G}$ don't necessarily share the same support. The GEBM defined by the base $\mathbb{G}$ and optimal energy $E^\star$ is the best model for the data, as measured by the generalized KL, within the family of models supported on $\mathbb{G}$ and with energies E in $\mathcal{E}$. However, if the base is far away from the support of the data, this is still not a good model. This already suggests the importance of learning a good base as we discuss in Section 3 .2.

The optimal solution $E^\star$ is well defined whenever the set of energies is suitably constrained. This is the case if the energies are parametrized by a compact set Ψ with $\psi \mapsto E_\psi$ continuous over Ψ. Estimating the likelihood is then achieved using i.i.d. samples $(X_n)_{1:N}, (Y_m)_{1:M}$ from $\mathbb{P}$ and $\mathbb{G}$ [Tsuboi et al., 2009, Sugiyama et al., 2012, Liu et al., 2017]:

$$\hat{\mathcal{L}}_{\mathbb{P},\mathbb{G}}(E) = -\frac{1}{N}\sum_{n=1}^{N} E(X_n) - \log\left(\frac{1}{M}\sum_{m=1}^{M}\exp(-E(Y_m))\right). \tag{5.8}$$

In the context of mini-batch stochastic gradient methods, however, M typically ranges from 10 to 1000, which can lead to a poor estimate for the log-partition function $A_{\mathbb{G},E}$. Moreover, (5.8) doesn't exploit estimates of $A_{\mathbb{G},E}$ from previous gradient iterations. Instead, we propose an estimator which introduces a variational

parameter $A \in \mathbb{R}$ meant to estimate $A_{\mathbb{G},E}$ in an amortized fashion. The key idea is to exploit the convexity of the exponential which directly implies $-A_{\mathbb{G},E} \geq -A - \exp(-A + A_{\mathbb{G},E}) + 1$ for any $A \in \mathbb{R}$, with equality only when $A = A_{\mathbb{G},E}$. Therefore, (5.6) admits a lower-bound of the form

$$\mathcal{L}_{\mathbb{P},\mathbb{G}}(E) \geq -\int (E + A)\,\mathrm{d}\mathbb{P} - \int \exp(-(E + A))\,\mathrm{d}\mathbb{G} + 1 := \mathcal{F}_{\mathbb{P},\mathbb{G}}(E + A),$$

where we introduced the functional $\mathcal{F}_{\mathbb{P},\mathbb{G}}$ for concision. Maximizing $\mathcal{F}_{\mathbb{P},\mathbb{G}}(E + A)$ over A recovers the likelihood $\mathcal{L}_{\mathbb{P},\mathbb{G}}(E)$. Moreover, jointly maximizing over E and A yields the maximum likelihood energy $E^\star$ and its corresponding log-partition function $A^\star = A_{\mathbb{G},E^\star}$. This optimization is well-suited for stochastic gradient methods using the following estimator Kanamori et al. [2011]:

$$\begin{aligned}\hat{\mathcal{F}}_{\mathbb{P},\mathbb{G}}(E + A) = &-\frac{1}{N}\sum_{n=1}^{N}(E(X_n) + A) \\ &-\frac{1}{M}\sum_{m=1}^{M}\exp(-(E(Y_m) + A)) + 1.\end{aligned} \tag{5.9}$$

Estimating the log-partition function. Optimizing (5.9) exactly over A yields (5.8), with the optimal A equal to $\tilde{A} = \log(\frac{1}{M}\sum_{m=1}^{M}\exp(-E(Y_m)))$. However, to maintain an amortized estimator of the log-partition we propose to optimize (5.9) iteratively using second order updates:

$$A_{k+1} = A_k - \lambda(\exp(A_k - \tilde{A}_{k+1}) - 1), \qquad A_0 = \tilde{A}_0 \tag{5.10}$$

where λ is a learning rate and $\tilde{A}_{k+1}$ is the empirical log-partition function estimated from a batch of new samples. By leveraging updates from previous iterations, A can yield much more accurate estimates of the log-partition function as confirmed empirically in Figure 5.7 of Section 6 .

3 .2 Learning the base

Unlike in Section 3 .1, varying the base $\mathbb{G}$ does not need to preserve the same support. Thus, it is generally not possible to use maximum likelihood methods for learning $\mathbb{G}$.

Instead, we propose to use the generalized likelihood (5.6) evaluated at the optimal energy $E^\star$ as a meaningful loss for learning $\mathbb{G}$, and refer to it as the *KL Approximate Lower-bound Estimate* (KALE),

$$\text{KALE}(\mathbb{P}||\mathbb{G}) = \sup_{(E,A)\in\mathcal{E}\times\mathbb{R}} \mathcal{F}_{\mathbb{P},\mathbb{G}}(E+A). \tag{5.11}$$

From Section 3 .1, $\text{KALE}(\mathbb{P}||\mathbb{G})$ is always a lower bound on $\text{KL}(\mathbb{P},\mathbb{G})$. The bound becomes tight whenever the negative log density of $\mathbb{P}$ w.r.t. $\mathbb{G}$ is well-defined and belongs to $\mathcal{E}$ (Section C of the supplementary material). Moreover, Proposition 33 shows that KALE is a reliable criterion for measuring convergence, and is a consequence of [Zhang et al., 2017, Theorem B.1], with a proof in Section A .1.1 of the supplementary material:

Proposition 33. *Assume all energies in $\mathcal{E}$ are L-Lipschitz and that any continuous function can be well approximated by linear combinations of energies in $\mathcal{E}$ (Assumptions **(A)** and **(B)** of Section A .1), then KALE$(\mathbb{P}||\mathbb{G}) \geq 0$ with equality only if $\mathbb{P}=\mathbb{G}$ and KALE$(\mathbb{P}||\mathbb{G}^n) \to 0$ iff $\mathbb{G}^n \to \mathbb{P}$ in distribution.*

The universal approximation assumption holds in particular when $\mathcal{E}$ contains feedforward networks. In fact networks with a single neuron are enough, as shown in [Zhang et al., 2017, Theorem 2.3]. The Lipschitz assumption holds when additional regularization of the energy is enforced during training by methods such as **spectral normalization** [Miyato et al., 2018] or additional regularization $I(\psi)$ on the energy E_ψ such as the **gradient penalty** [Gulrajani et al., 2017] as done in Section 6 .

Estimating KALE. According to Arora et al. [2017], accurate finite sample estimates of divergences that result from an optimization procedures (such as in (5.11)) depend on the richness of the class $\mathcal{E}$; and richer energy classes can result in slower convergence. Unlike divergences such as Jensen-Shannon, KL and the Wasserstein distance, which result from optimizing over a non-parametric and rich class of functions, KALE is restricted to a class of parametric energies E_ψ. Thus, [Arora et al., 2017, Theorem 3.1] applies, and guarantees good finite sample estimates, provided optimization is solved accurately. In Section C .1 of the supplementary material,

we provide an analysis for the more general case where energies are not necessarily parametric but satisfy some further smoothness properties; we emphasize that our rates do not require the strong assumption that the density ratio is bounded above and below as in [Nguyen et al., 2010].

Smoothness of KALE. Learning the base is achieved by minimizing $\mathcal{K}(\theta) := \mathrm{KALE}(\mathbb{P}||\mathbb{G}_\theta)$ over the set of parameters Θ of the generator G_θ. This requires $\mathcal{K}(\theta)$ to be smooth enough so that gradient methods converge to local minima and avoid instabilities during training [Chu et al., 2020]. Ensuring smoothness of losses that result from an optimization procedure, as in (5.11), can be challenging. Results for the regularized Wasserstein are provided by Sanjabi et al. [2018], while more general losses are considered by Chu et al. [2020], albeit under stronger conditions than for our setting.Theorem 34 shows that when E, G_θ and their gradients are all Lipschitz then $\mathcal{K}(\theta)$ is smooth enough.We provide a proof for Theorem 34 in Section A .1.1.

Theorem 34. *Under Assumptions **(I)** to **(III)** of Section A .1, sub-gradient methods on $\mathcal{K}$ converge to local optima. Moreover, $\mathcal{K}$ is Lipschitz and differentiable for almost all $\theta \in \Theta$ with:*

$$\nabla\mathcal{K}(\theta) = \exp(-A_{G_\theta,E^\star}) \int \nabla_x E^\star(G_\theta(z))\nabla_\theta G_\theta(z) \exp(-E^\star(G_\theta(z)))\eta(z)\,\mathrm{d}z. \tag{5.12}$$

Estimating the gradient in (5.12) is achieved by first optimizing over E_ψ and A using (5.9), with additional regularization $I(\psi)$. The resulting estimators $\hat{E}^\star$ and $\hat{A}^\star$ are plugged in (5.13) to estimate $\nabla\mathcal{K}(\theta)$ using samples $(Z_m)_{1:M}$ from η. Unlike for learning the energy $E^\star$, which benefits from using the amortized estimator of the log-partition function, we found that using the empirical log-partition for learning the base was more stable.

$$\widehat{\nabla\mathcal{K}(\theta)} = \frac{\exp(-\hat{A}^\star)}{M} \sum_{m=1}^{M} \nabla_x \hat{E}^\star(G_\theta(Z_m))\nabla_\theta G_\theta(Z_m) \exp(-\hat{E}^\star(G_\theta(Z_m))). \tag{5.13}$$

Training We summarize the training procedure in Algorithm 1, which alternates between learning the energy and the base in a similar fashion to *adversarial training*. An additional regularization, denoted by $I(\psi)$ is used to ensure conditions of Proposition 33 and Theorem 34 hold. $I(\psi)$ can include L_2 regularization over the parameters ψ, a gradient penalty as in Gulrajani et al. [2017] or Spectral normalization Miyato et al. [2018]. The energy can be trained either using the estimator in (5.8) (KALE-DV) or the one in (5.9) (KALE-F) depending on the variable $\mathcal{C}$.

Algorithm 1 Training GEBM

Input $\mathbb{P}$, N,M, n_b, n_e
Output Trained generator G_θ and energy E_ψ.
Initialize θ , ψ *and* A.
for $k = 1, \dots, n_b$ **do**
  for $j = 1, \dots, n_e$ **do**
    Sample $\{X_n\}_{1:N} \sim \mathbb{P}$ and $\{Y_n\}_{1:N} \sim \mathbb{G}_\theta$
    $g_\psi \leftarrow -\nabla_\psi \hat{\mathcal{F}}_{\mathbb{P},\mathbb{G}_\theta}(E_\psi + A) + I(\psi)$
    $\tilde{A} \leftarrow \log\left(\frac{1}{M}\sum_{m=1}^{M} \exp(-E_\psi(Y_m))\right)$
    $g_A \leftarrow \exp(A - \tilde{A}) - 1$
    Update ψ *and* A *using* g_ψ *and* g_A.
  end for
  Set $\hat{E}^\star \leftarrow E_\psi$ and $\hat{A}^\star \leftarrow A$.
  Update θ *using* $\widehat{\nabla \mathcal{K}(\theta)}$ *from (5.13)*
end for

4 Sampling from GEBMs

A simple estimate of the empirical distribution of observations under the GEBM is via importance sampling (IS). This consists in first sampling multiple points from the base $\mathbb{G}$, and then re-weighting the samples according to the energy E. Although straightforward, this approach can lead to highly unreliable estimates, a well known problem in the Sequential Monte Carlo (SMC) literature which employs IS extensively [Doucet et al., 2001, Del Moral et al., 2006]. Other methods such as rejection sampling are known to be inefficient in high dimensions Haugh [2017]. Instead, we propose to sample from the posterior ν using MCMC. Recall from (5.5) that a sample x from $\mathbb{Q}$ is of the form $x = G(z)$ with z sampled from the *posterior latent* ν of (5.4) instead of the prior η. While sampling from η is often

straightforward (for instance if η is a Gaussian), sampling from ν is generally harder, due to dependence of its density on complex functions E and G. It is still possible to use MCMC methods to sample from ν, however, since we have access to its density up to a normalizing constant (5.4). In particular, we are interested in methods that exploit the gradient of ν, and consider two classes of samplers: *Overdamped samplers* and *Kinetic samplers*.

Overdamped samplers are obtained as a time-discretization of the *Overdamped Langevin dynamics*:

$$dz_t = (\nabla_z \log \eta(z_t) - \nabla_z E(G(z_t))) + \sqrt{2}\, \mathrm{d}w_t, \tag{5.14}$$

where w_t is a standard Brownian motion. The simplest sampler arising from (5.14) is the Unadjusted Langevin Algorithm (ULA):

$$Z_{k+1} = Z_k + \lambda \left(\nabla_z \log \eta(Z_k) - \nabla_z E(G(Z_k))\right) + \sqrt{2\lambda} W_{k+1}, \qquad Z_0 \sim \eta,$$

where $(W_k)_{k\geq 0}$ are i.i.d. standard Gaussians and λ is the step-size. For large k, Z_k is an approximate sample from ν [Raginsky et al., 2017, Proposition 3.3]. Hence, setting $X = G(Z_k)$ for a large enough k provides an approximate sample from the GEBM $\mathbb{Q}$, as summarized in Algorithm 2.

Algorithm 2 Overdamped Langevin Algorithm

Input λ, γ, u,η,E,G
Ouput X_T
$Z_0 \sim \eta$
for $t = 0, \ldots, T$ **do**
  $Y_{t+1} \leftarrow \nabla_z \log \eta(Z_t) - \nabla_z E \circ B(Z_t)$
  $W_{t+1} \sim \mathcal{N}(0, I)$
  $Z_{t+1} \leftarrow Z_t + \lambda Y_{t+1} + \sqrt{2\lambda} W_{t+1}$
end for
$X_T \leftarrow G(Z_T)$

Kinetic samplers arise from the *Kinetic Langevin dynamics* which introduce a momentum variable:

$$\begin{cases} \mathrm{d}z_t &= v_t \,\mathrm{d}t \\ \mathrm{d}v_t &= -\gamma v_t \,\mathrm{d}t + u\left(\nabla \log \eta(z_t) - \nabla E(G(z_t))\right)\mathrm{d}t + \sqrt{2\gamma u}\,\mathrm{d}w_t. \end{cases} \tag{5.15}$$

with friction coefficient $\gamma \geq 0$, inverse mass $u \geq 0$, **momentum** vector v_t and standard Brownian motion w_t. When the mass u^{-1} becomes negligible compared to the friction coefficient γ, i.e. $u\gamma^{-2} \approx 0$, standard results show that (5.15) recovers the Overdamped dynamics (5.14).

Algorithm 3 Kinetic Langevin Algorithm

Input λ, γ, u,η,E,G
Ouput X_T
$Z_0 \sim \eta$
for $t = 0, \dots, T$ **do**
  $Z_{t+1} \leftarrow Z_t + \frac{\lambda}{2} V_t$
  $Y_{t+1} \leftarrow \nabla_z \log \eta(Z_{t+1}) - \nabla_z E \circ B(Z_{t+1})$
  $V_{t+1} \leftarrow V_t + \frac{u\lambda}{2} Y_{t+1}$.
  $W_{t+1} \sim \mathcal{N}(0, I)$
  $\tilde{V}_{t+1} \leftarrow \exp(-\gamma\lambda) V_{t+\frac{1}{2}} + \sqrt{u\left(1 - \exp(-2\gamma\lambda)\right)} W_{t+1}$
  $V_{t+1} \leftarrow \tilde{V}_{t+1} + \frac{u\lambda}{2} Y_{t+1}$
  $Z_{t+1} \leftarrow Z_{t+1} + \frac{\lambda}{2} V_{t+1}$
end for
$X_T \leftarrow G(Z_T)$

Discretization in time of (5.15) leads to Kinetic samplers similar to Hamiltonian Monte Carlo [Cheng et al., 2017, Sachs et al., 2017]. We consider a particular algorithm from Sachs et al. [2017] which we call Kinetic Langevin Algorithm (KLA) Algorithm 3. Kinetic samplers were shown to better explore the modes of the invariant distribution ν compared to Overdamped ones (see [Neal, 2010, Betancourt et al., 2017] for empirical results and [Cheng et al., 2017] for theory), as also confirmed empirically in Section 6 .1 for image generation tasks using GEBMs. Next, we provide the following convergence result:

Proposition 35. *Assume that* $\log \eta(z)$ *is strongly concave and has a Lipschitz gradient, that* E, G *and their gradients are all* L*-Lipschitz. Set* $x_t = G(z_t)$*, where* z_t *is*

given by (5.15) and call $\mathbb{P}_t$ the probability distribution of x_t. Then $\mathbb{P}_t$ converges to $\mathbb{Q}$ in the Wasserstein sense,

$$W_2(\mathbb{P}_t, \mathbb{Q}) \leq LCe^{-c\gamma t},$$

where c and C are positive constants independent of t, with $c = O(\exp(-dim(\mathcal{Z})))$.

Proposition 35 is proved in Section A .2 using [Eberle et al., 2017, Corollary 2.6], and implies that $(x_t)_{t\geq 0}$ converges at the same speed as $(z_t)_{t\geq 0}$. When the dimension q of $\mathcal{Z}$ is orders of magnitude smaller than the input space dimension d, the process $(x_t)_{t\geq 0}$ converges faster than typical sampling methods on $\mathcal{X}$, for which the exponent controlling the convergence rate is of order $O(\exp(-d))$.

Tempered GEBM. It can be preferable to sample from a *tempered* version of the model by rescaling the energy E by an *inverse temperature* parameter β, thus effectively sampling from $\exp^{-\beta E(x)} \mathrm{d}\mathbb{Q}(x)$. *High temperature* regimes ($\beta \to 0$) recover the base model $\mathbb{G}$ while *low temperature* regimes ($\beta \to \infty$) essentially sample from minima of the energy E. As shown in Section 6 , low temperatures tend to produce better sample quality for natural image generation tasks.

5 Related work

Energy based models. Usually, energy based models are required to have a density w.r.t. to a Lebesgue measure, and do not use a learnable base measure; in other words, models are supported on the whole space. Various methods have been proposed in the literature to learn EBMs. *Contrastive Divergence* [Hinton, 2002] approximates the gradient of the log-likelihood by sampling from the energy model with MCMC. More recently, [Belanger and McCallum, 2016, Xie et al., 2016, 2017, 2018a, 2019, Tu and Gimpel, 2018, Du and Mordatch, 2019, Deng et al., 2020] extend the idea using more sophisticated models and MCMC sampling strategies that lead to higher quality estimators. *Score Matching* [Hyvärinen, 2005] calculates an alternative objective (the *score*) to the log-likelihood which is independent of the partition function, and was recently used in the context non-parametric energy functions to provide estimators of the energy that are provably consistent as in Chapter 3 and

[Sriperumbudur et al., 2017, Sutherland et al., 2018, Wenliang et al., 2019]). In *Noise-Contrastive Estimation* [Gutmann and Hyvärinen, 2012], a classifier is trained to distinguish between samples from a fixed proposal distribution and the target $\mathbb{P}$. This provides an estimate for the density ratio between the optimal energy model and the proposal distribution. In a similar spirit, Cranmer et al. [2016] uses a classifier to learn likelihood ratios. Conversely, Grathwohl et al. [2020] interprets the logits of a classifier as an energy model obtained after marginalization over the classes. The resulting model is then trained using Contrastive Divergence. In more recent work, Dai et al. [2019a,b] exploit a dual formulation of the logarithm of the partition function as a supremum over the set of all probability distributions of some functional objective. Yu et al. [2020] explore methods for using general f-divergences, such as Jensen-Shannon, to train EBMs.

Generative Adversarial Networks. Recent work proposes using the discriminator of a trained GAN to improve the generator quality. Rejection sampling [Azadi et al., 2019] and Metropolis-Hastings correction [Turner et al., 2019, Neklyudov et al., 2019] perform sampling directly on the high-dimensional input space without using gradient information provided by the discriminator. Moreover, the data distribution is assumed to admit a density w.r.t. the generator. Ding et al. [2019] perform sampling on the feature space of some auxiliary pre-trained network; while Lawson et al. [2019] treat the sampling procedure as a model on its own, learned by maximizing the ELBO. In our case, no auxiliary model is needed. In the present work, sampling doesn't interfere with training, in contrast to recently considered methods to optimize over the latent space during training Wu et al. [2019a,b]. In Tanaka [2019], the discriminator is viewed as an optimal transport map between the generator and the data distribution and is used to compute optimized samples from latent space. This is in contrast to the diffusion-based sampling that we consider. In [Xie et al., 2018b,c], two independent models, a full support EBM and a generator network, are trained cooperatively using MCMC. By contrast, in the present work, the energy and base are part of the same model, and the model support is lower-dimensional than the target space $\mathcal{X}$. While we do not address the mode collapse problem, Xu et al. [2018],

Nguyen et al. [2017] showed that KL-based losses are resilient to it thanks to the zero-avoiding property of the KL, a good sign for KALE which is derived from KL by Fenchel duality.

The closest related approach appears in a study concurrent to the present work [Che et al., 2020], where the authors propose to use Langevin dynamics on the latent space of a GAN generator, but with a different discriminator to ours (derived from the Jensen-Shannon divergence or a Wasserstein-based divergence). Our theory results showing the existence of the loss gradient (Theorem 34), establishing weak convergence of distributions under KALE (Proposition 33), and demonstrating consistency of the KALE estimator (Section C .1) should transfer to the JS and Wasserstein criteria used in that work. Subsequent to the present work, an alternative approach has been recently proposed, based on normalising flows, to learn both the low-dimensional support of the data and the density on this support [Brehmer and Cranmer, 2020]. This approach maximises the explicit likelihood of a data projection onto a learned manifold, and may be considered complementary to our approach.

6 Experiments

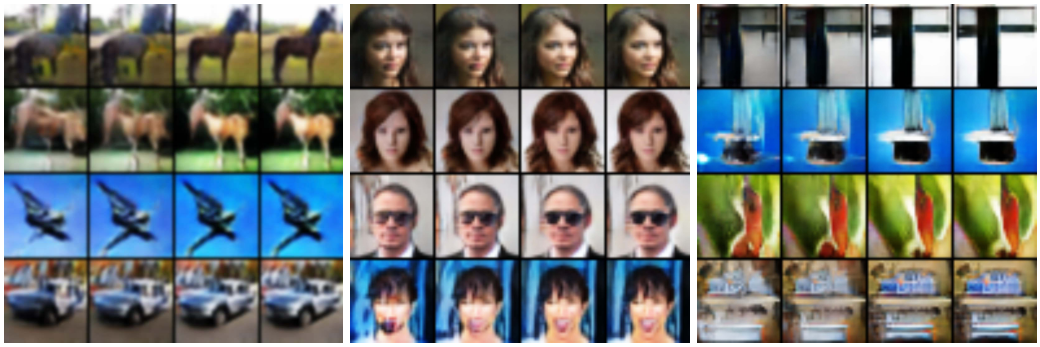


Figure 5.2: Samples at different iterations of the MCMC chain of Algorithm 3 (left to right).

6 .1 Image generation.

Experimental setting. We train a GEBM on **unsupervised** image generation tasks, and compare the quality of generated samples with other methods using the FID score [Heusel et al., 2017] computed on 5×10^4 generated samples. We consider `CIFAR-10` [Krizhevsky, 2009], LSUN [Yu et al., 2015], CelebA [Liu et al., 2015b] and ImageNet [Russakovsky et al., 2014] all downsampled to 32x32 resolution

to reduce computational cost. We consider two network architectures for each of the base and energy, a smaller one (SNGAN ConvNet) and a larger one (SNGAN ResNet), both of which are from Miyato et al. [2018]. For the base we used the SNGAN generator networks from Miyato et al. [2018] with a 100-dimensional Gaussian for the latent noise η. For the energy we used the SNGAN discriminator networks from Miyato et al. [2018]. (Details of the networks in Section B .2). We train the models for 150000 generator iterations using Algorithm 1. After training is completed, we rescale the energy by $\beta = 100$ to get a **colder** version of the GEBM and sample from it using either Algorithm 2 (ULA) or Algorithm 3 (KLA) with parameters $(\gamma = 100, u = 1)$. We perform 1000 MCMC iterations with initial step-size of $\lambda = 10^{-4}$ decreased by 10 every 200 iterations. As a baseline we consider samples generated from the base of the GEBM only (without using information from the energy) and call this KALE-GAN. More details are given in Section B .

Results: Table 5.1 shows that GEBM outperforms both KALE and standard GANs when using the same networks for the base/generator and energy/critic. Moreover, KALE-GAN matches the performance of a standard GAN (with Jensen-Shannon critic), showing that the improvement of GEBM cannot be explained by the switch from Jensen-Shannon to a KALE-based critic. Rather, the improvement is largely due to incorporating the energy function into the model, and sampling using Algorithm 3.

This finding experimentally validates our claim that incorporating the energy improves the model, and that all else being equal, a GEBM outperforms a GAN with the same generator and critic architecture. Indeed, if the critic is not zero at convergence, then by definition it contains information on the remaining mismatch between the generator (base) and data mass, which the GEBM incorporates, but the GAN does not. The GEBM also outperforms an EBM even when the latter was trained using a larger network (ResNet) with supervision (S) on ImageNet, which is an easier task (Chen et al. [2019]). More comparisons on Cifar10 and ImageNet are provided in Table 5.2.

Table 5.3 shows different sampling methods using the same trained networks (generator and critic), with KALE-GAN as a baseline. All energy-exploiting methods

	SNGAN (ConvNet)			SNGAN (ResNet)			
	GEBM	KALE-GAN	GAN	GEBM	KALE-GAN	GAN	EBM
Cifar10	23.02	32.03	29.9	**19.31**	20.19	21.7	38.2
ImageNet	**13.94**	19.37	20.66	20.33	21.00	20.50	14.31 (S)

Table 5.1: FID scores for two versions of SNGAN from [Miyato et al., 2018] on Cifar10 and ImageNet. GEBM: training using Algorithm 1 and sampling using Algorithm 3. KALE-GAN: Only the base of a GEBM is retained for sampling. GAN: training as in [Miyato et al., 2018] with $q = 128$ for the latent dimension as it worked best. EBM: results from Du and Mordatch [2019] with *supervised* training on ImageNet (S).

Model	FID
Cifar10 Unsupervised	
PixelCNN Oord et al. [2016]	65.93
PixelIQN Ostrovski et al. [2018]	49.46
EBM Radford et al. [2016]	38.2
WGAN-GP Gulrajani et al. [2017]	36.4
NCSN Ho and Ermon [2016]	25.32
SNGAN Miyato et al. [2018]	21.7
GEBM (ours)	**19.31**
Cifar10 Supervised	
BigGAN Donahue and Simonyan [2019]	14.73
SAGAN Zenke et al. [2017]	13.4
ImageNet Supervised	
PixelCNN	33.27
PixelIQN	22.99
EBM	14.31
ImageNet Unsupervised	
SNGAN	20.50
GEBM (ours)	**13.94**

Table 5.2: FID scores on ImageNet and CIFAR-10.

outperform the unmodified KALE-GAN with the same architecture. That said, our method (both ULA and KLA) outperforms both (IHM) [Turner et al., 2019] and (DOT) [Tanaka, 2019], which both use the energy information.

Effect of the temperature and sampler convergence. Using a colder temperature leads to an improved FID score, and needs relatively few MCMC iterations, as shown

	Cifar10	LSUN	CelebA	ImageNet
KALE-GAN	32.03	21.67	6.91	19.37
IHM	30.47	20.63	6.39	18.15
DOT	26.35	20.41	5.93	16.21
GEBM (ULA)	**23.02**	16.23	**5.21**	14.00
GEBM (KLA)	24.29	**15.25**	5.38	**13.94**

Table 5.3: FID scores for different sampling methods using the same trained SNGAN (ConvNet): KALE-GAN as a baseline w/o critic information.

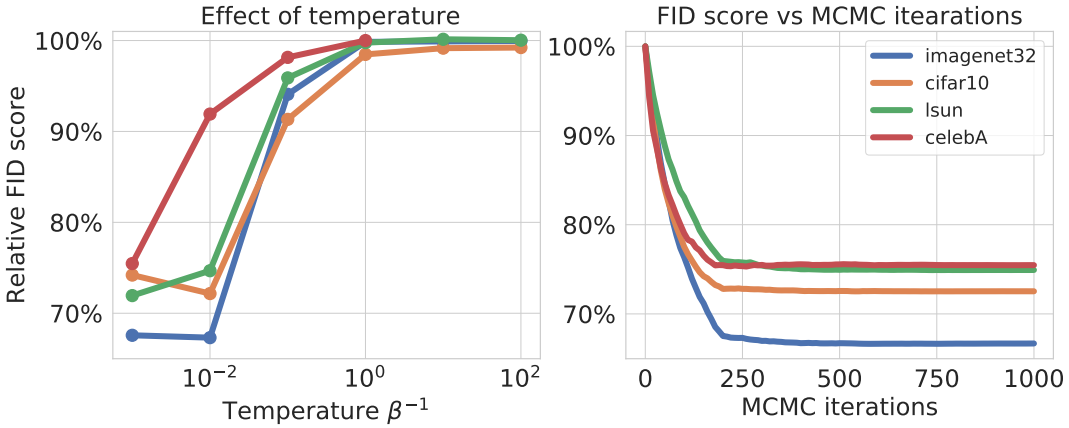


Figure 5.3: Relative FID score: ratio between FID score of the GEBM $\mathbb{Q}_{\mathbb{G},E}$ and its base $\mathbb{G}$. (Left) Evolution of the ratio for increasing temperature on the 4 datasets after 1000 iterations of (5.15). (Right) Evolution of the same ratio during MCMC iteration using (5.15) for $\beta = 100$.

in Figure 5.3. Sampler convergence to visually plausible modes at low temperatures is demonstrated in Figure 5.2.

Mode exploration: KLA vs ULA sampler. In Table 5.3, KLA was used in the high friction regime $\gamma = 100$ and thus behaves like ULA. This allows to obtain sharper samples concentrated around the modes of the GEBM thus improving the FID score. If, instead, the goal is to encourage more exploration of the modes of the GEBM, then KLA with a smaller γ is a better alternative than ULA. Figures 5.4 and 5.5 show sample trajectories using Algorithm 3 with no friction $\gamma = 0$ for the 4 datasets. It is clear that along the same MCMC chain, several image modes are explored. We also notice the transition from a mode to another happens almost at the same time for all chains and corresponds to the gray images. This is unlike Langevin or when the friction coefficient γ is large as in Figure 5.6. In that case each chain remains

within the same mode.

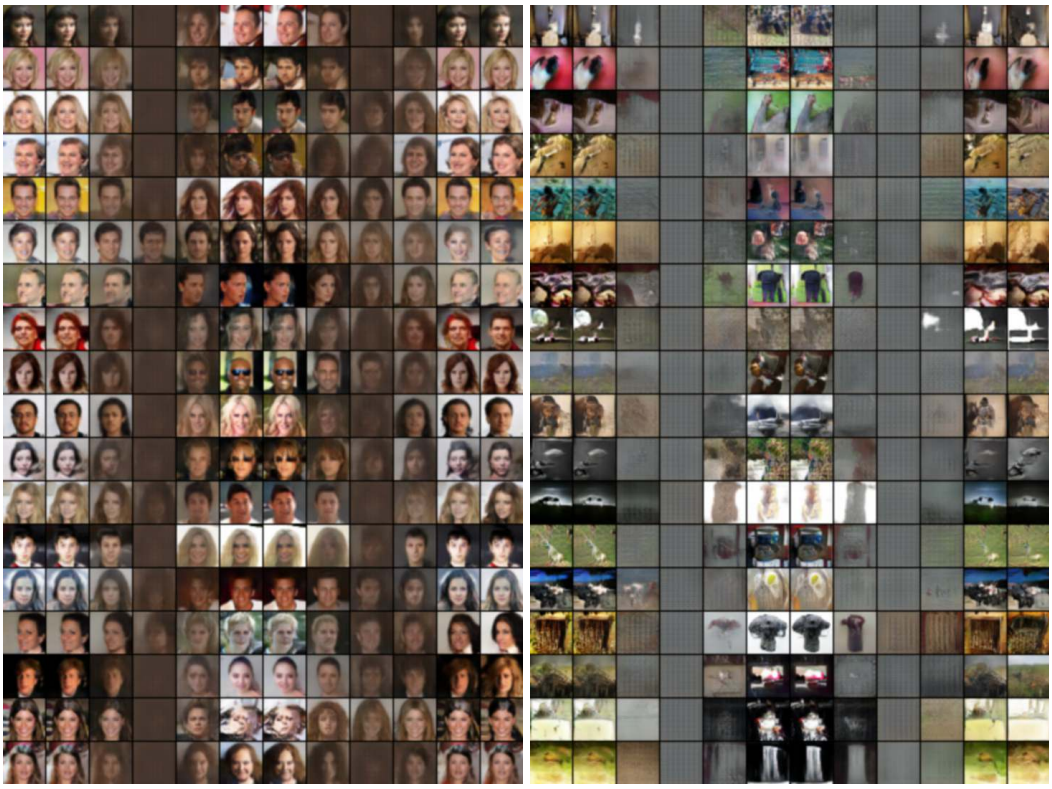


Figure 5.4: Samples from the GEBM at different stages of sampling using Algorithm 3 and inverse temperature $\beta = 1$, on CelebA (Left), Imagenet (Right). Each row represents a sampling trajectory from early stages (leftmost images) to later stages (rightmost images). The samples were obtained in the low friction regime of Algorithm 3 $\gamma \simeq 0$ which yields a near conservation of the total hamiltonian and thus exhibits near periodic trajectories within the chains as illustrated by the samples jumping between two modes.

6 .2 Density Estimation

Motivation. We next consider the particular setting where the likelihood of the model is well-defined, and admits a closed form expression. This is intended principally as a sanity check that our proposed training method in Algorithm 1 succeeds in learning maximum likelihood solutions. Outside of this setting, closed form expressions of the normalizing constant are not available for generic GEBMs. While this is not an issue (since the proposed method doesn't require a closed form expression for the normalizing constant), in this experiment only, we want to have access to closed form expressions, as they enable a direct comparison with other density estimation methods.

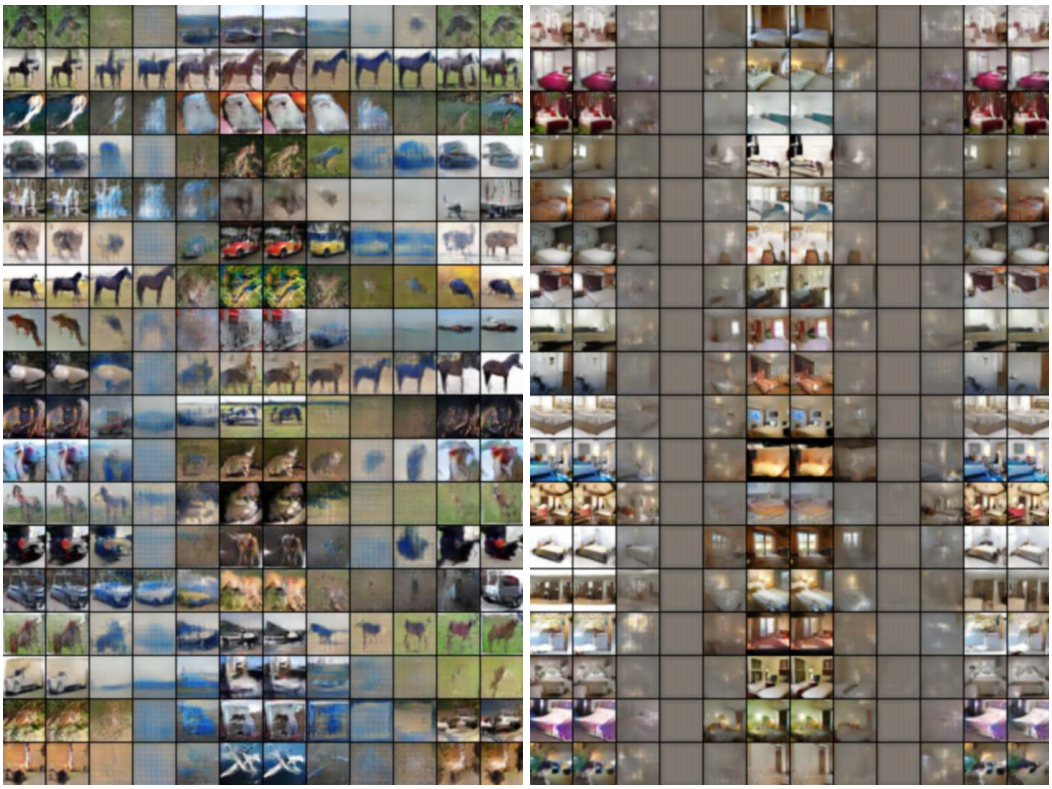


Figure 5.5: Samples from the GEBM at different stages of sampling using Algorithm 3 and inverse temperature $\beta = 1$, on Cifar10 and LSUN (Right). Each row represents a sampling trajectory from early stages (leftmost images) to later stages (rightmost images).

Experimental setting. To have a closed-form likelihood, we consider the case where the dimension of the latent space is equal to data-dimension, and choose the base $\mathbb{G}$ of the GEBM to be a Real NVP (Ding et al. [2019]) with density $\exp(-r(x))$ and energy $E(x) = h(x) - r(x)$. Thus, in this particular case, the GEBM has a well defined likelihood over the whole space, and we are precisely in the setting of Proposition 31, which shows that this GEBM is equal to an EBM with density proportional to $exp(-h)$. We further require the EBM to be a second Real NVP so that its density has a closed form expression. We consider 5 UCI datasets for which we use the same pre-processing as in [Wenliang et al., 2019]. For comparison, we train the EBM by direct maximum likelihood (ML) and contrastive divergence (CD). To train the GEBM, we use Algorithm 1, which doesn't directly exploit the closed-form expression of the likelihood (unlike direct ML). We thus use either (5.8) (KALE-DV) or (5.9) (KALE-F) to estimate the normalizing constant. More details

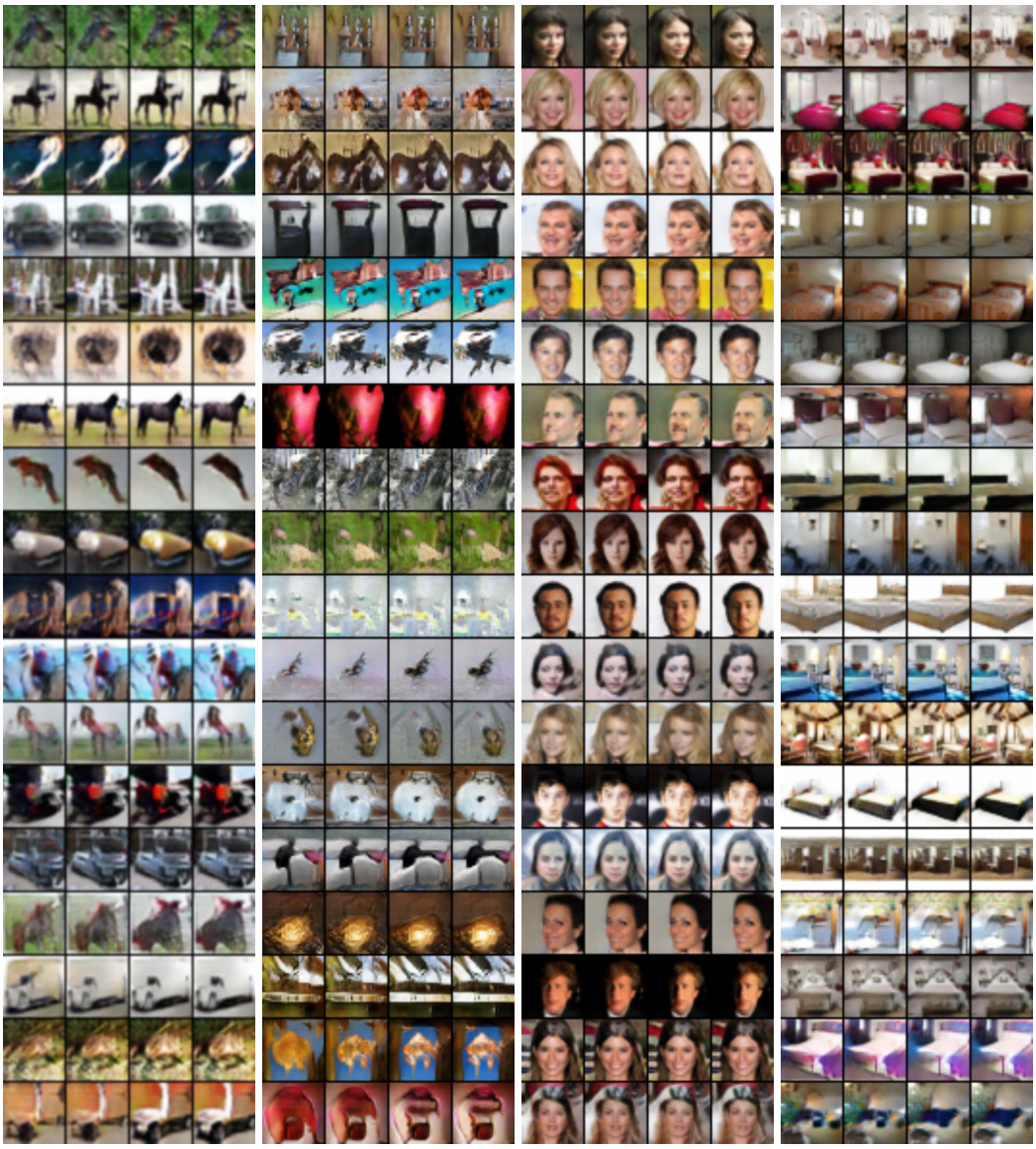


Figure 5.6: Samples from the tempered GEBM at different stages of sampling using langevin and inverse temperature $\beta = 100$, on Cifar10 (Left), Imagenet (Middle-left), CelebA (Middle-Right) and LSUN (Right). Each row represents a sampling trajectory from early stages (leftmost images) to later stages (rightmost images).

are given in Section B .3.

Results. Table 5.4 reports the Negative Log-Likelihood (NLL) evaluated on the test set and corresponding to the best performance on the validation set. Training the GEBM using Algorithm 1 leads to comparable performance to (CD) and (ML).

Amortized estimation of the normalizing constant. Figure Figure 5.7 (left) shows the error in the estimation of the log-partition function using both methods (KALE-

	RedWine $d = 11$ $N \sim 10^3$	**Whitewine** $d = 11$ $N \sim 10^3$	**Parkinsons** $d = 15$ $N \sim 10^3$	**Hepmass** $d = 22$ $N \sim 10^5$	**Miniboone** $d = 43$ $N \sim 10^4$
NVP w ML	11.98	13.05	14.5	24.89	42.28
NVP w CD	11.88	13.01	14.06	**22.89**	39.36
NVP w KALE (DV)	11.6	12.77	**13.26**	26.56	46.48
NVP w KALE (F)	**11.19**	**12.66**	**13.26**	24.66	**38.35**

Table 5.4: UCI datasets: Negative log-likelihood computed on the test set and corresponding to the best performance on the validation set. Best method in boldface.

DV and KALE-F). (KALE-F) leads to more accurate estimates of the log-partition function, with a relative error of order 0.1% compared to 10% for (KALE-DV). This result illustrates the advantage of performing an amortized estimation of the log-partition function (KALE-F) rather than directly estimating it on small batch sizes (KALE-DV).

Figure Figure 5.7 (right) shows the evolution of the negative log-likelihood (NLL) on both training and test sets per epochs for RedWine and Whitewine datasets. The error decreases steadily in the case of KALE-DV and KALE-F while the error gap between the training and test set remains controlled. Larger gaps are observed for both direct maximum likelihood estimation and Contrastive divergence although the training NLL tends to decrease faster than for KALE.

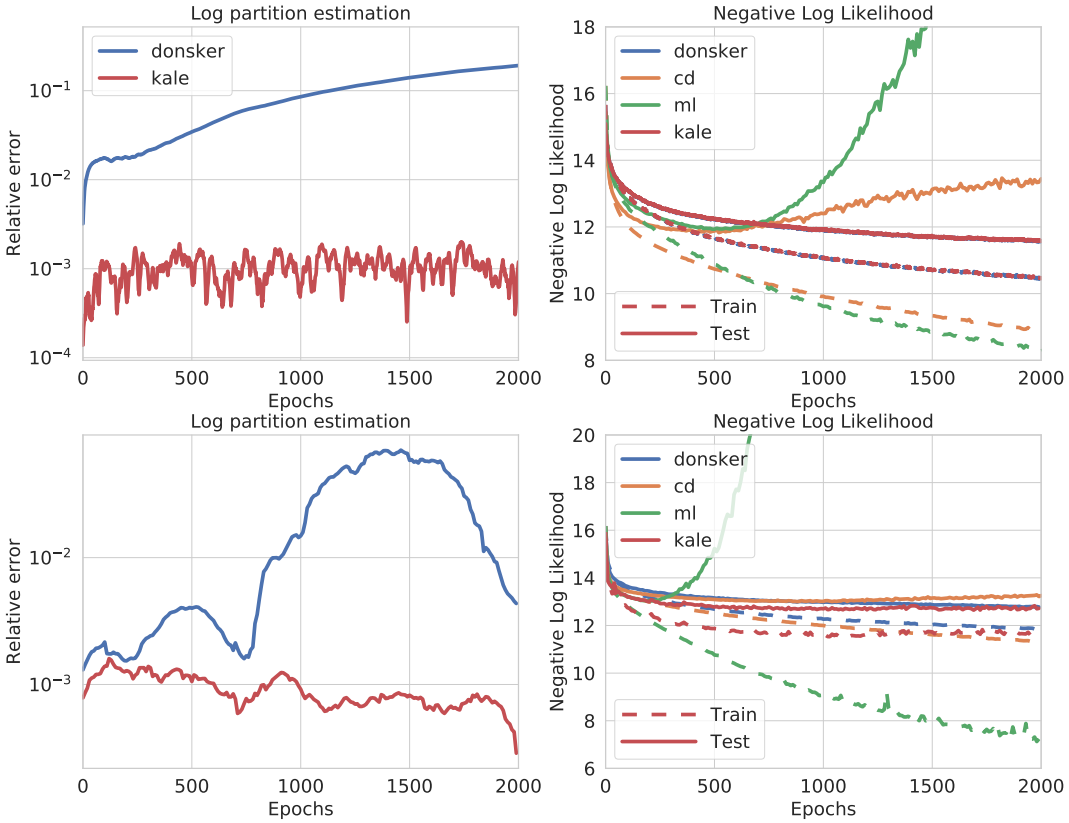


Figure 5.7: (Left): Relative error $\frac{|\hat{c}-c^\star|}{|\hat{c}|+|c^\star|}$ on the estimation of the ground truth log-partition function c^* by $\hat{c}$ using either KALE-DV or KALE-F vs training Epochs on RedWine (Top) and WhiteWine (Bottom) datasets. In both cases the batch-size is 100. (Right): Negative log likelihood vs training epochs on both training and test set for 4 different learning methods (KALE-DV,KALE-F, CD and ML) on RedWine dataset.

Supplementary

A Proofs

A .1 Topological and smoothness properties of KALE

Topological properties of KALE. Denseness and smoothness of the energy class $\mathcal{E}$ are the key to guarantee that KALE is a reliable criterion for measuring convergence. We thus make the following assumptions on $\mathcal{E}$:

(**A**) For all $E \in \mathcal{E}$, $-E \in \mathcal{E}$ and there is $C_E > 0$ such that $cE \in \mathcal{E}$ for $0 \leq c \leq C_E$. For any continuous function g, any compact support K in $\mathcal{X}$ and any precision $\epsilon > 0$, there exists a finite linear combination of energies $G = \sum_{i=1}^{r} a_i E_i$ such that $\sup_{x \in K} |f(x) - G(x)| \leq \epsilon$.

(**B**) All energies E in $\mathcal{E}$ are Lipschitz in their input with the same Lipschitz constant $L > 0$.

Assumption (**A**) holds in particular when $\mathcal{E}$ contains feedforward networks with a given number of parameters. In fact networks with a single neuron are enough, as shown in [Zhang et al., 2017, Theorem 2.3]. Assumption (**B**) holds when additional regularization of the energy is enforced during training by methods such as **spectral normalization** Miyato et al. [2018] or **gradient penalty** Gulrajani et al. [2017] as done in Section 6 . Proposition 33 states the topological properties of KALE ensuring that it can be used as a criterion for weak convergence. A proof is given in Section A .1.1 and is a consequence of [Zhang et al., 2017, Theorem B.1].

Proposition 36. *Under Assumptions (**A**) and (**B**) it holds that:*

1. *KALE*$(\mathbb{P}||\mathbb{G}) \geq 0$ *with KALE*$(\mathbb{P}||\mathbb{G}) = 0$ *if and only if* $\mathbb{P} = \mathbb{G}$.

2. *KALE*$(\mathbb{P}||\mathbb{G}^n) \to 0$ *if and only if* $\mathbb{G}^n \to \mathbb{P}$ *under the weak topology.*

A .1.1 Topological properties of KALE

In this section we prove Proposition 33. We first start by recalling the required assumptions and make them more precise:

Assumption 1. *Assume the following holds:*

- *The set* $\mathcal{X}$ *is compact.*
- *For all* $E \in \mathcal{E}$, $-E \in \mathcal{E}$ *and there is* $C_E > 0$ *such that* $cE \in \mathcal{E}$ *for* $0 \leq c \leq C_E$. *For any continuous function* g, *any compact support* K *in* $\mathcal{X}$ *and any precision* $\epsilon > 0$, *there exists a finite linear combination of energies* $G = \sum_{i=1}^{r} a_i E_i$ *such that* $|f(x) - G(x)| \leq \epsilon$ *on* K.
- *All energies* E *in* $\mathcal{E}$ *are Lipschitz in their input with the same Lipschitz constant* $L > 0$.

For simplicity we consider the set $\mathcal{H} = \mathcal{E} + \mathbb{R}$, i.e.: $\mathcal{H}$ is the set of functions h of the form $h = E + c$ where $E \in \mathcal{E}$ and $c \in \mathbb{R}$. In all what follows $\mathcal{P}_1$ is the set of probability distributions with finite first order moments. We consider the notion of weak convergence on $\mathcal{P}_1$ as defined in [Villani, 2009, Definition 6.8] which is equivalent to convergence in the Wasserstein-1 distance W_1.

Proof of Proposition 33 . We proceed by proving the **separation** properties (1^{st} statement), then the **metrization of the weak topology** (2^{nd} statement).

Separation. We have by Assumption 1 that $0 \in \mathcal{E}$, hence by definition $\text{KALE}(PP||\mathbb{G}) \geq \mathcal{F}_{\mathbb{P},\mathbb{G}}(0) = 0$. On the other hand, whenever $\mathbb{P} = \mathbb{G}$, it holds that:

$$\mathcal{F}_{\mathbb{P},\mathbb{G}}(h) = -\int (\exp(-h) + h - 1)\, d\mathbb{P}, \qquad \forall h \in \mathcal{H}.$$

Moreover, by convexity of the exponential, we know that $\exp(-x) + x - 1 \geq 0$ for all $x \in \mathbb{R}$. Hence, $\mathcal{F}_{\mathbb{P},\mathbb{G}}(h) \leq \mathcal{F}_{\mathbb{P},\mathbb{G}}(0) = 0$ for all $h \in \mathcal{H}$. This directly implies that $\text{KALE}(\mathbb{P}|\mathbb{G}) = 0$. For the converse, we will use the same argument as in the proof

of [Zhang et al., 2017, Theorem B.1]. Assume that $\text{KALE}(\mathbb{P}|\mathbb{G}) = 0$ and let h be in $\mathcal{H}$. By Assumption 1, there exists $C_h > 0$ such that $ch \in \mathcal{H}$ and we have:

$$\mathcal{F}(ch) \leq \text{KALE}(\mathbb{P}||\mathbb{G}) = 0.$$

Now dividing by c and taking the limit to 0, it is easy to see that $-\int h \, \mathrm{d}\mathbb{P} + \int h \, \mathrm{d}\mathbb{G} \leq 0$. Again, by Assumption 1, we also know that $-h \in \mathcal{H}$, hence, $\int h \, \mathrm{d}\mathbb{P} - \int h \, \mathrm{d}\mathbb{G} \leq 0$. This necessarily implies that $\int h \, \mathrm{d}\mathbb{P} - \int h \, \mathrm{d}\mathbb{G} = 0$ for all $h \in \mathcal{H}$. By the density of $\mathcal{H}$ in the set continuous functions on compact sets, we can conclude that the equality holds for any continuous and bounded function, which in turn implies that $\mathbb{P} = \mathbb{G}$.

Metrization of the weak topology. We first show that for any $\mathbb{P}$ and $\mathbb{G}$ with finite first moment, it holds that $\text{KALE}(\mathbb{P}|\mathbb{G}) \leq LW_1(\mathbb{P}, \mathbb{G})$, where $W_1(\mathbb{P}, \mathbb{G})$ is the Wasserstein-1 distance between $\mathbb{P}$ and $\mathbb{G}$. For any $h \in \mathcal{H}$ the following holds:

$$\begin{aligned}
\mathcal{F}(h) &= -\int h d\mathbb{P} - \int \exp(-h) d\mathbb{G} + 1 \\
&= \int h(x) d\mathbb{G}(x) - h(x') d\mathbb{P}(x') \\
&\quad - \int \underbrace{(\exp(-h) + h - 1)}_{\geq 0} d\mathbb{G} \\
&\leq \int h(x) d\mathbb{G}(x) - h(x') d\mathbb{P}(x') \leq LW_1(\mathbb{P}, \mathbb{G})
\end{aligned}$$

The first inequality results from the convexity of the exponential while the last one is a consequence of h being L-Lipschitz. This allows to conclude that $\text{KALE}(\mathbb{P}||\mathbb{G}) \leq LW_1(\mathbb{P}, \mathbb{G})$ after taking the supremum over all $h \in \mathcal{H}$. Moreover, since W_1 metrizes the weak convergence on $\mathcal{P}_1$ [Villani, 2009, Theorem 6.9], it holds that whenever a sequence $\mathbb{G}^n$ converges weakly towards $\mathbb{P}$ in $\mathcal{P}_1$ we also have $W_1(\mathbb{P}, \mathbb{G}^n) \to 0$ and thus $\text{KALE}(\mathbb{P}||\mathbb{G}^n) \to 0$. The converse is a direct consequence of [Liu et al., 2017, Theorem 10] since by assumption $\mathcal{X}$ is compact. □

A .1.2 Smoothness properties of KALE

We will now prove Theorem 34. We begin by stating the assumptions that will be used in this section:

(I) $\mathcal{E}$ is parametrized by a compact set of parameters Ψ.

(II) Functions in $\mathcal{E}$ are jointly continuous w.r.t. (ψ, x) and are L-lipschitz and L-smooth w.r.t. the input x:

$$\|E_\psi(x) - E_\psi(x')\| \leq L_e\|x - x'\|,$$
$$\|\nabla_x E_\psi(x) - \nabla_x E_\psi(x')\| \leq L_e\|x - x'\|.$$

(III) $(\theta, z) \mapsto G_\theta(z)$ is jointly continuous in θ and z, with $z \mapsto G_\theta(z)$ uniformly Lipschitz w.r.t. z:

$$\|G_\theta(z) - G_\theta(z')\| \leq L_b\|z - z'\|, \qquad \forall z, z' \in \mathcal{Z}, \theta \in \Theta.$$

There exists non-negative functions a and b defined from $\mathcal{Z}$ to $\mathbb{R}$ such that $\theta \mapsto G_\theta(z)$ are a-Lipschitz and b-smooth in the following sense:

$$\|G_\theta(z) - G_{\theta'}(z)\| \leq a(z)\|\theta - \theta'\|,$$
$$\|\nabla_\theta G_\theta(z) - \nabla_\theta G_{\theta'}(z)\| \leq b(z)\|\theta - \theta'\|.$$

Moreover, a and b are integrable in the following sense:

$$\int a(z)^2 \exp(2L_e L_b\|z\|)d\eta(z) < \infty, \qquad \int \exp(L_e L_b\|z\|)d\eta(z) < \infty,$$
$$\int b(z) \exp(L_e L_b\|z\|)d\eta(z) < \infty.$$

To simplify notation, we will denote by $\mathcal{L}_\theta(f)$ the expected $\mathbb{G}_\theta$ log-likelihood under $\mathbb{P}$. In other words,

$$\mathcal{L}_\theta(E) := \mathcal{L}_{\mathbb{P},\mathbb{G}_\theta}(E) = -\int E d\mathbb{P} - \log \int \exp(-E) d\mathbb{G}_\theta.$$

We also denote by $p_{E,\theta}$ the density of the model w.r.t. $\mathbb{G}_\theta$,

$$p_{E,\theta} = \frac{\exp(-E)}{Z_{\mathbb{G}_\theta,E}}, \qquad Z_{\mathbb{G}_\theta,E} = \int \exp(-E) d\mathbb{G}_\theta.$$

We write $\mathcal{K}(\theta) := \text{KALE}(\mathbb{P}||\mathbb{G}_\theta)$ to emphasize the dependence on θ.

Proof of Theorem 34. To show that sub-gradient methods converge to local optima, we only need to show that $\mathcal{K}$ is Lipschitz continuous and weakly convex. This directly implies convergence to local optima for sub-gradient methods, according to Davis and Drusvyatskiy [2018], Thekumparampil et al. [2019]. Lipschitz continuity ensures that $\mathcal{K}$ is differentiable for almost all $\theta \in \Theta$, and weak convexity simply means that there exits some positive constant $C \geq 0$ such that $\theta \mapsto \mathcal{K}(\theta) + C\|\theta\|^2$ is convex. We now proceed to show these two properties.

We will first prove that $\theta \mapsto \mathcal{K}(\theta)$ is weakly convex in θ. By Lemma 37, we know that for any $E \in \mathcal{E}$, the function $\theta \mapsto \mathcal{L}_\theta(E)$ is M-smooth for the same positive constant M. This directly implies that it is also weakly convex and the following inequality holds:

$$\mathcal{L}_{\theta_t}(E) \leq t\mathcal{L}_\theta(E) + (1-t)\mathcal{L}_{\theta'}(E) + \frac{M}{2}t(1-t)\|\theta - \theta'\|^2.$$

Taking the supremum w.r.t. E, it follows that

$$\mathcal{K}(\theta_t) \leq t\mathcal{K}(\theta) + (1-t)\mathcal{K}(\theta') + \frac{M}{2}t(1-t)\|\theta - \theta'\|^2.$$

This means precisely that $\mathcal{K}$ is weakly convex in θ.

To prove that $\mathcal{K}$ is Lipschitz, we will also use Lemma 37, which states that $\mathcal{L}_\theta(E)$ is Lipschitz in θ uniformly on $\mathcal{E}$. Hence, the following holds:

$$\mathcal{L}_\theta(E) \leq \mathcal{L}_\theta(E) + LC\|\theta - \theta'\|.$$

Again, taking the supremum over E, it follows directly that

$$\mathcal{K}(\theta) \leq \mathcal{K}(\theta') + LC\|\theta - \theta'\|.$$

We conclude that $\mathcal{K}$ is Lipschitz by exchanging the roles of θ and θ' to get the other side of the inequality. Hence, by the Rademacher theorem, $\mathcal{K}$ is differentiable for almost all θ.

We will now provide an expression for the gradient of $\mathcal{K}$. By Lemma 38 we know that $\psi \mapsto \mathcal{L}_\theta(E_\psi)$ is continuous and by Assumption **(I)** Ψ is compact. Therefore, the supremum $\sup_{E\in\mathcal{E}} \mathcal{L}_\theta(E)$ is achieved for some function $E^\star_\theta$. Moreover, we know by Lemma 37 that $\mathcal{L}_\theta(E)$ is smooth uniformly on $\mathcal{E}$, therefore the family $(\partial_\theta \mathcal{L}_\theta(E))_{E\in\mathcal{E}}$ is equi-differentiable. We are in position to apply Milgrom and Segal [2002](Theorem 3) which ensures that $\mathcal{K}(\theta)$ admits left and right partial derivatives given by

$$\begin{aligned} \partial_e^+ \mathcal{K}(\theta) &= \lim_{\substack{t>0 \\ t\to 0}} \partial_\theta \mathcal{L}_\theta(E^\star_{\theta+te})^\top e, \\ \partial_e^- \mathcal{K}(\theta) &= \lim_{\substack{t<0 \\ t\to 0}} \partial_\theta \mathcal{L}_\theta(E^\star_{\theta+te})^\top e, \end{aligned} \tag{5.16}$$

where e is a given direction in $\mathbb{R}^r$. Moreover, the theorem also states that $\mathcal{K}(\theta)$ is differentiable iff $t \mapsto E^\star_{\theta+te}$ is continuous at $t = 0$. Now, recalling that $\mathcal{K}(\theta)$ is actually differentiable for almost all θ, it must hold that $E^\star_{\theta+te} \to_{t\to 0} E^\star_\theta$ and $\partial_e^+ \mathcal{K}(\theta) = \partial_e^- \mathcal{K}(\theta)$ for almost all θ. This implies that the two limits in (5.16) are actually equal to $\partial_\theta \mathcal{L}_\theta(E^\star_\theta)^\top e$. The gradient of $\mathcal{K}$, whenever defined, in therefore given by

$$\nabla_\theta \mathcal{K}(\theta) = Z^{-1}_{\mathbb{G}_\theta, E^\star_\theta} \int \nabla_x E^\star_\theta(G_\theta(z)) \nabla_\theta G_\theta(z) \exp(-E^\star_\theta(G_\theta(z)))\eta(z)\,\mathrm{d}z.$$

□

Lemma 37. *Under Assumptions **(I)** to **(III)**, the functional $\mathcal{L}_\theta(E)$ is Lipschitz and*

smooth in θ uniformly on $\mathcal{E}$:

$$|\mathcal{L}_\theta(E) - \mathcal{L}_{\theta'}(E)| \leq LC\|\theta - \theta'\|,$$
$$\|\partial_\theta \mathcal{L}_\theta(E) - \partial_\theta \mathcal{L}_{\theta'}(E))\| \leq 2CL(1+L)\|\theta - \theta'\|.$$

Proof. By Lemma 38, we have that $\mathcal{L}_\theta(E)$ is differentiable, and that

$$\partial_\theta \mathcal{L}_\theta(E) := \int (\nabla_x E \circ G_\theta)\, \nabla_\theta G_\theta\, (p_{E,\theta} \circ G_\theta)\, d\eta.$$

Lemma 38 ensures that $\|\partial_\theta \mathcal{L}_\theta(E)\|$ is bounded by some positive constant C that is independent from E and θ. This implies in particular that $\mathcal{L}_\theta(E)$ is Lipschitz with a constant C. We will now show that it is also smooth. For this, we need to control the difference

$$D := \|\partial_\theta \mathcal{L}_\theta(E) - \partial_\theta \mathcal{L}_{\theta'}(E)\|.$$

We have by triangular inequality:

$$\begin{aligned} D \leq & \underbrace{\int \|\nabla_x E \circ G_\theta - \nabla_x E \circ G_{\theta'}\|\, \|\nabla_\theta G_\theta\|\, (p_{E,\theta} \circ G_\theta)\, d\eta}_{I} \\ & + \underbrace{\int \|\nabla_x E \circ G_\theta\| \|\nabla_\theta G_\theta - \nabla_\theta G_{\theta'}\|\, (p_{E,\theta} \circ G_\theta)\, d\eta}_{II} \\ & + \underbrace{\int \|\nabla_x E \circ G_\theta \nabla_\theta G_\theta\| |p_{E,\theta} \circ G_\theta - p_{E,\theta'} \circ G_{\theta'}| d\eta}_{III}. \end{aligned}$$

The first term can be upper-bounded using L_e-smoothness of E and the fact that G_θ is Lipschitz in θ:

$$\begin{aligned} I &\leq L_e \|\theta - \theta'\| \int |a|^2 (p_{E,\theta} \circ G_\theta) d\eta \\ &\leq L_e C \|\theta - \theta'\|. \end{aligned}$$

The last inequality was obtained by Lemma 39. Similarly, using that $\nabla_\theta G_\theta$ is

Lipschitz, it follows by Lemma 39 that

$$\begin{aligned} II &\leq L_e\|\theta-\theta'\| \int |b|(p_{E,\theta}\circ G_\theta)d\eta \\ &\leq L_e C\|\theta-\theta'\|. \end{aligned}$$

Finally, for the last term III, we first consider a path $\theta_t = t\theta+(1-t)\theta'$ for $t\in[0,1]$, and introduce the function $s(t) := p_{E,\theta_t}\circ G_{\theta_t}$. We will now control the difference $p_{E,\theta}\circ G_\theta - p_{E,\theta'}\circ G_{\theta'}$, also equal to $s(1)-s(0)$. Using the fact that s_t is absolutely continuous we have that $s(1)-s(0) = \int_0^1 s'(t)dt$. The derivative $s'(t)$ is simply given by $s'(t) = (\theta-\theta')^\top(M_t-\bar{M}_t)s(t)$ where $M_t = (\nabla_x E\circ B_{\theta_t})\nabla_\theta G_{\theta_t}$ and $\bar{M}_t = \int M_t p_{E,\theta_t}\circ G_{\theta_t}d\eta$. Hence,

$$s(1)-s(0) = (\theta-\theta')^\top \int_0^1 (M_t-\bar{M}_t)s(t)dt.$$

We also know that M_t is upper-bounded by $La(z)$, which implies

$$\begin{aligned} III &\leq L_e^2\|\theta-\theta'\| \int_0^1 \left(\int |a(z)|^2 s(t)(z)d\eta(z) + \left(\int a(z)s(t)(z)d\eta(z)\right)^2\right) \\ &\leq L_e^2(C+C^2)\|\theta-\theta'\|, \end{aligned}$$

where the last inequality is obtained using Lemma 39. This allows us to conclude that $\mathcal{L}_\theta(E)$ is smooth for any $E\in\mathcal{E}$ and $\theta\in\Theta$. □

Lemma 38. *Under Assumptions **(II)** and **(III)**, it holds that $\psi\mapsto\mathcal{L}_\theta(E_\psi)$ is continuous, and that $\theta\mapsto\mathcal{L}_\theta(E_\psi)$ is differentiable in θ with gradient given by*

$$\partial_\theta\mathcal{L}_\theta(E) := \int (\nabla_x E\circ G_\theta)\,\nabla_\theta G_\theta\,(p_{E,\theta}\circ G_\theta)\,d\eta.$$

Moreover, the gradient is bounded uniformly in θ and E:

$$\|\nabla_\theta\mathcal{L}_\theta(E)\| \leq L_e\left(\int \exp(-L_e L_b\|z\|)\,\mathrm{d}\eta(z)\right)^{-1}\int a(z)\exp(L_e L_b\|z\|)\,\mathrm{d}\eta(z).$$

Proof. To show that $\psi\mapsto\mathcal{L}_\theta(E_\psi)$ is continuous, we will use the dominated conver-

gence theorem. We fix ψ_0 in the interior of Ψ and consider a compact neighborhood W of ψ_0. By assumption, we have that $(\psi, x) \mapsto E_\psi(x)$ and $(\psi, z) \mapsto E_\psi(G_\theta(z))$ are jointly continuous. Hence, $|E_\psi(0)|$ and $|E_\psi(G_\theta(0))|$ are bounded on W by some constant C. Moreover, by Lipschitz continuity of $x \mapsto E_\psi$, we have

$$|E_\psi(x)| \leq |E_\psi(0)| + L_e\|x\| \leq C + L_e\|x\|,$$
$$\exp(-E(G_\theta(z))) \leq \exp(-E(G_\theta(0)))\exp(L_eL_b\|z\|) \leq \exp(C)\exp(L_eL_b\|z\|).$$

Recalling that $\mathbb{P}$ admits a first order moment and that by Assumption **(III)**, $\exp(L_eL_b\|z\|)$ is integrable w.r.t. η, it follows by the dominated convergence theorem and by composition of continuous functions that $\psi \mapsto \mathcal{L}_\theta(E_\psi)$ is continuous in ψ_0.

To show that $\theta \mapsto \mathcal{L}_\theta(E_\psi)$ is differentiable in θ, we will use the differentiation lemma in [Klenke, 2008, Theorem 6.28]. We first fix θ_0 in the interior of Θ, and consider a compact neighborhood V of θ_0. Since $\theta \mapsto |E(G_\theta(0))|$ is continuous on the compact neighborhood V it admits a maximum value C; hence we have using Assumptions **(II)** and **(III)** that

$$\exp(-E(G_\theta(z))) \leq \exp(-E(G_\theta(0)))\exp(L_eL_b\|z\|) \leq \exp(C)\exp(L_eL_b\|z\|).$$

Along with the integrability assumption in Assumption **(III)**, this ensures that $z \mapsto \exp(-E(G_\theta(z)))$ is integrable w.r.t η for all θ in V. We also have that $\exp(-E(G_\theta(z)))$ is differentiable, with gradient given by

$$\nabla_\theta \exp(-E(G_\theta(z))) = \nabla_x E(G_\theta(z))\nabla_\theta G_\theta(z)\exp(-E(G_\theta(z))).$$

Using that E is Lipschitz in its inputs and $G_\theta(z)$ is Lipschitz in θ, and combining with the previous inequality, it follows that

$$\|\nabla_\theta \exp(-E(G_\theta(z)))\| \leq \exp(C)L_e a(z)\exp(L_eL_b\|z\|),$$

where $a(z)$ is the location dependent Lipschitz constant introduced in Assump-

tion (III). The r.h.s. of the above inequality is integrable by Assumption (III) and is independent of θ on the neighborhood V. Thus [Klenke, 2008, Theorem 6.28] applies, and it follows that

$$\nabla_\theta \int \exp(-E(G_{\theta_0}(z)))\,\mathrm{d}\eta(z) = \int \nabla_x E(G_{\theta_0}(z))\nabla_\theta G_{\theta_0}(z)\exp(-E(G_{\theta_0}(z)))\,\mathrm{d}\eta(z).$$

We can now directly compute the gradient of $\mathcal{L}_\theta(E)$,

$$\nabla_\theta \mathcal{L}_\theta(E) = \left(\int \exp(-E(G_{\theta_0}))\,\mathrm{d}\eta\right)^{-1} \int \nabla_x E(G_{\theta_0})\nabla_\theta G_{\theta_0}\exp(-E(G_{\theta_0}))\,\mathrm{d}\eta.$$

Since E and G_θ are Lipschitz in x and θ respectively, it follows that $\|\nabla_x E(G_{\theta_0}(z))\| \leq L_e$ and $\|\nabla_\theta G_{\theta_0}(z)\| \leq a(z)$. Hence, we have

$$\|\nabla_\theta \mathcal{L}_\theta(E)\| \leq L_e \int a(z)(p_{E,\theta} \circ G_\theta(z))d\eta(z).$$

Finally, Lemma 39 allows us to conclude that $\|\nabla_\theta \mathcal{L}_\theta(E)\|$ is bounded by a positive constant C independently from θ and E. □

Lemma 39. *Under Assumptions (II) and (III), there exists a constant C independent from θ and E such that*

$$\begin{aligned} \int a^i(z)(p_{E,\theta} \circ G_\theta(z))d\eta(z) < C, \\ \int b(z)(p_{E,\theta} \circ G_\theta(z))d\eta(z) < C, \end{aligned} \tag{5.17}$$

for $i \in 1, 2$.

Proof. By Lipschitzness of E and G_θ, we have $\exp(-L_e L_b\|z\|) \leq \exp(E(G_\theta(0)) - E(G_\theta(z)) \leq \exp(L_e L_b\|z\|)$, thus introducing the factor $\exp(E(G_\theta(0))$ in (5.17) we get

$$\begin{aligned} \mathbb{E}\left[a^i(Z)(p_{E,\theta} \circ G_\theta(Z))\right] &\leq L_e \left(\mathbb{E}\left[\exp(-L_e L_b\|Z\|)\right]\right)^{-1} \mathbb{E}\left[a(Z)^i \exp(L_e L_b\|Z\|)\right], \\ \mathbb{E}\left[b(Z)(p_{E,\theta} \circ G_\theta(Z))\right] &\leq L_e \left(\mathbb{E}\left[\exp(-L_e L_b\|Z\|)\right]\right)^{-1} \mathbb{E}\left[b(Z)\exp(L_e L_b\|Z\|)\right]. \end{aligned}$$

where the expectation is taken w.r.t. $Z \sim \eta$. The r.h.s. of both inequalities is independent of θ and E, and finite by the integrability assumptions in Assumption (III). □

A .2 Latent space sampling

Here we prove Proposition 35 for which we make the assumptions more precise:

Assumption 2. *We make the following assumption:*

- $\log \eta$ *is strongly concave and admits a Lipschitz gradient.*
- *There exists a non-negative constant* L *such that for any* $x, x' \in \mathcal{X}$ *and* $z, z' \in \mathcal{Z}$*:*

$$|E(x) - E(x')| \leq \|x - x'\|, \qquad \|\nabla_x E(x) - \nabla_x E(x')\| \leq \|x - x'\|$$
$$|G(z) - G(z')| \leq \|z - z'\|, \qquad \|\nabla_z G(z) - \nabla_z G(z')\| \leq \|z - z'\|$$

Throughout this section, we introduce $U(z) := -\log(\eta(z)) + E(G(z))$ for simplicity.

Proof of Proposition 30 . To sample from $\mathbb{Q}_{\mathbb{G},E}$, we first need to identify the *posterior latent* distribution $\nu_{\mathbb{G},E}$ used to produce those samples. We rely on (5.18) which holds by definition of $\mathbb{Q}_{\mathbb{G},E}$ for any test function h on $\mathcal{X}$:

$$\int h(x)\,\mathrm{d}\mathbb{Q}(x) = \int h(G(z)) f(G(z)) \eta(z)\,\mathrm{d}z, \tag{5.18}$$

Hence, the posterior latent distribution is given by $\nu(z) = \eta(z) f(G(z))$, and samples from GEBM are produced by first sampling from $\nu_{\mathbb{G},E}$, then applying the implicit map G,

$$X \sim \mathbb{Q} \quad \Longleftrightarrow \quad X = G(Z), \quad Z \sim \nu.$$

□

Proof of Proposition 31. the base distribution $\mathbb{G}$ admits a density on the whole space denoted by $\exp(-r(x))$ and the energy $\tilde{E}$ is of the form $\tilde{E}(x) = E(x) - r(x)$ for

some parametric function E, it is easy to see that $\mathbb{Q}$ has a density proportional to $\exp(-E)$ and is therefore equivalent to a standard EBM with energy E.

The converse holds as well, meaning that for any EBM with energy E, it is possible to construct a GEBM using an *importance weighting* strategy. This is achieved by first choosing a base $\mathbb{G}$, which is required to have an explicit density $\exp(-r)$ up to a normalizing constant, then defining the energy of the GEBM to be $\tilde{E}(x) = E(x) - r(x)$ so that:

$$\mathrm{d}\mathbb{Q}(x) \propto \exp(-\tilde{E}(x))\,\mathrm{d}\mathbb{G}_\theta(x) \propto \exp(-E(x))\,\mathrm{d}x \tag{5.19}$$

Equation (5.19) effectively depends only on $E(x)$ and not on $\mathbb{G}$ since the factor $\exp(r)$ exactly compensates for the density of $\mathbb{G}$. The requirement that the base also admits a tractable implicit map G can be met by choosing $\mathbb{G}$ to be a *normalizing flow* [Rezende and Mohamed, 2015] and does not restrict the class of possible EBMs that can be expressed as GEBMs. □

Proof of Proposition 35. Let π_t be the probability distribution of (z_t, v_t) at time t of the diffusion in (5.15), which we recall that

$$dz_t = v_t dt, \qquad dv_t \;\; = -\left(\gamma v_t + u\nabla U(z_t)\right) + \sqrt{2\lambda u}dw_t,$$

We call π_∞ its corresponding invariant distribution given by

$$\pi_\infty(z, v) \propto \exp\left(-U(z) - \frac{1}{2}\|v\|^2\right)$$

By Lemma 40 we know that U is dissipative, bounded from below, and has a Lipschitz gradient. This allows to directly apply [Eberle et al., 2017](Corollary 2.6.) which implies that

$$W_2(\pi_t, \pi_\infty) \leq C\exp(-tc),$$

where c is a positive constant and C only depends on π_∞ and the initial distribution

π_0. Moreover, the constant c is given explicitly in [Eberle et al., 2017, Theorem 2.3] and is of order $0(e^{-q})$ where q is the dimension of the latent space $\mathcal{Z}$.

We now consider an optimal coupling Π_t between π_t and π_0. Given joints samples $((z_t, v_t), (z, v))$ from Π_t, we consider the following samples in input space $(x_t, x) := (G(z_t), G(z))$. Since z_t and z have marginals π_t and π_∞, it is easy to see that $x_t \sim \mathbb{P}_t$ and $x \sim \mathbb{Q}$. Therefore, by definition of the W_2 distance, we have the following bound:

$$\begin{aligned} W_2^2(\mathbb{P}_t, \mathbb{Q}) &\leq \mathbb{E}\left[\|x_t - x\|^2\right] \\ &\leq \int \|G(z_t) - G(z)\|^2 d\Pi_t(z_t, z) \\ &\leq L^2 \int \|z_t - z\|^2 d\Pi_t(z_t, z) \\ &\leq L^2 W_2^2(\pi_t, \pi_\infty) \leq C^2 L^2 \exp(-2tc). \end{aligned}$$

The second line uses the definition of (x_t, x) as joint samples obtained by mapping (z_t, z). The third line uses the assumption that B is L-Lipschitz. Finally, the last line uses that Π_t is an optimal coupling between π_t and π_∞. □

Lemma 40. *Under Assumption 2, there exists $A > 0$ and $\lambda \in (0, \frac{1}{4}]$ such that*

$$\frac{1}{2} z^{\top t} \nabla U(z) \geq \lambda \left(U(z) + \frac{\gamma^2}{4u} \|z\|^2 \right) - A, \qquad \forall z \in \mathcal{Z}, \tag{5.20}$$

where γ and u are the coefficients appearing in (5.15). Moreover, U is bounded bellow and has a Lipschitz gradient.

Proof. For simplicity, let's call $u(z) = -\log \eta(z)$, $w(z) = E^\star \circ B_{\theta^\star}(z)$, and denote by M an upper-bound on the Lipschitz constant of w and ∇w which is guaranteed to be finite by assumption. Hence $U(z) = u(z) + w(z)$. Equation (5.20) is equivalent to having

$$z^\top \nabla u(z) - 2\lambda u(z) - \frac{\gamma^2}{2u} \|z\|^2 \geq 2\lambda w(z) - z^\top \nabla w(z) - 2A. \tag{5.21}$$

Using that w is Lipschitz, we have that $w(z) \leq w(0) + M\|z\|$ and $-z^\top \nabla w(z) \leq M\|z\|$. Hence, $2\lambda w(z) - z^\top \nabla w(z) - 2A \leq 2\lambda w(0) + (2\lambda+1)M\|z\| - 2A$. Therefore, a sufficient condition for (5.21) to hold is

$$z^\top \nabla u(z) - 2\lambda u(z) - \frac{\gamma^2}{2u}\|z\|^2 \geq +(2\lambda+1)M\|z\| - 2A + 2\lambda w(0). \tag{5.22}$$

We will now rely on the strong convexity of u, which holds by assumption, and implies the existence of a positive constant $m > 0$ such that

$$\begin{aligned} -u(z) &\geq -u(0) - z^\top \nabla u(z) + \frac{m}{2}\|z\|^2, \\ z^\top \nabla u(z) &\geq -\|z\|\|\nabla u(0)\| + m\|z\|^2. \end{aligned}$$

This allows to write the following inequality,

$$\begin{aligned} z^\top \nabla u(z) - 2\lambda u(z) - \frac{\gamma^2}{2u} \geq & (1-2\lambda) z^\top \nabla u(z) + \lambda(m + \frac{\gamma^2}{2u})\|z\|^2 - 2\lambda u(0) \\ \geq & (1 - \lambda(m + \frac{\gamma^2}{2u}))\|z\|^2 - 2\lambda u(0) \\ & - (1-2\lambda)\|z\|\|\nabla u(0)\|. \end{aligned}$$

Combining the previous inequality with (5.22) and denoting $M' = \|\nabla u(0)\|$, it is sufficient to find A and λ satisfying

$$\xi_1 \|z\|^2 - \xi_2 \|z\| + \xi_3 \geq 0. \tag{5.23}$$

with ξ_1, ξ_2 and ξ_3 being real number defined by:

$$\begin{aligned} \xi_1 &= \left(1 - \lambda\left(m + \frac{\gamma^2}{2u}\right)\right) \\ \xi_2 &= (M + M' + 2\lambda(M - M')) \\ \xi_3 &= 2A - 2\lambda(u(0) + w(0)) \end{aligned}$$

The l.h.s. in (5.23) is a quadratic function in $\|z\|$ and admits a global minimum when

$\lambda < \left(m + \frac{\gamma^2}{2u}\right)^{-1}$. The global minimum is always positive provided that A is large enough.

To see that U is bounded below, it suffice to note, by Lipschitzness of w, that $w(z) \geq w(0) - M\|z\|$ and by strong convexity of u that

$$u(z) \geq u(0) + M'\|z\| + \frac{m}{2}\|z\|^2.$$

Hence, U is lower-bounded by a quadratic function in $\|z\|$ with positive leading coefficient $\frac{m}{2}$, hence it must be lower-bounded by a constant. Finally, by assumption, u and w have Lipschitz gradients, which directly implies that U has a Lipschitz gradient. □

Proof of Proposition 32. By assumption $KL(\mathbb{P}||\mathbb{G}) < +\infty$, this implies that $\mathbb{P}$ admits a density w.r.t. $\mathbb{G}$ which we call $r(x)$. As a result $\mathbb{P}$ admits also a density w.r.t. $\mathbb{Q}$ given by:

$$Z \exp(E^\star(x))r(x).$$

We can then compute the $KL(\mathbb{P}||\mathbb{Q})$ explicitly:

$$\begin{aligned} KL(\mathbb{P}||\mathbb{Q}) &= \mathbb{E}_{\mathbb{P}}[E] + \log(Z) + \mathbb{E}_{\mathbb{P}}[\log(r)] \\ &= -\mathcal{L}_{\mathbb{P},\mathbb{G}}(E^\star) + KL(\mathbb{P}||\mathbb{G}). \end{aligned}$$

Since 0 belongs to $\mathcal{E}$ and by optimality of $E^\star$, we know that $\mathcal{L}_{\mathbb{P},\mathbb{G}}(E^\star) \geq \mathcal{L}_{\mathbb{P},\mathbb{G}}(0) = 0$. The result then follows directly. □

B Experimental details

In all experiments, we use **regularization** which is a combination of L_2 norm and a variant of the gradient penalty Gulrajani et al. [2017]. For the image generation tasks, we also employ spectral normalization Miyato et al. [2018]. This is to ensure that the conditions in Proposition 33 and Theorem 34 hold. We **pre-condition** the gradient as proposed in Simsekli et al. [2020] to stabilize training, and to avoid taking large

noisy gradient steps due to the exponential terms in (5.8) and (5.9). We also use the second-order updates in (5.10) for the variational constant c whenever it is learned.

B .1 Illustrative example in Figure 5.1

The ground truth distribution $\mathbb{P}$ in Figure 5.1(a) follows a simple generative process where each data point $X = (X_1, X_2)$ is obtained as follows:

$$Z \sim Uniform[0, 1]$$
$$X_1 = h_1(Z), \qquad X_2 = h_2(h_1(Z))$$

We choose h_1 and h_2 to be of the form:

$$h_1(z) = \frac{1}{2}\left(z + \frac{1}{1 + \exp(-9(tan(\pi(z - \frac{1}{2})))}\right),$$
$$h_2(x) = sin(8\pi x)/(1 + 4\pi x),$$

Hence, the data is supported on a line defined by the equation $X_2 = h_2(X_1)$ and possesses two modes due to the effect of the distortion introduced by the function h_1.

We provide the details of the models used in Figure 5.1.

GAN For the generator we sample Z uniformly from $[0, 1]$ then generate a sample $(X_1, X_2) = (G_\theta^{(1)}(Z), G_\theta^{(2)}(Z))$:

$$G_\theta^{(1)}(z) = 4\pi W_1 x + b_1, \qquad G_\theta^{(2)}(z) = sin(8\pi W_2 z)/(1 + 4\pi b_2 z).$$

The goal is to learn $\theta = (W_1, b_1, W_2, b_2)$. For the discriminator, we used an MLP with 6 layers and 10 hidden units.

GEBM For the base we use the same generator as in the GAN model. For the energy we use the same MLP as discriminator of the GAN model.

EBM To ensure tractability of the likelihood, we use the following model:

$$X_2|X_1 \sim \mathcal{N}(G_\theta^{(2)}(X_1), \sigma_0)$$
$$X_1 \sim MoG((\mu_1, \sigma_1), (\mu_2, \sigma_2))$$

$z \in \mathbb{R}^{100} \sim \mathcal{N}(0, I)$
dense $\rightarrow M_g \times M_g \times 512$
4×4, stride= 2 deconv. BN 256 ReLU
4×4, stride= 2 deconv. BN 128 ReLU
4×4, stride= 2 deconv. BN 64 ReLU
3×3, stride= 1 conv. 3 Tanh

Table 5.5: Base/Generator of SNGAN ConvNet: $M_g = 4$.

RGB image $x \in \mathbb{R}^{M \times M \times 3}$
3×3, stride= 1 conv 64 lReLU
4×4, stride= 2 conv 64 lReLU
3×3, stride= 1 conv 128 lReLU
4×4, stride= 2 conv 128 lReLU
3×3, stride= 1 conv 256 lReLU
4×4, stride= 2 conv 256 lReLU
3×3, stride= 1 conv 512 lReLU
dense $\rightarrow$ 1.

Table 5.6: Energy / Discriminator of SNGAN ConvNet: $M = 32$.

$MoG((\mu_1, \sigma_1), (\mu_2, \sigma_2))$ refers to a Mixture of two gaussians with mean and variances μ_i and σ_i. We learn each of the parameters $(\theta, \sigma_0, \mu_1, \sigma_1, \mu_2, \sigma_2)$ by maximizing the likelihood.

Both GAN and GEBM have the capacity to recover the the exact support by finding the optimal parameter $\theta^\star$. For the EBM, when $\theta = \theta^\star$, the mean $G_{\theta^\star}(X_1)$ of the conditional gaussian $X_2|X_1$ draws a line which matches the data support exactly, i.e.: $X_2 = G^{(2)}_{\theta^\star}(X_1)$.

B .2 Image generation

Network Architecture Table 5.5 and Table 5.6 show the network architectures used for the GEBM in the case of SNGAN ConvNet. Table 5.5 and Table 5.6 show the network architectures used for the GEBM in the case of SNGAN ResNet. The residual connections of each residual block consists of two convolutional layers proceeded by a BatchNormalization and ReLU activation: **BN+ReLU+Conv+BN+ReLU+Conv** as in [Miyato et al., 2018, Figure 8].

Training: We train both base and energy by alternating 5 gradient steps to learn the energy vs 1 gradient step to learn the base. For the first two gradient iterations and after every 500 gradient iterations on base, we train the energy for 100 gradient steps instead of 5. We then train the model up to 150000 gradient iterations on the base using a batch-size of 128 and Adam optimizer with initial learning rate of 10^{-4} and parameters $(0.5, .999)$ for both energy and base.

RGB image $x \in \mathbb{R}^{M \times M \times 3}$
ResBlock down 128
ResBlock down 128
ResBlock 128
ResBlock 128
ReLu
Global sum pooling
dense $\rightarrow 1$

Table 5.7: Energy / Discriminator of SNGAN ResNet.

$z \in \mathbb{R}^{100} \sim \mathcal{N}(0, I)$
dense, $4 \times 4 \times 256$
ResBlock up 256
ResBlock up 256
ResBlock up 256
BN, ReLu, 3×3 conv, Tanh

Table 5.8: Base/Generator of SNGAN ResNet.

Scheduler: We decrease the learning rate using a scheduler that monitors the FID score in a similar way as in Chapter 4 and Bińkowski* et al. [2018]. More precisely, every 2000 gradient iterations on the base, we evaluate the FID score on the training set using 50000 generated samples from the base and check if the current score is larger than the score 20000 iterations before. The learning rate is decreased by a factor of 0.8 if the FID score fails to decrease for 3 consecutive times.

Sampling: For (DOT) Tanaka [2019], we use the following objective:

$$z \mapsto \|z - z_y + \epsilon\| + \frac{1}{k_{eff}} E \circ G(z) \tag{5.24}$$

where z_y is sampled from a standard Gaussian, ϵ is a perturbation meant to stabilize sampling and k_{eff} is the estimated Lipschitz constant of $E \circ B$. Note that (5.24) uses a flipped sign for the $E \circ B$ compared to Tanaka [2019]. This is because E plays the role of $-D$ where D is the discriminator in Tanaka [2019]. Introducing the minus sign in (5.24) leads to a degradation in performance. We perform 1000 gradient iterations with a step-size of 0.0001 which is also decreased by a factor of 10 every 200 iterations as done for the proposed method. As suggested by the authors of Tanaka [2019] we perform the following projection for the gradient before applying it:

$$g \leftarrow g - \frac{(g^\top z)}{\sqrt{q}} z.$$

We set the perturbation ϵ to 0.001 and k_{eff} to 1 which was also shown in Tanaka [2019] to perform well. In fact, we found that estimating the Lipschitz constant by taking the maximum value of $\|\nabla E \circ G(z)\|$ over 1000 latent samples according to η lead to higher values for k_{eff}: (Cifar10: 9.4, CelebA : 7.2, ImageNet: 4.9, Lsun: 3.8). However, those higher values did not perform as well as setting $k_{eff} = 1$.

For (IHM) Turner et al. [2019] we simply run the MCMC chain for 1000 iterations.

B .3 Density estimation

Pre-processing We use code and pre-processing steps from Wenliang et al. [2019] which we describe here for completeness. For RedWine and WhiteWine, we added uniform noise with support equal to the median distances between two adjacent values. That is to avoid instabilities due to the quantization of the datasets. For Hepmass and MiniBoone, we removed ill-conditioned dimensions as also done in Papamakarios et al. [2017]. We split all datasets, except HepMass into three splits. The test split consists of 10% of the total data. For the validation set, we use 10% of the remaining data with an upper limit of 1000 to reduce the cost of validation at each iteration. For HepMass, we used the sample splitting as done in Papamakarios et al. [2017]. Finally, the data is whitened before fitting and the whitening matrix was computed on at most 10000 data points.

Regularization: We set the regularization parameter to 0.1 and use a combination of L_2 norm and a variant of the gradient penalty Gulrajani et al. [2017]:

$$I(\psi)^2 = \frac{1}{d_\psi}\|\psi\|^2 + \mathbb{E}\left[\|\nabla_x f_\psi(\widetilde{X})\|^2\right]$$

Network Architecture. For both base and energy, we used an NVP Dinh et al. [2016] with 5 NVP layers each consisting of a shifting and scaling layer with two hidden layers of 100 neurons. We do not use Batch-normalization.

Training: In all cases we use Adam optimizer with learning rate of 0.001 and momentum parameters $(0.5, 0.9)$. For both KALE-DV and KALE-F, we used a batch-size of 100 data samples vs 2000 generated samples from the base in order to

reduce the variance of the estimation of the energy. We alternate 50 gradient steps on the energy vs 1 step on the base and further perform 50 additional steps on the energy for the first two gradient iterations and after every 500 gradient iterations on base. For Contrastive divergence, each training step is performed by first producing 100 samples from the model using 100 Langevin iterations with a step-size of 10^{-2} and starting from a batch of 100 data-samples. The resulting samples are then used to estimate the gradient of the of the loss.

For (CD), we used 100 Langevin iterations for each learning step to sample from the EBM. This translates into an improved performance at the expense of increased computational cost compared to the other methods. All methods are trained for 2000 epochs with batch-size of 100 (1000 on Hepmass and Miniboone datasets) and fixed learning rate 0.001, which was sufficient for convergence.

C The KL Approximate Lower-bound Estimate

We discuss the relation between KALE (5.11) and the Kullback-Leibler divergence via Fenchel duality. Recall that a distribution $\mathbb{P}$ is said to admit a density w.r.t. $\mathbb{G}$ if there exists a real-valued measurable function r_0 that is integrable w.r.t. $\mathbb{G}$ and satisfies $d\mathbb{P} = r_0 d\mathbb{G}$. Such a density is also called the *Radon-Nikodym derivative* of $\mathbb{P}$ w.r.t. $\mathbb{G}$. In this case, we have:

$$\mathrm{KL}(\mathbb{P}||\mathbb{G}) = \int r_0 \log(r_0) d\mathbb{G}. \tag{5.25}$$

Nguyen et al. [2010], Nowozin et al. [2016] derived a variational formulation for the KL using Fenchel duality. By the duality theorem [Rockafellar, 1970], the convex and lower semi-continuous function $\zeta : u \mapsto u\log(u)$ that appears in (5.25) can be expressed as the supremum of a concave function:

$$\zeta(u) = \sup_v uv - \zeta^\star(v).$$

The function $\zeta^\star$ is called the *Fenchel dual* and is defined as $\zeta^\star(v) = \sup_u uv - \zeta(u)$. By convention, the value of the objective is set to $-\infty$ whenever u is outside of the

domain of definition of $\zeta^\star$. When $\zeta(u) = u\log(u)$, the Fenchel dual $\zeta^\star(v)$ admits a closed form expression of the form $\zeta^\star(v) = \exp(v-1)$. Using the expression of ζ in terms of its Fenchel dual $\zeta^\star$, it is possible to express $\mathrm{KL}(\mathbb{P}||\mathbb{G})$ as the supremum of the variational objective (5.26) over all measurable functions h.

$$\mathcal{F}(h) := -\int h d\mathbb{P} - \int \exp(-h) d\mathbb{G} + 1. \tag{5.26}$$

Nguyen et al. [2010] provided the variational formulation for the reverse KL using a different choice for ζ: $(\zeta(u) = -\log(u))$. We refer to [Nowozin et al., 2016] for general f-divergences. Choosing a smaller set of functions $\mathcal{H}$ in the variational objective (5.26) will lead to a lower bound on the KL. This is the *KL Approximate Lower-bound Estimate* (KALE):

$$\mathrm{KALE}(\mathbb{P}||\mathbb{G}) = \sup_{h\in\mathcal{H}} \mathcal{F}(h) \tag{5.27}$$

In general, $\mathrm{KL}(\mathbb{P}||\mathbb{G}) \geq \mathrm{KALE}(\mathbb{P}||\mathbb{G})$. The bound is tight whenever the negative log-density $h_0 = -\log r_0$ belongs to $\mathcal{H}$; however, we do not require r_0 to be well-defined in general. Equation (5.27) has the advantage that it can be estimated using samples from $\mathbb{P}$ and $\mathbb{G}$. Given i.i.d. samples $(X_1, ..., X_N)$ and $(Y_1, ..., Y_M)$ from $\mathbb{P}$ and $\mathbb{G}$, we denote by $\hat{\mathbb{P}}$ and $\hat{\mathbb{G}}$ the corresponding empirical distributions. A simple approach to estimate $\mathrm{KALE}(\mathbb{P}||\mathbb{G})$ is to use an M-estimator. This is achieved by optimizing the penalized objective

$$\hat{h} := \arg\max_{h\in\mathcal{H}} \widehat{\mathcal{F}}(h) - \frac{\lambda}{2} I^2(h), \tag{5.28}$$

where $\widehat{\mathcal{F}}$ is an empirical version of $\mathcal{F}$ and $I^2(h)$ is a penalty term that prevents overfitting due to finite samples. The penalty $I^2(h)$ acts as a regularizer favoring smoother solutions while the parameter λ determines the strength of the smoothing and is chosen to decrease as the sample size N and M increase. The M-estimator of $\mathrm{KALE}(\mathbb{P}||\mathbb{G})$ is obtained simply by plugging in $\hat{h}$ into the empirical objective

$\widehat{\mathcal{F}}(h)$:

$$\widehat{\mathrm{KALE}}(\mathbb{P}||\mathbb{G}) := \widehat{\mathcal{F}}(\hat{h}). \tag{5.29}$$

We defer the consistency analysis of (5.29) to Section C .1 where we provide convergence rates in a setting where the set of functions $\mathcal{H}$ is a Reproducing Kernel Hilbert Space and under weaker assumptions that were not covered by the framework of Nguyen et al. [2010].

C .1 Convergence rate of KALE

In this section, we provide a convergence rate for the estimator in (5.29) when $\mathcal{H}$ is an RKHS. The theory remains the same whether $\mathcal{H}$ contains constants or not. With this choice, the Representer Theorem allows us to reduce the potentially infinite-dimensional optimization problem in (5.28) to a convex finite-dimensional one. We further restrict ourselves to the *well-specified* case where the density r_0 of $\mathbb{P}$ w.r.t. $\mathbb{G}$ is well-defined and belongs to $\mathcal{H}$, so that KALE matches the KL. While Nguyen et al. [2010] (Theorem 3) provides a convergence rate of $1/\sqrt{N}$ for a related M-estimator, this requires the density r_0 to be lower-bounded by 0 as well as (generally) upper-bounded. This can be quite restrictive if, for instance, r_0 is the density ratio of two gaussians. In Theorem 41, we provide a similar convergence rate for the estimator defined in (5.29) without requiring r_0 to be bounded. We start by briefly introducing some notations, the working assumptions and the statement of the convergence result and defer the proofs to Section C .2.

We recall that an RKHS $\mathcal{H}$ of functions defined on a domain $\mathcal{X} \subset \mathbb{R}^d$ and with kernel k is a Hilbert space with dot product $\langle ., . \rangle$, such that $y \mapsto k(x, y)$ belongs to $\mathcal{H}$ for any $x \in \mathcal{X}$, and

$$k(x, y) = \langle k(x, .), k(y, .) \rangle, \qquad \forall x, y \in \mathcal{X}.$$

Any function h in $\mathcal{H}$ satisfies the reproducing property $f(x) = \langle f, k(x, .) \rangle$ for any $x \in \mathcal{X}$.

Recall that KALE($\mathbb{P}||\mathbb{G}$) is obtained as an optimization problem

$$\text{KALE}(\mathbb{P}||\mathbb{G}) = \sup_{h\in\mathcal{H}} \mathcal{F}(h) \tag{5.30}$$

where $\mathcal{F}$ is given by:

$$\mathcal{F}(h) := -\int h d\mathbb{P} - \int \exp(-h) d\mathbb{G} + 1.$$

Since the negative log density ratio h_0 is assumed to belong to $\mathcal{H}$, this directly implies that the supremum of $\mathcal{F}$ is achieved at h_0 and $\mathcal{F}(h_0) = \text{KALE}(\mathbb{P}||\mathbb{G})$. We are interested in estimating KALE($\mathbb{P}||\mathbb{G}$) using the empirical distributions $\hat{\mathbb{P}}$ and $\hat{\mathbb{G}}$,

$$\hat{\mathbb{P}} := \frac{1}{N}\sum_{n=1}^{N} \delta_{X_n}, \qquad \hat{\mathbb{G}} := \frac{1}{N}\sum_{n=1}^{N} \delta_{Y_n},$$

where $(X_n)_{1\leq n\leq N}$ and $(Y_n)_{1\leq n\leq N}$ are i.i.d. samples from $\mathbb{P}$ and $\mathbb{G}$. For this purpose we introduce the empirical objective functional,

$$\widehat{\mathcal{F}}(h) := -\int h d\hat{\mathbb{P}} - \int \exp(-h) d\hat{\mathbb{G}} + 1.$$

The proposed estimator is obtained by solving a regularized empirical problem,

$$\sup_{h\in\mathcal{H}} \widehat{\mathcal{F}}(h) - \frac{\lambda}{2}\|h\|^2, \tag{5.31}$$

with a corresponding population version,

$$\sup_{h\in\mathcal{H}} \mathcal{F}(h) - \frac{\lambda}{2}\|h\|^2. \tag{5.32}$$

Finally, we introduce $D(h,\delta)$ and $\Gamma(h,\delta)$:

$$D(h,\delta) = \int \delta \exp(-h) d\mathbb{G} - \int \delta d\mathbb{P},$$
$$\Gamma(h,\delta) = -\int\int_0^1 (1-t)\delta^2 \exp(-(h+t\delta)) d\mathbb{G}.$$

The empirical versions of $D(h,\delta)$ and $\Gamma(h,\delta)$ are denoted $\hat{D}(h,\delta)$ and $\hat{\Gamma}(h,\delta)$. Later, we will show that $D(h,\delta)$ $\hat{D}(h,\delta)$ are in fact the gradients of $\mathcal{F}(h)$ and $\widehat{\mathcal{F}}(h)$ along the direction δ.

We state now the working assumptions:

(i) The supremum of $\mathcal{F}$ over $\mathcal{H}$ is attained at h_0.

(ii) The following quantities are finite for some positive ϵ:

$$\int \sqrt{k(x,x)}\, d\mathbb{P}(x),$$
$$\int \sqrt{k(x,x)} \exp((\|h_0\| + \epsilon)\sqrt{k(x,x)})\, d\mathbb{G}(x),$$
$$\int k(x,x) \exp((\|h_0\| + \epsilon)\sqrt{k(x,x)})\, d\mathbb{G}(x).$$

(iii) For any $h \in \mathcal{H}$, if $D(h,\delta) = 0$ for all δ then $h = h_0$.

Theorem 41. *Fix any $1 > \eta > 0$. Under Assumptions (i) to (iii), and provided that $\lambda = \frac{1}{\sqrt{N}}$, it holds with probability at least $1 - 2\eta$ that*

$$|\widehat{\mathcal{F}}(\hat{h}) - \mathcal{F}(h_0)| \leq \frac{M'(\eta, h_0)}{\sqrt{N}}$$

for a constant $M'(\eta, h_0)$ that depends only on η and h_0.

The assumptions in Theorem 41 essentially state that the kernel associated to the RKHS $\mathcal{H}$ needs to satisfy some integrability requirements. That is to guarantee that the gradient $\delta \mapsto \nabla\mathcal{F}(h)(\delta)$ and its empirical version are well-defined and continuous. In addition, the optimality condition $\nabla\mathcal{F}(h) = 0$ is assumed to characterize the global solution h_0. This will be the case if the kernel is characteristic Simon-Gabriel and Scholkopf [2018]. The proof of Theorem 41, in Section C .2, takes advantage of the Hilbert structure of the set $\mathcal{H}$, the convexity of the functional $\mathcal{F}$ and the optimality condition $\nabla\widehat{\mathcal{F}}(\hat{h}) = \lambda\hat{h}$ of the regularized problem, all of which turn out to be sufficient for controlling the error of (5.29).

C .2 Proofs

We state now the proof of Theorem 41 with subsequent lemmas and propositions.

Proof of Theorem 41. We begin with the following inequalities:

$$\frac{\lambda}{2}(\|\hat{h}\|^2 - \|h_0\|^2) \leq \widehat{\mathcal{F}}(\hat{h}) - \widehat{\mathcal{F}}(h_0) \leq \langle \nabla\widehat{\mathcal{F}}(h_0), \hat{h} - h_0\rangle.$$

The first inequality is by definition of $\hat{h}$ while the second is obtained by concavity of $\widehat{\mathcal{F}}$. For simplicity we write $\mathcal{B} = \|\hat{h} - h_0\|$ and $\mathcal{C} = \|\nabla\widehat{\mathcal{F}}(h_0) - \mathcal{L}(h_0)\|$. Using Cauchy-Schwarz and triangular inequalities, it is easy to see that

$$-\frac{\lambda}{2}\left(\mathcal{B}^2 + 2\mathcal{B}\|h_0\|\right) \leq \widehat{\mathcal{F}}(\hat{h}) - \widehat{\mathcal{F}}(h_0) \leq \mathcal{C}\mathcal{B}.$$

Moreover, by triangular inequality, it holds that

$$\mathcal{B} \leq \|h_\lambda - h_0\| + \|\hat{h} - h_\lambda\|.$$

Lemma 45 ensures that $\mathcal{A}(\lambda) = \|h_\lambda - h_0\|$ converges to 0 as $\lambda \to 0$. Furthermore, by Proposition 46, we have $\|\hat{h} - h_\lambda\| \leq \frac{1}{\lambda}\mathcal{D}$ where $\mathcal{D}(\lambda) = \|\nabla\widehat{\mathcal{F}}(h_\lambda) - \nabla\mathcal{L}(h_\lambda)\|$. Now choosing $\lambda = \frac{1}{\sqrt{N}}$ and applying Chebychev inequality in Lemma 42, it follows that for any $1 > \eta > 0$, we have with probability greater than $1 - 2\eta$ that both

$$\mathcal{D}(\lambda) \leq \frac{C(\|h_0\|\eta)}{\sqrt{N}}, \qquad \mathcal{C} \leq \frac{C(\|h_0\|, \eta)}{\sqrt{N}},$$

where $C(\|h_0\|, \eta)$ is defined in Lemma 42. This allows to conclude that for any $\eta > 0$, it holds with probability at least $1 - 2\eta$ that $|\widehat{\mathcal{F}}(\hat{h}) - \widehat{\mathcal{F}}(h_0)| \leq \frac{M'(\eta, h_0)}{\sqrt{N}}$ where $M'(\eta, h_0)$ depends only on η and h_0. □

We proceed using the following lemma, which provides an expression for $D(h, \delta)$ and $\hat{D}(h, \delta)$ along with a probabilistic bound:

Lemma 42. *Under Assumptions **(i)** and **(ii)**, for any $h \in \mathcal{H}$ such that $\|h\| \leq \|h_0\| + \epsilon$,*

there exists $\mathcal{D}(h)$ *in* $\mathcal{H}$ *satisfying*

$$D(h, \delta) = \langle \delta, \mathcal{D}(h) \rangle,$$

and for any $h \in \mathcal{H}$*, there exists* $\widehat{\mathcal{D}}(h)$ *satisfying*

$$\widehat{D}(h, \delta) = \langle \delta, \widehat{\mathcal{D}}(h) \rangle.$$

Moreover, for any $0 < \eta < 1$ *and any* $h \in \mathcal{H}$ *such that* $\|h\| \leq \|h_0\| + \epsilon := M$*, it holds with probability greater than* $1 - \eta$ *that*

$$\|\mathcal{D}(h) - \widehat{\mathcal{D}}(h)\| \leq \frac{C(M, \eta)}{\sqrt{N}},$$

where $C(M, \eta)$ *depends only on* M *and* η*.*

Proof. First, we show that $\delta \mapsto D(h, \delta)$ is a bounded linear operator. Indeed, Assumption **(ii)** ensures that $k(x, .)$ and $k(x, .) \exp(-h(x))$ are Bochner integrable w.r.t. $\mathbb{P}$ and $\mathbb{G}$ (Retherford [1978]), hence $D(h, \delta)$ is obtained as

$$D(h, \delta) := \langle \delta, \mu_{\exp(-h)\mathbb{G}} - \mu_{\mathbb{P}} \rangle,$$

where $\mu_{\exp(-h)\mathbb{G}} = \int k(x, .) \exp(-h(x)) d\mathbb{G}$ and $\mu_{\mathbb{P}} = \int k(x, .) d\mathbb{P}$. Defining $\mathcal{D}(h)$ to be $= \mu_{\exp(-h)\mathbb{G}} - \mu_{\mathbb{P}}$ leads to the desired result. $\widehat{\mathcal{D}}(h)$ is simply obtained by taking the empirical version of $\mathcal{D}(h)$.

Finally, the probabilistic inequality is a simple consequence of Chebychev's inequality. □

The next lemma states that $\mathcal{F}(h)$ and $\widehat{\mathcal{F}}(h)$ are Frechet differentiable.

Lemma 43. *Under Assumptions* ***(i)*** *and* ***(ii)*** *,* $h \mapsto \mathcal{F}(h)$ *is Frechet differentiable on the open ball of radius* $\|h_0\| + \epsilon$ *while* $h \mapsto \widehat{\mathcal{F}}(h)$ *is Frechet differentiable on* $\mathcal{H}$*. Their gradients are given by* $\mathcal{D}(h)$ *and* $\widehat{\mathcal{D}}(h)$ *as defined in Lemma 42,*

$$\nabla \mathcal{F}(h) = \mathcal{D}(h), \qquad \nabla \widehat{\mathcal{F}}(h) = \widehat{\mathcal{D}}(h)$$

Proof. The empirical functional $\widehat{\mathcal{F}}(h)$ is differentiable since it is a finite sum of differentiable functions, and its gradient is simply given by $\widehat{\mathcal{D}}(h)$. For the population functional, we use second order Taylor expansion of $\exp$ with integral remainder, which gives

$$\mathcal{F}(h+\delta) = \mathcal{F}(h) - D(h,\delta) + \Gamma(h,\delta).$$

By Assumption **(ii)** we know that $\frac{\Gamma(h,\delta)}{\|\delta\|}$ converges to 0 as soon as $\|\delta\| \to 0$. This allows to directly conclude that $\mathcal{F}$ is Frechet differentiable, with differential given by $\delta \mapsto D(h,\delta)$. By Lemma 42, we conclude the existence of a gradient $\nabla\mathcal{F}(h)$ which is in fact given by $\nabla\mathcal{F}(h) = \mathcal{D}(h)$.

□

From now on, we will only use the notation $\nabla\mathcal{F}(h)$ and $\nabla\widehat{\mathcal{F}}(h)$ to refer to the gradients of $\mathcal{F}(h)$ and $\widehat{\mathcal{F}}(h)$. The following lemma states that (5.31) and (5.32) have a unique global optimum, and gives a first order optimality condition.

Lemma 44. *The problems (5.31) and (5.32) admit unique global solutions $\hat{h}$ and h_λ in $\mathcal{H}$. Moreover, the following first order optimality conditions hold:*

$$\lambda\hat{h} = \nabla\widehat{\mathcal{F}}(\hat{h}), \qquad \lambda h_\lambda = \nabla\mathcal{F}(h_\lambda).$$

Proof. For (5.31), existence and uniqueness of a minimizer $\hat{h}$ is a simple consequence of continuity and strong concavity of the regularized objective. We now show the existence result for (5.32). Let's introduce $\mathcal{G}_\lambda(h) = -\mathcal{F}(h) + \frac{\lambda}{2}\|h\|^2$ for simplicity. Uniqueness is a consequence of the strong convexity of $\mathcal{G}_\lambda$. For the existence, consider a sequence of elements $f_k \in \mathcal{H}$ such that $\mathcal{G}_\lambda(f_k) \to \inf_{h\in\mathcal{H}} \mathcal{G}_\lambda(h)$. If h_0 is not the global solution, then it must hold for k large enough that $\mathcal{G}_\lambda(f_k) \leq \mathcal{G}_\lambda(h_0)$. We also know that $\mathcal{F}(f_k) \leq \mathcal{F}(h_0)$, hence, it is easy to see that $\|f_k\| \leq \|h_0\|$ for k large enough. This implies that f_k is a bounded sequence, therefore it admits a weakly convergent sub-sequence by weak compactness. Without loss of generality we assume that f_k weakly converges to some element $h_\lambda \in \mathcal{H}$ and that $\|f_k\| \leq \|h_0\|$. Hence, $\|h_\lambda\| \leq \liminf_k \|f_k\| \leq \|h_0\|$. Recall now that by definition of weak

convergence, we have $f_k(x) \to_k h_\lambda(x)$ for all $x \in \mathcal{X}$. By Assumption (ii), we can apply the dominated convergence theorem to ensure that $\mathcal{F}(f_k) \to \mathcal{F}(h_\lambda)$. Taking the limit of $\mathcal{G}_\lambda f_k$, the following inequality holds:

$$\sup_{h \in \mathcal{H}} \mathcal{G}_\lambda(h) = \limsup_k \mathcal{G}_\lambda(f_k) \leq \mathcal{G}_\lambda(h_\lambda).$$

Finally, by Lemma 43 we know that $\mathcal{F}$ is Frechet differentiable, hence we can use Ekeland and Témam [1999] (Proposition 2.1) to conclude that $\nabla\mathcal{F}(h_\lambda) = \lambda h_\lambda$. We use exactly the same arguments for (5.31). □

Next, we show that h_λ converges towards h_0 in $\mathcal{H}$.

Lemma 45. *Under Assumptions (i) to (iii) it holds that:*

$$\mathcal{A}(\lambda) := \|h_\lambda - h_0\| \to 0.$$

Proof. We will first prove that h_λ converges weakly towards h_0, and then conclude that it must also converge strongly. We start with the following inequalities:

$$0 \geq \mathcal{F}(h_\lambda) - \mathcal{F}(h_0) \geq \frac{\lambda}{2}(\|h_\lambda\|^2 - \|h_0\|^2).$$

These are simple consequences of the definitions of h_λ and h_0 as optimal solutions to (5.30) and (5.31). This implies that $\|h_\lambda\|$ is always bounded by $\|h_0\|$. Consider now an arbitrary sequence $(\lambda_m)_{m \geq 0}$ converging to 0. Since $\|h_{\lambda_m}\|$ is bounded by $\|h_0\|$, it follows by weak-compactness of balls in $\mathcal{H}$ that h_{λ_m} admits a weakly convergent sub-sequence. Without loss of generality we can assume that h_{λ_m} is itself weakly converging towards an element h^*. We will show now that h^* must be equal to h_0. Indeed, by optimality of h_{λ_m}, it must hold that

$$\lambda_m h_{\lambda_m} = \nabla\mathcal{F}(h_m).$$

This implies that $\nabla\mathcal{F}(h_m)$ converges weakly to 0. On the other hand, by Assumption (ii), we can conclude that $\nabla\mathcal{F}(h_m)$ must also converge weakly towards $\nabla\mathcal{F}(h^*)$,

hence $\nabla\mathcal{F}(h^*) = 0$. Finally by Assumption **(iii)** we know that h_0 is the unique solution to the equation $\nabla\mathcal{F}(h) = 0$, hence $h^* = h_0$. We have shown so far that any subsequence of h_{λ_m} that converges weakly, must converge weakly towards h_0. This allows to conclude that h_{λ_m} actually converges weakly towards h_0. Moreover, we also have by definition of weak convergence that:

$$\|h_0\| \leq \lim\inf_{m\to\infty} \|h_{\lambda_m}\|.$$

Recalling now that $\|h_{\lambda_m}\| \leq \|h_0\|$ it follows that $\|h_{\lambda_m}\|$ converges towards $\|h_0\|$. Hence, we have the following two properties:

- h_{λ_m} converges weakly towards h_0,
- $\|h_{\lambda_m}\|$ converges towards $\|h_0\|$.

This allows to directly conclude that $\|h_{\lambda_m} - h_0\|$ converges to 0. □

Proposition 46. *We have that:*

$$\|\hat{h} - h_\lambda\| \leq \frac{1}{\lambda}\|\nabla\hat{\mathcal{F}}(h_\lambda) - \nabla\mathcal{F}(h_\lambda)\|$$

Proof. By definition of $\hat{h}$ and h_λ the following optimality conditions hold:

$$\lambda\hat{h} = \nabla\widehat{\mathcal{F}}(\hat{h}), \qquad \lambda h_\lambda = \nabla\mathcal{F}(h_\lambda).$$

We can then simply write:

$$\lambda(\hat{h} - h_\lambda) - (\nabla\widehat{\mathcal{F}}(\hat{h}) - \nabla\widehat{\mathcal{F}}(h_\lambda)) = \nabla\widehat{\mathcal{F}}(h_\lambda) - \nabla\mathcal{F}(h_\lambda).$$

Now introducing $\delta := \hat{h} - h_\lambda$ and $E := \nabla\widehat{\mathcal{F}}(\hat{h}) - \nabla\widehat{\mathcal{F}}(h_\lambda)$ for simplicity and taking the squared norm of the above equation, it follows that

$$\lambda^2\|\delta\|^2 + \|E\|^2 - 2\lambda\langle\delta, E\rangle = \|\nabla\widehat{\mathcal{F}}(h_\lambda) - \nabla\mathcal{F}(h_\lambda)\|^2.$$

By concavity of $\widehat{\mathcal{F}}$ on $\mathcal{H}$ we know that $-\langle \hat{h} - h_\lambda, E \rangle \geq 0$. Therefore:

$$\lambda^2 \|\hat{h} - h_\lambda\|^2 \leq \|\nabla \widehat{\mathcal{F}}(h_\lambda) - \nabla \mathcal{F}(h_\lambda)\|^2.$$

□

Part III

Optimizing implicit generative models

Chapter 6

Wasserstein gradient flow of the Maximum Mean Discrepancy

We construct a Wasserstein gradient flow of the maximum mean discrepancy (MMD) and study its convergence properties. The MMD is an integral probability metric defined for a reproducing kernel Hilbert space (RKHS), and serves as a metric on probability measures for a sufficiently rich RKHS. We obtain conditions for convergence of the gradient flow towards a global optimum, that can be related to particle transport when optimizing neural networks. We also propose a way to regularize this MMD flow, based on an injection of noise in the gradient. This algorithmic fix comes with theoretical and empirical evidence. The practical implementation of the flow is straightforward, since both the MMD and its gradient have simple closed-form expressions, which can be easily estimated with samples.

1 Introduction

We address the problem of defining a gradient flow on the space of probability distributions endowed with the Wasserstein metric, which transports probability mass from a starting distribtion ν to a target distribution μ. Our flow is defined on the maximum mean discrepancy (MMD) Gretton et al. [2012], an integral probability metric Müller [1997] which uses the unit ball in a characteristic RKHS Sriperumbudur et al. [2010] as its witness function class. Specifically, we choose the function in the witness class that has the largest difference in expectation under ν and μ: this difference constitutes

the MMD. The idea of descending a gradient flow over the space of distributions can be traced back to the seminal work of Jordan et al. [1998], who revealed that the Fokker-Planck equation is a gradient flow of the Kullback-Leibler divergence. Its time-discretization leads to the celebrated Langevin Monte Carlo algorithm, which comes with strong convergence guarantees (see Durmus et al. [2018], Dalalyan and Karagulyan [2019]), but requires the knowledge of an analytical form of the target μ. A more recent gradient flow approach, Stein Variational Gradient Descent (SVGD) Liu [2017], also leverages this analytical μ.

The study of particle flows defined on the MMD relates to two important topics in modern machine learning. The first is in training Implicit Generative Models, notably generative adversarial networks Goodfellow et al. [2014]. Integral probability metrics have been used extensively as critic functions in this setting: these include the Wasserstein distance Arjovsky and Bottou [2017], Gulrajani et al. [2017], Genevay et al. [2018] and the maximum mean discrepancy used in Chapter 4 and in Dziugaite et al. [2015], Li et al. [2015, 2017], Bellemare et al. [2017], Bińkowski* et al. [2018]. In [Mroueh et al., 2019, Section 3.3], a connection between IGMs and particle transport is proposed, where it is shown that gradient flow on the witness function of an integral probability metric takes a similar form to the generator update in a GAN. The critic IPM in this case is the Kernel Sobolev Discrepancy (KSD), which has an additional gradient norm constraint on the witness function compared with the MMD. It is intended as an approximation to the negative Sobolev distance from the optimal transport literature Otto and Villani [2000], Villani [2009], Peyre [2018]. There remain certain differences between gradient flow and GAN training, however. First, and most obviously, gradient flow can be approximated by representing ν as a set of particles, whereas in a GAN ν is the output of a generator network. The requirement that this generator network be a smooth function of its parameters causes a departure from pure particle flow. Second, in modern implementations as in Chapter 4 and Li et al. [2017], Bińkowski* et al. [2018], the kernel used in computing the critic witness function for an MMD GAN critic is parametrized by a deep network, and an alternating optimization between the critic parameters and the generator parameters

is performed. Despite these differences, we anticipate that the theoretical study of MMD flow convergence will provide helpful insights into conditions for GAN convergence, and ultimately, improvements to GAN training algorithms.

Regarding the second topic, we note that the properties of gradient descent for large neural networks have been modeled using the convergence towards a global optimum of particle transport in the population limit, when the number of particles goes to infinity Rotskoff and Vanden-Eijnden [2018], Chizat and Bach [2018a], Mei et al. [2018], Sirignano and Spiliopoulos [2018]. In particular, Rotskoff et al. [2019] show that gradient descent on the parameters of a neural network can also be seen as a particle transport problem, which has as its population limit a gradient flow of a functional defined for probability distributions over the parameters of the network. This functional is in general non-convex, which makes the convergence analysis challenging. The particular structure of the MMD allows us to relate its gradient flow to neural network optimization in a well-specified regression setting similar to Rotskoff et al. [2019], Chizat and Bach [2018a] (we make this connection explicit in Section C).

Our main contribution in this work is to establish conditions for convergence of MMD gradient flow to its *global optimum*. We give detailed descriptions of MMD flow for both its continuous-time and discrete instantiations in Section 2 . In particular, the MMD flow may employ a sample approximation for the target μ: unlike e.g. Langevin Monte Carlo or SVGD, it does not require μ in analytical form. Global convergence is especially challenging to prove: while for functionals that are *displacement convex*, the gradient flow can be shown to converge towards a global optimum Ambrosio et al. [2008], the case of non-convex functionals, like the MMD, requires different tools. A modified gradient flow is proposed in Rotskoff et al. [2019] that uses particle birth and death to reach global optimality. Global optimality may also be achieved simply by teleporting particles from ν to μ, as occurs for the Sobolev Discrepancy flow absent a kernel regulariser [Mroueh et al., 2019, Theorem 4, Appendix D]. Note, however, that the regularised Kernel Sobolev Discrepancy flow does not rely on teleportation.

Our approach takes inspiration in particular from Bottou et al. [2018], where it is shown that although the 1-Wasserstein distance is non-convex, it can be optimized up to some barrier that depends on the diameter of the domain of the target distribution. Similarly to Bottou et al. [2018], we provide in Section 3 a barrier on the gradient flow of the MMD, although the tightness of this barrier in terms of the target diameter remains to be established. We obtain a further condition on the evolution of the flow to ensure global optimality, and give rates of convergence in that case, however the condition is a strong one: it implies that the negative Sobolev distance between the target and the current particles remains bounded at all times.

We thus propose a way to regularize the MMD flow, based on a noise injection (Section 4) in the gradient, with more tractable theoretical conditions for convergence. Encouragingly, the noise injection is shown in practice to ensure convergence in a simple illustrative case where the original MMD flow fails. Finally, while our emphasis has been on establishing conditions for convergence, we note that MMD gradient flow has a simple $O(MN + N^2)$ implementation for N ν-samples and M μ-samples, and requires only evaluating the gradient of the kernel k on the given samples.

2 Gradient flow of the MMD in W_2

2 .1 Construction of the gradient flow

In this section we introduce the gradient flow of the Maximum Mean Discrepancy (MMD) and highlight some of its properties. We start by briefly reviewing the MMD introduced in Gretton et al. [2012]. We define $\mathcal{X} \subset \mathbb{R}^d$ as the closure of a convex open set, and $\mathcal{P}_2(\mathcal{X})$ as the set of probability distributions on $\mathcal{X}$ with finite second moment, equipped with the 2-Wassertein metric denoted W_2. For any $\nu \in \mathcal{P}_2(\mathcal{X})$, $L_2(\nu)$ is the set of square integrable functions w.r.t. ν.

Maximum Mean Discrepancy. Given a characteristic kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we denote by $\mathcal{H}$ its corresponding RKHS (see Smola and Scholkopf [1998]). The space $\mathcal{H}$ is a Hilbert space with inner product $\langle ., . \rangle_{\mathcal{H}}$ and norm $\|.\|_{\mathcal{H}}$. We will rely on specific assumptions on the kernel which are given below:

(A) k is continuously differentiable on $\mathcal{X}$ with L-Lipschitz gradient: $\|\nabla k(x,x') - \nabla k(y,y')\| \leq L(\|x-y\| + \|x'-y'\|)$ for all $x, x', y, y' \in \mathcal{X}$.

(B) k is twice differentiable on $\mathcal{X}$.

(C) $\|Dk(x,y)\| \leq \lambda$ for all $x, y \in \mathcal{X}$, where $Dk(x,y)$ is an $\mathbb{R}^{d^2} \times \mathbb{R}^{d^2}$ matrix with entries given by $\partial_{x_i}\partial_{x_j}\partial_{x'_i}\partial_{x'_j}k(x,y)$.

(D) $\sum_{i=1}^{d} \|\partial_i k(x,.) - \partial_i k(y,.)\|_{\mathcal{H}}^2 \leq \lambda^2\|x-y\|^2$ for all $x, y \in \mathcal{X}$.

In particular, Assumption **(A)** states that the gradient of the kernel, ∇k, is Lipschitz with constant L. For such kernels, it is possible to define the Maximum Mean Discrepancy as a distance on $\mathcal{P}_2(\mathcal{X})$. The MMD can be written as the RKHS norm of the unnormalised *witness function* $f_{\mu,\nu}$ between μ and ν, which is the difference between the mean embeddings of ν and μ,

$$\begin{aligned} MMD(\mu,\nu) &= \|f_{\mu,\nu}\|_{\mathcal{H}} \\ f_{\nu,\mu}(z) &= \int k(x,z)\,\mathrm{d}\nu(x) - \int k(x,z)\,\mathrm{d}\mu(x) \quad \forall z \in \mathcal{X} \end{aligned} \tag{6.1}$$

Throughout this chapter, μ will be fixed and ν can vary, hence we will only consider the dependence in ν and denote by $\mathcal{F}(\nu) = \frac{1}{2}MMD^2(\mu,\nu)$. A direct computation [Mroueh et al., 2019, Appendix B] shows that for any finite measure χ such that $\nu + \epsilon\chi \in \mathcal{P}_2(\mathcal{X})$, we have

$$\lim_{\epsilon\to 0} \epsilon^{-1}(\mathcal{F}(\nu+\epsilon\chi) - \mathcal{F}(\nu)) = \int f_{\mu,\nu}(x) d\chi(x). \tag{6.2}$$

This means that $f_{\mu,\nu}$ is the differential of $\mathcal{F}(\nu)$. Interestingly, $\mathcal{F}(\nu)$ admits a *free-energy* expression:

$$\mathcal{F}(\nu) = \int V(x)\,\mathrm{d}\nu(x) + \frac{1}{2}\int W(x,y)\,\mathrm{d}\nu(x)\,\mathrm{d}\nu(y) + C. \tag{6.3}$$

where V is a confinement potential, W an interaction potential and C a constant defined by:

$$V(x) = -\int k(x,x')\,\mathrm{d}\mu(x'), \quad W(x,x') = k(x,x') \tag{6.4}$$
$$C = \frac{1}{2}\int k(x,x')\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x').$$

Formulation (6.3) and the simple expression of the differential in (6.2) will be key to construct a gradient flow of $\mathcal{F}(\nu)$, to transport particles. In (6.4), V reflects the potential generated by μ and acting on each particle, while W reflects the potential arising from the interactions between those particles.

Gradient flow of the MMD. We consider now the problem of transporting mass from an initial distribution ν_0 to a target distribution μ, by finding a continuous path ν_t starting from ν_0 that converges to μ while decreasing $\mathcal{F}(\nu_t)$. Such a path should be physically plausible, in that teleportation phenomena are not allowed. For instance, the path $\nu_t = (1-e^{-t})\mu + e^{-t}\nu_0$ would constantly teleport mass between μ and ν_0 although it decreases $\mathcal{F}$ since $\mathcal{F}(\nu_t) = e^{-2t}\mathcal{F}(\nu_0)$ [Mroueh et al., 2019, Section 3.1, Case 1]. The physicality of the path is understood in terms of classical statistical physics: given an initial configuration ν_0 of N particles, these can move towards a new configuration μ through successive small transformations, without jumping from one location to another.

Optimal transport theory provides a way to construct such a continuous path by means of the *continuity equation*. Given a vector field V_t on $\mathcal{X}$ and an initial condition ν_0, the continuity equation is a partial differential equation which defines a path ν_t evolving under the action of the vector field V_t, and reads $\partial_t \nu_t = -div(\nu_t V_t)$ for all $t \geq 0$. The reader can find more detailed discussions in Definition 1 or Santambrogio [2015]. Following Ambrosio et al. [2008], a natural choice is to choose V_t as the negative gradient of the differential of $\mathcal{F}(\nu_t)$ at ν_t, since it corresponds to a gradient flow of $\mathcal{F}$ associated with the W_2 metric (see Section 2 .3). By (6.2), we know that the differential of $\mathcal{F}(\nu_t)$ at ν_t is given by f_{μ,ν_t}, hence $V_t(x) = -\nabla f_{\mu,\nu_t}(x)$.[1] The

[1] Also, $V_t = \nabla V + \nabla W \star \nu_t$ (see Section 2 .3) where $\star$ denotes the classical convolution.

gradient flow of $\mathcal{F}$ is then defined by the solution $(\nu_t)_{t\geq 0}$ of

$$\partial_t \nu_t = div(\nu_t \nabla f_{\mu,\nu_t}). \tag{6.5}$$

Equation (6.5) is non-linear in that the vector field depends itself on ν_t. This type of equation is associated in the probability theory literature to the so-called McKean-Vlasov process Kac [1956], McKean Jr [1966],

$$dX_t = -\nabla f_{\mu,\nu_t}(X_t)dt \qquad X_0 \sim \nu_0. \tag{6.6}$$

In fact, (6.6) defines a process $(X_t)_{t\geq 0}$ whose distribution $(\nu_t)_{t\geq 0}$ satisfies (6.5), as shown in Proposition 47. $(X_t)_{t\geq 0}$ can be interpreted as the trajectory of a single particle, starting from an initial random position X_0 drawn from ν_0. The trajectory is driven by the velocity field $-\nabla f_{\mu,\nu_t}$, and is affected by other particles. These interactions are captured by the velocity field through the dependence on the current distribution ν_t of all particles. Existence and uniqueness of a solution to (6.5) and (6.6) are guaranteed in the next proposition, whose proof is given Section A .1.1.

Proposition 47. *Let $\nu_0 \in \mathcal{P}_2(\mathcal{X})$. Then, under Assumption (A), there exists a unique process $(X_t)_{t\geq 0}$ satisfying the McKean-Vlasov equation in (6.6) such that $X_0 \sim \nu_0$. Moreover, the distribution ν_t of X_t is the unique solution of (6.5) starting from ν_0, and defines a gradient flow of $\mathcal{F}$.*

Besides existence and uniqueness of the gradient flow of $\mathcal{F}$, one expects $\mathcal{F}$ to decrease along the path ν_t and ideally to converge towards 0. The first property, stated in the next proposition, is rather easy to get and is the object of Proposition 48, similar to the result for KSD flow in [Mroueh et al., 2019, Section 3.1].

Proposition 48. *Under Assumption (A), $\mathcal{F}(\nu_t)$ is decreasing in time and satisfies:*

$$\frac{d\mathcal{F}(\nu_t)}{dt} = -\int \|\nabla f_{\mu,\nu_t}(x)\|^2 \, \mathrm{d}\nu_t(x). \tag{6.7}$$

This property results from (6.5) and the energy identity in [Ambrosio et al., 2008, Theorem 11.3.2] and is proved in Section A .1.1. From (6.7), $\mathcal{F}$ can be seen

as a Lyapunov functional for the dynamics defined by (6.5), since it is decreasing in time. Hence, the continuous-time gradient flow introduced in (6.5) allows to formally consider the notion of gradient descent on $\mathcal{P}_2(\mathcal{X})$ with $\mathcal{F}$ as a cost function. A time-discretized version of the flow naturally follows, and is provided in the next section.

2 .2 Euler scheme

We consider here a forward-Euler scheme of (6.5). For any $T : \mathcal{X} \to \mathcal{X}$ a measurable map, and $\nu \in \mathcal{P}_2(\mathcal{X})$, we denote the pushforward measure by $T_{\#}\nu$ (see Section 2). Starting from $\nu_0 \in \mathcal{P}_2(\mathcal{X})$ and using a step-size $\gamma > 0$, a sequence $\nu_n \in \mathcal{P}_2(\mathcal{X})$ is given by iteratively applying

$$\nu_{n+1} = (I - \gamma \nabla f_{\mu,\nu_n})_{\#}\nu_n. \tag{6.8}$$

For all $n \geq 0$, equation (6.8) is the distribution of the process defined by

$$X_{n+1} = X_n - \gamma \nabla f_{\mu,\nu_n}(X_n) \qquad X_0 \sim \nu_0. \tag{6.9}$$

The asymptotic behavior of (6.8) as $n \to \infty$ will be the object of Section 3 . For now, we provide a guarantee that the sequence $(\nu_n)_{n\in\mathbb{N}}$ approaches $(\nu_t)_{t\geq 0}$ as the step-size $\gamma \to 0$.

Proposition 49. *Let $n \geq 0$. Consider ν_n defined in (6.8), and the interpolation path ρ_t^γ defined as: $\rho_t^\gamma = (I - (t - n\gamma)\nabla f_{\mu,\nu_n})_{\#}\nu_n$, $\forall t \in [n\gamma, (n+1)\gamma)$. Then, under Assumption (A), $\forall\, T > 0$,*

$$W_2(\rho_t^\gamma, \nu_t) \leq \gamma C(T) \quad \forall t \in [0, T]$$

where $C(T)$ is a constant that depends only on T.

A proof of Proposition 49 is provided in Section A .1.2 and relies on standard techniques to control the discretization error of a forward-Euler scheme. Proposition 49 means that ν_n can be linearly interpolated giving rise to a path ρ_t^γ which

gets arbitrarily close to ν_t on bounded intervals. Note that as $T \to \infty$ the bound $C(T)$ it is expected to blow up. However, this result is enough to show that (6.8) is indeed a discrete-time flow of $\mathcal{F}$. In fact, provided that γ is small enough, $\mathcal{F}(\nu_n)$ is a decreasing sequence, as shown in Proposition 50.

Proposition 50. *Under Assumption (A), and for $\gamma \leq 2/3L$, the sequence $\mathcal{F}(\nu_n)$ is decreasing, and*

$$\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq -\gamma(1 - \frac{3\gamma}{2}L) \int \|\nabla f_{\mu,\nu_n}(x)\|^2 \, \mathrm{d}\nu_n(x), \quad \forall n \geq 0.$$

Proposition 50, whose proof is given in Section A .1.2, is a discrete analog of Proposition 48. In fact, (6.8) is intractable in general as it requires the knowledge of $\nabla f_{\mu,\nu_n}$ (and thus of ν_n) exactly at each iteration n. Nevertheless, we present in Section 4 .2 a practical algorithm using a finite number of samples which is provably convergent towards (6.8) as the sample-size increases. We thus begin by studying the convergence properties of the time discretized MMD flow (6.8) in the next section.

3 Convergence properties of the MMD flow

We are interested in analyzing the asymptotic properties of the gradient flow of $\mathcal{F}$. Although we know from Propositions 48 and 50 that $\mathcal{F}$ decreases in time, it can very well converge to local minima. One way to see this is by looking at the equilibrium condition for (6.7). As a non-negative and decreasing function, $t \mapsto \mathcal{F}(\nu_t)$ is guaranteed to converge towards a finite limit $l \geq 0$, which implies in turn that the r.h.s. of (6.7) converges to 0. If ν_t happens to converge towards some distribution ν^*, it is possible to show that the equilibrium condition (6.10) must hold [Mei et al., 2018, Prop. 2] ,

$$\int \|\nabla f_{\mu,\nu^*}(x)\|^2 \, \mathrm{d}\nu^*(x) = 0. \tag{6.10}$$

Condition (6.10) does not necessarily imply that ν^* is a global optimum. Thus convergence to a global optimum is not guaranteed unless suitable initial conditions hold and if the loss function has a particular structure Chizat and Bach [2018b].

For instance, the homogeneity condition on the loss function from Chizat and Bach [2018b] would hold if the kernel is linear in at least one of its dimensions. However, when a characteristic kernel is required (to ensure the MMD is a distance), such a structure can't be exploited. Similarly, the claim that KSD flow converges globally, [Mroueh et al., 2019, Prop. 3, Appendix B.1], requires an assumption [Mroueh et al., 2019, Assump. A] that excludes local minima which are not global (see Section D ; recall KSD is related to MMD). Global convergence of the flow is harder to obtain, and will be the topic of this section. The main challenge is the lack of convexity of $\mathcal{F}$ w.r.t. the Wassertein metric. We show that $\mathcal{F}$ is merely Λ-convex, and that standard optimization techniques only provide a loose bound on its asymptotic value. We next exploit a Lojasiewicz type inequality to prove convergence to the global optimum provided that a particular quantity remains bounded at all times.

3 .1 Optimization in a (W_2) non-convex setting

The *displacement convexity* of a functional $\mathcal{F}$ is an important criterion in characterizing the convergence of its Wasserstein gradient flow. Displacement convexity states that $t \mapsto \mathcal{F}(\rho_t)$ is a convex function whenever $(\rho_t)_{t\in[0,1]}$ is a path of minimal length between two distributions μ and ν (see Definition 2). Displacement convexity should not be confused with *mixture convexity*, which corresponds to the usual notion of convexity. As a matter of fact, $\mathcal{F}$ is mixture convex in that it satisfies: $\mathcal{F}(t\nu + (1-t)\nu') \leq t\mathcal{F}(\nu) + (1-t)\mathcal{F}(\nu')$ for all $t \in [0,1]$ and $\nu, \nu' \in \mathcal{P}_2(\mathcal{X})$ (see Lemma 67). Unfortunately, *$\mathcal{F}$ is not displacement convex*. Instead, $\mathcal{F}$ only satisfies a weaker notion of displacement convexity called Λ-displacement convexity, given in Definition 4 (Section 2 .4).

Proposition 51. *Under Assumptions (A) to (C), $\mathcal{F}$ is Λ-displacement convex, and satisfies*

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\nu) + t\mathcal{F}(\nu') - \int_0^1 \Lambda(\rho_s, v_s) G(s,t)\,\mathrm{d}s$$

for all $\nu, \nu' \in \mathcal{P}_2(\mathcal{X})$ and any displacement geodesic $(\rho_t)_{t\in[0,1]}$ from ν to ν' with velocity vectors $(v_t)_{t\in[0,1]}$. The functional Λ is defined for any pair (ρ, v) with

$\rho \in \mathcal{P}_2(\mathcal{X})$ and $\|v\| \in L_2(\rho)$,

$$\Lambda(\rho, v) = \left\| \int v(x).\nabla_x k(x, .)\, \mathrm{d}\rho(x) \right\|_{\mathcal{H}}^2 - \sqrt{2}\lambda d\mathcal{F}(\rho)^{\frac{1}{2}} \int \|v(x)\|^2\, \mathrm{d}\rho(x), \quad (6.11)$$

where $(s,t) \mapsto G(s,t) = s(1-t)\mathbb{1}\{s \leq t\} + t(1-s)\mathbb{1}\{s \geq t\}$ and λ is defined in Assumption (C).

Proposition 51 can be obtained by computing the second time derivative of $\mathcal{F}(\rho_t)$, which is then lower-bounded by $\Lambda(\rho_t, v_t)$ (see Section A .2.1). In (6.11), the map Λ is a difference of two non-negative terms: thus $\int_0^1 \Lambda(\rho_s, v_s)G(s,t)\,\mathrm{d}s$ can become negative, and displacement convexity does not hold in general. [Carrillo et al., 2006, Theorem 6.1] provides a convergence when only Λ-displacement convexity holds as long as either the potential or the interaction term is convex enough. In fact, as mentioned in [Carrillo et al., 2006, Remark 6.4], the convexity of either term could compensate for a lack of convexity of the other. Unfortunately, this cannot be applied for MMD since both terms involve the same kernel but with opposite signs. Hence, even under convexity of the kernel, a concave term appears and cancels the effect of the convex term. Moreover, the requirement that the kernel be positive semi-definite makes it hard to construct interesting convex kernels. However, it is still possible to provide an upper bound on the asymptotic value of $\mathcal{F}(\nu_n)$ when $(\nu_n)_{n\in\mathbb{N}}$ are obtained using (6.8). This bound is given in Theorem 52, and depends on a scalar $K(\rho^n) := \int_0^1 \Lambda(\rho_s^n, v_s^n)(1-s)\,\mathrm{d}s$, where $(\rho_s^n)_{s\in[0,1]}$ is a *constant speed displacement geodesic* from ν_n to the optimal value μ, with velocity vectors $(v_s^n)_{s\in[0,1]}$ of constant norm.

Theorem 52. *Let $\bar{K}$ be the average of $(K(\rho^j))_{0\leq j\leq n}$. Under Assumptions (A) to (C) and if $\gamma \leq 1/3L$,*

$$\mathcal{F}(\nu_n) \leq \frac{W_2^2(\nu_0, \mu)}{2\gamma n} - \bar{K}.$$

Theorem 52 is obtained using techniques from optimal transport and optimization. It relies on Proposition 51 and Proposition 50 to prove an *extended variational inequality* (see Proposition 61), and concludes using a suitable Lyapunov function.

A full proof is given in Section A .2.2. When $\bar{K}$ is non-negative, one recovers the usual convergence rate as $O(\frac{1}{n})$ for the gradient descent algorithm. However, $\bar{K}$ can be negative in general, and would therefore act as a barrier on the optimal value that $\mathcal{F}(\nu_n)$ can achieve when $n \to \infty$. In that sense, the above result is similar to [Bottou et al., 2018, Theorem 6.9]. Theorem 52 only provides a loose bound, however. In Section 3 .2 we show global convergence, under the boundedness at all times t of a specific distance between ν_t and μ.

3 .2 A condition for global convergence

The lack of convexity of $\mathcal{F}$, as shown in Section 3 .1, suggests that a finer analysis of the convergence should be performed. One strategy is to provide estimates for the dynamics in Proposition 48 using differential inequalities which can be solved using the Gronwall's lemma (see Oguntuase [2001]). Such inequalities are known in the optimization literature as Lojasiewicz inequalities (see Blanchet and Bolte [2018]), and upper-bound $\mathcal{F}(\nu_t)$ by the absolute value of its time derivative $\int \|\nabla f_{\mu,\nu_t}(x)\|^2 \,\mathrm{d}\nu_t(x)$. The latter is the squared *weighted Sobolev semi-norm* of f_{μ,ν_t} (see Section A .2.3), also written $\|f_{\mu,\nu_t}\|_{\dot{H}(\nu_t)}$. Thus one needs to find a relationship between $\mathcal{F}(\nu_t) = \frac{1}{2}\|f_{\mu,\nu_t}\|^2_{\mathcal{H}}$ and $\|f_{\mu,\nu_t}\|_{\dot{H}(\nu_t)}$. For this purpose, we consider the *weighted negative Sobolev distance* on $\mathcal{P}_2(\mathcal{X})$, defined by duality using $\|.\|_{\dot{H}(\nu)}$ (see also Peyre [2018]).

Definition 7. *Let $\nu \in \mathcal{P}_2(\mathcal{X})$, with its corresponding weighted Sobolev semi-norm $\|.\|_{\dot{H}(\nu)}$. The weighted negative Sobolev distance $\|p-q\|_{\dot{H}^{-1}(\nu)}$ between any p and q in $\mathcal{P}_2(\mathcal{X})$ is defined as*

$$\|p-q\|_{\dot{H}^{-1}(\nu)} = \sup_{f \in L_2(\nu), \|f\|_{\dot{H}(\nu)} \leq 1} \left| \int f(x) \,\mathrm{d}p(x) - \int f(x) \,\mathrm{d}q(x) \right| \tag{6.12}$$

with possibly infinite values.

Equation (6.12) plays a fundamental role in dynamic optimal transport. It can be seen as the minimum kinetic energy needed to advect the mass ν to q (see Mroueh

et al. [2019]). It is shown in Section A .2.3 that

$$\|f_{\mu,\nu_t}\|_{\mathcal{H}}^2 \leq \|f_{\mu,\nu_t}\|_{\dot{H}(\nu_t)}\|\mu-\nu_t\|_{\dot{H}^{-1}(\nu_t)}. \tag{6.13}$$

Provided that $\|\mu-\nu_t\|_{\dot{H}^{-1}(\nu_t)}$ remains bounded by some positive constant C at all times, (6.13) leads to a functional version of Lojasiewicz inequality for $\mathcal{F}$. It is then possible to use the general strategy explained earlier to prove the convergence of the flow to a global optimum:

Proposition 53. *Under Assumption (A),*

(i) If $\|\mu-\nu_t\|_{\dot{H}^{-1}(\nu_t)}^2 \leq C$, *for all* $t \geq 0$, *then:* $\mathcal{F}(\nu_t) \leq \frac{C}{C\mathcal{F}(\nu_0)^{-1}+4t}$,

(ii) If $\|\mu-\nu_n\|_{\dot{H}^{-1}(\nu_n)}^2 \leq C$ *for all* $n \geq 0$, *then:* $\mathcal{F}(\nu_n) \leq \frac{C}{C\mathcal{F}(\nu_0)^{-1}+4\gamma(1-\frac{3}{2}\gamma L)n}$.

Proofs of Proposition 53 (i) and (ii) are direct consequences of Propositions 48 and 50 and the bounded energy assumption: see Section A .2.3. The fact that (6.12) appears in the context of Wasserstein flows of $\mathcal{F}$ is not a coincidence. Indeed, (6.12) is a linearization of the Wasserstein distance (see Peyre [2018], Otto and Villani [2000] and Section E). Gradient flows of $\mathcal{F}$ defined under different metrics would involve other kinds of distances instead of (6.12). For instance, Rotskoff et al. [2019] consider gradient flows under a hybrid metric (a mixture between the Wasserstein distance and KL divergence), where convergence rates can then be obtained provided that the chi-square divergence $\chi^2(\mu\|\nu_t)$ remains bounded. As shown in Section E , $\chi^2(\mu\|\nu_t)^{\frac{1}{2}}$ turns out to linearize $KL(\mu\|\nu_t)^{\frac{1}{2}}$ when μ and ν_t are close. Hence, we conjecture that gradient flows of $\mathcal{F}$ under a metric d can be shown to converge when the linearization of the metric remains bounded. This can be verified on simple examples for $\|\mu-\nu_t\|_{\dot{H}^{-1}(\nu_t)}$ as discussed in Section B . However, it remains hard to guarantee this condition in general. One possible approach could be to regularize $\mathcal{F}$ using an estimate of (6.12). Indeed, Mroueh et al. [2019] considers the gradient flow of a regularized version of the negative Sobolev distance which can be written in closed form, and shows that this decreases the MMD. Combing both losses could improve the overall convergence properties of the MMD, albeit at additional

computational cost. In the next section, we propose a different approach to improve the convergence, and a particle-based algorithm to approximate the MMD flow in practice.

4 A practical algorithm to descend the MMD flow

4 .1 A noisy update as a regularization

We showed in Section 3 .1 that $\mathcal{F}$ is a non-convex functional, and derived a condition in Section 3 .2 to reach the global optimum. We now address the case where such a condition does not necessarily hold, and provide a regularization of the gradient flow to help achieve global optimality in this scenario. Our starting point will be the equilibrium condition in (6.10). If an equilibrium ν^* that satisfies (6.10) happens to have a positive density, then f_{μ,ν^*} would be constant everywhere. This in turn would mean that $f_{\mu,\nu^*} = 0$ when the RKHS does not contain constant functions, as for a gaussian kernel [Steinwart and Christmann, 2008, Corollary 4.44]. Hence, ν^* would be a global optimum since $\mathcal{F}(\nu^*) = 0$. The limit distribution ν^* might be singular, however, and can even be a dirac distribution [Mei et al., 2018, Theorem 6]. Although the gradient $\nabla f_{\mu,\nu^*}$ is not identically 0 in that case, (6.10) only evaluates it on the support ν^*, on which $\nabla f_{\mu,\nu^*} = 0$ holds. Hence a possible fix would be to make sure that the unnormalised witness gradient is also evaluated at points outside of the support of ν^*. Here, we propose to regularize the flow by injecting noise into the gradient during updates of (6.9),

$$X_{n+1} = X_n - \gamma \nabla f_{\mu,\nu_n}(X_n + \beta_n U_n), \qquad n \geq 0, \tag{6.14}$$

where U_n is a standard gaussian variable and β_n is the noise level at n. Compared to (6.8), the sample here is first blurred before evaluating the gradient. Intuitively, if ν_n approaches a local optimum ν^*, $\nabla f_{\mu,\nu_n}$ would be small on the support of ν_n but it might be much larger outside of it, hence evaluating $\nabla f_{\mu,\nu_n}$ outside the support of ν_n can help in escaping the local minimum. The stochastic process (6.14) is different from adding a diffusion term to (6.5). The latter case would correspond to

regularizing $\mathcal{F}$ using an entropic term as in Mei et al. [2018], Şimşekli et al. [2019] and was shown to converge to a global optimum that is in general different from the global minmum of the un-regularized loss. Eq. (6.14) is also different from Craig and Bertozzi [2016], Carrillo et al. [2019], where $\mathcal{F}$ (and thus its associated velocity field) is regularized by convolving the interaction potential W in (6.4) with a mollifier. The optimal solution of a regularized version of the functional $\mathcal{F}$ will be generally different from the non-regularized one, however, which is not desirable in our setting. Eq. (6.14) is more closely related to the *continuation methods* Gulcehre et al. [2016a,b], Chaudhari et al. [2017] and *graduated optimization* Hazan et al. [2016] used for non-convex optimization in Euclidian spaces, which inject noise into the gradient of a loss function F at each iteration. The key difference is the dependence of f_{μ,ν_n} of ν_n, which is inherently due to functional optimization. We show in Proposition 54 that (6.14) attains the global minimum of $\mathcal{F}$ provided that the level of the noise is well controlled, with the proof given in Section A .2.4.

Proposition 54. *Let $(\nu_n)_{n\in\mathbb{N}}$ be defined by (6.14) with an initial ν_0. Denote $\mathcal{D}_{\beta_n}(\nu_n) = \mathbb{E}_{x\sim\nu_n,u\sim g}[\|\nabla f_{\mu,\nu_n}(x+\beta_n u)\|^2]$ with g the density of the standard gaussian distribution. Under Assumptions (A) and (D), and for a choice of β_n such that*

$$8\lambda^2\beta_n^2\mathcal{F}(\nu_n) \leq \mathcal{D}_{\beta_n}(\nu_n), \tag{6.15}$$

the following inequality holds:

$$\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq -\frac{\gamma}{2}(1-3\gamma L)\mathcal{D}_{\beta_n}(\nu_n), \tag{6.16}$$

where λ and L are defined in Assumptions (A) and (D) and depend only on the choice of the kernel. Moreover if $\sum_{i=0}^{n}\beta_i^2 \to \infty$, then

$$\mathcal{F}(\nu_n) \leq \mathcal{F}(\nu_0)e^{-4\lambda^2\gamma(1-3\gamma L)\sum_{i=0}^{n}\beta_i^2}.$$

A particular case where $\sum_{i=0}^{n}\beta_i^2 \to \infty$ holds is when β_n decays as $1/\sqrt{n}$ while still satisfying (6.15). In this case, convergence occurs in polynomial time. At each

iteration, the level of the noise needs to be adjusted such that the gradient is not too blurred. This ensures that each step decreases the loss functional. However, β_n does not need to decrease at each iteration: it could increase adaptively whenever needed. For instance, when the sequence gets closer to a local optimum, it is helpful to increase the level of the noise to probe the gradient in regions where its value is not flat. Note that for $\beta_n = 0$ in (6.16) , we recover a similar bound to Proposition 50.

4 .2 The sample-based approximate scheme

We now provide a practical algorithm to implement the noisy updates in the previous section, which employs a discretization in space. The update (6.14) involves computing expectations of the gradient of the kernel k w.r.t the target distribution μ and the current distribution ν_n at each iteration n. This suggests a simple approximate scheme, based on samples from these two distributions, where at each iteration n, we model a system of N interacting particles $(X_n^i)_{1\leq i\leq N}$ and their empirical distribution in order to approximate ν_n. More precisely, given i.i.d. samples $(X_0^i)_{1\leq i\leq N}$ and $(Y^m)_{1\leq m\leq M}$ from ν_0 and μ and a step-size γ, the approximate scheme iteratively updates the i-th particle as

$$X_{n+1}^i = X_n^i - \gamma \nabla f_{\hat{\mu},\hat{\nu}_n}(X_n^i + \beta_n U_n^i), \tag{6.17}$$

where U_n^i are i.i.d standard gaussians and $\hat{\mu}$, $\hat{\nu}_n$ denote the empirical distributions of $(Y^m)_{1\leq m\leq M}$ and $(X_n^i)_{1\leq i\leq N}$, respectively. It is worth noting that for $\beta_n = 0$, (6.17) is equivalent to gradient descent over the particles (X_n^i) using a sample based version of the MMD. Implementing (6.17) is straightforward as it only requires to evaluate the gradient of k on the current particles and target samples. Pseudocode is provided in Algorithm 4. The overall computational cost of the algorithm at each iteration is $O((M+N)N)$ with $O(M+N)$ memory. The computational cost becomes $O(M+N)$ when the kernel is approximated using random features, as is the case for regression with neural networks (Section C). This is in contrast to the cubic cost of the flow of the KSD Mroueh et al. [2019], which requires solving a linear system at each iteration. The cost can also be compared to the algorithm

in Şimşekli et al. [2019], which involves computing empirical CDF and quantile functions of random projections of the particles.

The approximation scheme in (6.17) is a particle version of (6.14), so one would expect it to converge towards its population version (6.14) as M and N goes to infinity. This is shown below.

Theorem 55. *Let $n \geq 0$ and $T > 0$. Let ν_n and $\hat{\nu}_n$ defined by (6.8) and (6.17) respectively. Suppose Assumption (A) holds and that $\beta_n < B$ for all n, for some $B > 0$. Then for any $\frac{T}{\gamma} \geq n$:*

$$\mathbb{E}\left[W_2(\hat{\nu}_n, \nu_n)\right] \leq \frac{1}{4}\left(\frac{1}{\sqrt{N}}(B + var(\nu_0)^{\frac{1}{2}})e^{2LT} + \frac{1}{\sqrt{M}}var(\mu)^{\frac{1}{2}})\right)(e^{4LT} - 1)$$

Theorem 55 controls the propagation of the chaos at each iteration, and uses techniques from Jourdain et al. [2007]. Notice also that these rates remain true when no noise is added to the updates, i.e. for the original flow when $B = 0$. A proof is provided in Section A .3.

5 Experiments

5 .1 Student-Teacher networks

Experimental setting. We consider a student-teacher network setting similar to Chizat and Bach [2018b]. More precisely, using the notation from Section C , we denote by $\Psi(z, \nu)$ the neural network of the form: $\Psi(z, \nu) = \int \psi(z, x)\, \mathrm{d}\nu(x)$ where z is an input vector in $\mathbb{R}^p$ and ν is a probability distribution over the parameters x. Hence Ψ is an expectation over sub-networks $\psi(z, x)$ with parameters x. Here, we choose ψ of the form:

$$\psi(z, x) = G\left(b^1 + W^1\sigma(W^0 z + b^0)\right).$$

where x is obtained as the concatenation of the parameters $(b^1, W^1, b^0, W^0) \in \mathcal{X}$, σ is the ReLU non-linearity while G is a fixed function and is defined later. Note that using x to denote the parameters of a neural network is unusual, however, we

prefer to keep a notation which is consistent with the rest of this chapter. We will only consider the case when ν is given by an empirical distribution of N particles $X = (x^1, ...x^N)$ for some $N \in \mathbb{N}$. In that case, we denote by ν_X such distribution to stress the dependence on the particles X, i.e.: $\nu := \nu_X = \frac{1}{N}\sum_{i=1}^{N} \delta_{x^i}$. The teacher network $\Psi_T(z, \nu_\Xi)$ is given by M particles $\Xi = (\Xi_1, ..., \Xi_M)$ which are fixed during training and are initially drawn according to a normal distribution $\mathcal{N}(0, 1)$. Similarly, the student network $\Psi_S(z, \nu_X)$ has N particles $X = (x^1, ..., x^N)$ that are initialized according to a normal distribution $\mathcal{N}(10^{-3}, 1)$. Here we choose $M = 1$ and $N = 1000$. The inputs z are drawn from a uniform distribution $\mathbb{S}$ on the sphere in $\mathbb{R}^p$ as in Chizat and Bach [2018b] with $p = 50$. The number of hidden layers H is set to 3 and the output dimension is 1. The parameters of the student networks are trained to minimize the risk in (6.18) using SGD with mini-batches of size $n_b = 10^2$ and optimal step-size γ selected from: $\{10^{-3}, 10^{-2}, 10^{-1}\}$.

$$\min_X \mathbb{E}_{z\sim\mathbb{S}} \left[(\Psi_T(z, \nu_\Xi) - \Psi_S(z, \nu_X))^2\right] \tag{6.18}$$

When G is simply the identity function and no bias is used, one recovers the setting in Chizat and Bach [2018a]. In that case the network is partially 1-homogeneous and [Chizat and Bach, 2018a, Theorem 3.5] applies ensuring global optimality. Here, we are interested in the case when global optimality is not guaranteed by the homogeneity structure, hence we choose G to be a gaussian with fixed bandwidth $\sigma = 2$. As shown in Section C , performing gradient descent to minimize (6.18) can be seen as a particle version of the gradient flow of the MMD with a kernel given by $k(x, x') = \mathbb{E}_{z\sim\mathbb{S}}[\psi(z, x)\psi(z, x')]$ and target distribution μ given by $\mu = \nu_\Xi$. Hence one can use the noise injection algorithm defined in (6.17) to train the parameters of the student network. Since k is defined through an expectation over the data, it can be approximated using n_b data samples $\{z_1, ..., z_B\}$:

$$\hat{k}(x, x') = \frac{1}{n_b}\sum_{b=1}^{n_b} \psi(z_b, x)\psi(z_b, x'). \tag{6.19}$$

Such approximation of the kernel leads to a simple expression for the gradient

of the un-normalised witness function between ν_Ξ and ν_X defined for any $x \in \mathcal{X}$:

$$\nabla \hat{f}_{\nu_\Xi,\nu_X}(x) = \frac{1}{n_b}\sum_{b=1}^{n_b}\left(\frac{1}{M}\sum_{j=1}^{M}\psi(z_b,\Xi^j) - \frac{1}{N}\sum_{i=1}^{N}\psi(z_b,x^i)\right)\nabla_x\psi(z_b,x).$$

Algorithm 5, provides the main steps to train the parameters of the student network using the noisy gradient flow of the MMD proposed in (6.17). It can be easily implemented using automatic differentiation packages like `PyTorch`. Indeed, one only needs to compute an auxiliary loss function $\mathcal{F}_{aux}$ instead of the actual MMD loss $\mathcal{F}$ and perform gradient descent using $\mathcal{F}_{aux}$. Such function is given by:

$$\mathcal{F}_{aux} = \frac{1}{n_b}\sum_{i=1}^{N}\sum_{b=1}^{n_b}\left(\text{NoGrad}\left(y_S^b\right) - y_T^b\right)\psi(z^b,\widetilde{x}_n^i)$$

To compute $\mathcal{F}_{aux}$, two forward passes on the student network are required. A first forward pass using the current parameter values $X_n = (x_n^1,...,x_n^N)$ of the student network is used to compute the predictions y_S^b given an input z^b. For such forward pass, the gradient w.r.t to the parameters X_n is not used. This is enforced, here, formally by calling the function `NoGrad`. The second forward pass is performed using the noisy parameters $\widetilde{x}_n^i = x_n^i + \beta_n u_n^i$ and requires implementing special layers which can inject noise to the weights. This second forward pass will be used to provide a gradient to update the particles using back-propagation. Indeed, it is easy to see that $\nabla_{x_n^i}\mathcal{F}_{aux}$ gives exactly the gradient $\nabla \hat{f}_{\nu_\Xi,\nu_X}(\widetilde{x}_n^i)$ used in Algorithm 5.

Results. Figure 6.1 illustrates the behavior of the proposed algorithm (6.17) in a simple setting and compares it with three other methods: MMD without noise injection (blue traces), MMD with diffusion (green traces) and KSD (purple traces, Mroueh et al. [2019]). Here, a student network is trained to produce the outputs of a teacher network using gradient descent. More details on the experiment are provided in Section 5 .1. As discussed in Section C , this setting can be seen as a *stochastic* version of the MMD flow since the kernel is estimated using random features at each iteration ((6.19) in Section 5 .1). Here, the MMD flow fails to converge towards the global optimum. Such behavior is consistent with the observations in Chizat

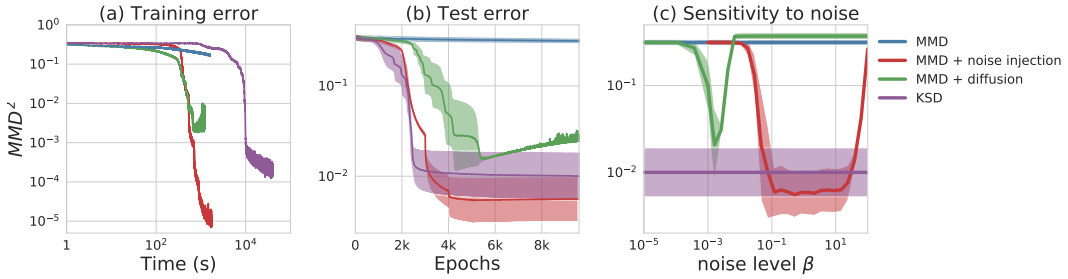


Figure 6.1: Comparison between different training methods for student-teacher ReLU networks with gaussian output non-linearity and synthetic data uniform on a hypersphere. In blue, (6.17) is used without noise $\beta_n = 0$ while in red noise is added with the following schedule: $\beta_0 > 0$ and β_n is decreased by half after every 10^3 epochs. In green, a diffusion term is added to the particles with noise level kept fixed during training ($\beta_n = \beta_0$). In purple, the KSD is used as a cost function instead of the MMD. In all cases, the kernel is estimated using random features (RF) with a batch size of 10^2. Best step-size was selected for each method from $\{10^{-3}, 10^{-2}, 10^{-1}\}$ and was used for 10^4 epochs on a dataset of 10^3 samples (RF). Initial parameters of the networks are drawn from i.i.d. gaussians: $\mathcal{N}(0, 1)$ for the teacher and $\mathcal{N}(10^{-3}, 1)$ for the student. Results are averaged over 10 different runs.

and Bach [2018b] when the parameters are initialized from a gaussian noise with relatively high variance (which is the case here). On the other hand, adding noise to the gradient seems to lead to global convergence. Indeed, the training error decreases below 10^{-5} and leads to much better validation error. While adding a small diffusion term (green) help convergence, the noise-injection (red) still outperforms it. This also holds for KSD (purple) which leads to a good solution (b) although at a much higher computational cost (a). Our noise injection method (red) is also robust to the amount of noise and achieves best performance over a wide region (c). On the other hand, MMD + diffusion (green) performs well only for much smaller values of noise that are located in a narrow region. This is expected since adding a diffusion changes the optimal solution, unlike the injection where the global optimum of the MMD remains a fixed point of the algorithm.

Another illustrative experiment on a simple flow between Gaussians is given in Section 5 .2.

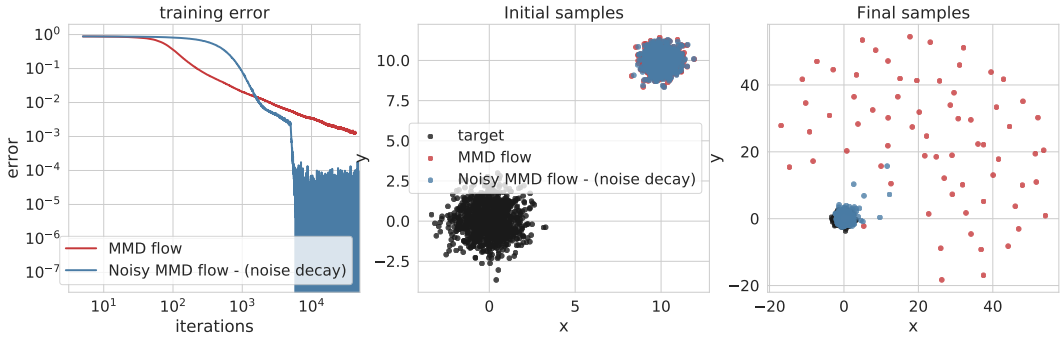


Figure 6.2: Gradient flow of the MMD from a gaussian initial distributions $\nu_0 \sim \mathcal{N}(10, 0.5)$ towards a target distribution $\mu \sim \mathcal{N}(0, 1)$ using $N = M = 1000$ samples from μ and ν_0 and a gaussian kernel with bandwidth $\sigma = 2$. (6.17) is used without noise $\beta_n = 0$ in red and with noise $\beta_n = 10$ up to $n = 5000$, then $\beta_n = 0$ afterwards in blue. The left figure shows the evolution of the MMD at each iteration. The middle figure shows the initial samples (black for μ), and the right figure shows the final samples after 10^5 iterations with step-size $\gamma = 0.1$.

5 .2 Learning gaussians

Figure 6.2 illustrates the behavior of the proposed algorithm (6.17) in a simple setting, and compares it with the gradient flow of the MMD without noise injection. In this setting, the MMD flow fails to converge to the global optimum. Indeed, as shown in Figure 6.2(right), some of the final samples (in red) obtained using noise-free gradient updates tend to get further away from the target samples (in black). Most of the remaining samples collapse to a unique point at the center near the origin. This can also be seen from Figure 6.2(left) where the training error fails to decrease below 10^{-3}. On the other hand, adding noise to the gradient seems to lead to global convergence, as seen visually from the samples. The training error decreases below 10^{-4} and oscillates between 10^{-8} and 10^{-4}. The oscillation is due to the step-size, which remained fixed while the noise was set to 0 starting from iteration 5000. It is worth noting that adding noise to the gradient slows the speed of convergence, as one can see from Figure 6.2(left). This is expected since the algorithm doesn't follow the path of steepest descent. The noise helps in escaping local optima, however, as illustrated here.

Supplementary

A Proofs

A .1 Construction of the W_2 gradient flow of the MMD

A .1.1 Existence of the continuous time flow

Existence and uniqueness of a solution to (6.5) and (6.6) is guaranteed under Lipschitz regularity of ∇k.

Proof of Proposition 47. [Existence and uniqueness] Under Assumption (**A**), the map $(x, \nu) \mapsto \nabla f_{\mu,\nu}(x) = \int \nabla k(x, .)d\nu - \int \nabla k(x, .)d\mu$ is Lipschitz continuous on $\mathcal{X} \times \mathcal{P}_2(\mathcal{X})$ (endowed with the product of the canonical metric on $\mathcal{X}$ and W_2 on $\mathcal{P}_2(\mathcal{X})$), see Proposition 63. Hence, we benefit from standard existence and uniqueness results of McKean-Vlasov processes (see Jourdain et al. [2007]). Then, it is straightforward to verify that the distribution of (6.6) is solution of (6.5) by ItÃŽ's formula (see Itô [1951]). The uniqueness of the gradient flow, given a starting distribution ν_0, results from the λ-convexity of $\mathcal{F}$ (for $\lambda = 3L$) which is given by Lemma 59, and [Ambrosio et al., 2008, Theorem 11.1.4]. The existence derive from the fact that the sub-differential of $\mathcal{F}$ is single-valued, as stated by (6.2), and that any ν_0 in $\mathcal{P}_2(\mathcal{X})$ is in the domain of $\mathcal{F}$. One can then apply [Ambrosio et al., 2008, Theorem 11.1.6 and Corollary 11.1.8]. □

Proof of Proposition 48. [Decay of the MMD] Recalling the discussion in Section 2 .3, the time derivative of $\mathcal{F}(\nu_t)$ along the flow is formally given by (2.8)

$$\frac{d\mathcal{F}(\nu_t)}{dt} = -D(\nu_t) \quad \text{with } D(\nu) = \int \left\| \nabla \frac{\partial \mathcal{F}(\nu_t(x))}{\partial \nu} \right\|^2 \nu_t(x)dx.$$

But we know from (6.2) that the strong differential $\nabla \frac{\delta \mathcal{F}(\nu)}{\delta \nu}$ is given by $\nabla f_{\mu,\nu}$. Therefore, one formally obtains the desired expression by exchanging the order of derivation and integration, performing an integration by parts and using the continuity equation (see (2.7)) that we recall here:

$$\frac{\partial \nu}{\partial t} = div(\nu \nabla \frac{\partial \mathcal{F}}{\partial \nu}) = div(\nu \nabla (U'(\nu) + V + W * \nu)).$$

We refer to Mroueh et al. [2019] for similar calculations. One can also obtain directly the same result using the energy identity in [Ambrosio et al., 2008, Theorem 11.3.2] which holds for λ-displacement convex functionals. The result applies here since, by Lemma 59, we know that $\mathcal{F}$ is λ-displacement convex with $\lambda = 3L$. □

A .1.2 Time-discretized flow

We prove that (6.8) approximates (6.5). To make the dependence on the step-size γ explicit, we will write: $\nu^\gamma_{n+1} = (I - \gamma \nabla f_{\mu,\nu^\gamma_n})_\# \nu^\gamma_n$ (so $\nu^\gamma_n = \nu_n$ for any $n \geq 0$). We start by introducing an auxiliary sequence $\bar{\nu}^\gamma_n$ built by iteratively applying $\nabla f_{\mu,\nu_{\gamma n}}$ where $\nu_{\gamma n}$ is the solution of (6.5) at time $t = \gamma n$:

$$\bar{\nu}^\gamma_{n+1} = (I - \gamma \nabla f_{\mu,\nu_{\gamma n}})_\# \bar{\nu}^\gamma_n \tag{6.20}$$

with $\bar{\nu}_0 = \nu_0$. Note that the latter sequence involves the continuous-time process ν_t of (6.5) with $t = \gamma n$. Using ν^γ_n, we also consider the interpolation path $\rho^\gamma_t = (I - (t - n\gamma)\nabla f_{\mu,\nu^\gamma_n})_\# \nu^\gamma_n$ for all $t \in [n\gamma, (n+1)\gamma)$ and $n \in \mathbb{N}$, which is the same as in Proposition 49.

Proof of Proposition 49. Let π be an optimal coupling between ν^γ_n and $\nu_{\gamma n}$, and (x, y) a sample from π. For $t \in [n\gamma, (n+1)\gamma)$ we write $y_t = y_{n\gamma} - \int_{n\gamma}^t \nabla f_{\mu,\nu_s}(y_u)\, du$ and $x_t = x - (t - n\gamma)\nabla f_{\mu,\nu^\gamma_n}(x)$ where $y_{n\gamma} = y$. We also introduce the approximation error $E(t, n\gamma) := y_t - y + (t - n\gamma)\nabla f_{\mu,\nu_{\gamma n}}(y)$ for which we know by Lemma 58 that $\mathcal{E}(t, n\gamma) := \mathbb{E}[E(t, n\gamma)^2]^{\frac{1}{2}}$ is upper-bounded by $(t - n\gamma)^2 C$ for some positive constant C that depends only on T and the Lipschitz constant L. This allows to

write:

$$
\begin{aligned}
W_2(\rho_t^\gamma, \nu_t) &\leq \mathbb{E}\left[\left\|y - x + (t - n\gamma)(\nabla f_{\mu,\nu_n^\gamma}(x) - \nabla f_{\mu,\nu_{\gamma n}}(y)) + E(t, n\gamma)\right\|^2\right]^{\frac{1}{2}} \\
&\leq W_2(\nu_n^\gamma, \nu_{\gamma n}) + 4L(t - n\gamma)W_2(\nu_n^\gamma, \nu_{\gamma n}) + \mathcal{E}(t, n\gamma) \\
&\leq (1 + 4\gamma L)W_2(\nu_n^\gamma, \nu_{\gamma n}) + (t - \gamma n)^2 C \\
&\leq (1 + 4\gamma L)\left(W_2(\nu_n^\gamma, \bar{\nu}_n^\gamma) + W_2(\nu_{\gamma n}, \bar{\nu}_n^\gamma)\right) + \gamma^2 C \\
&\leq \gamma\left[(1 + 4\gamma L)\, M(T) + \gamma C\right]
\end{aligned}
$$

The second line is obtained using that $\nabla f_{\mu,\nu_{\gamma n}}(x)$ is jointly $2L$-Lipschitz in x and ν (see Proposition 63) and by the fact that $W_2(\nu_n^\gamma, \nu_{\gamma n}) = \mathbb{E}_\pi[\|y - x\|^2]^{\frac{1}{2}}$. The third one is obtained using $t - n\gamma \leq \gamma$. For the last inequality, we used Lemmas 56 and 57 where $M(T)$ is a constant that depends only on T. Hence for $\gamma \leq \frac{1}{4L}$ we get $W_2(\rho_t^\gamma, \nu_t) \leq \gamma(\frac{C}{4L} + 2M(T))$. □

Lemma 56. *For any $n \geq 0$:*

$$
W_2(\nu_{\gamma n}, \bar{\nu}_n^\gamma) \leq \gamma \frac{C}{2L}(e^{n\gamma 2L} - 1)
$$

Proof. Let π be an optimal coupling between $\bar{\nu}_n^\gamma$ and $\nu_{\gamma n}$ and $(\bar{x},\ x)$ a joint sample from π. Consider also the joint sample $(\bar{y}, y)$ obtained from $(\bar{x}\ , x)$ by applying the gradient flow of $\mathcal{F}$ in continuous time to get $y := x_{(n+1)\gamma} = x_{n\gamma} - \int_{n\gamma}^{(n+1)\gamma} \nabla f_{\mu,\nu_s}(x_u)\,\mathrm{d}u$ with $x_{n\gamma} = x$ and by taking a discrete step from $\bar{x}$ to write $\bar{y} = \bar{x} - \gamma\nabla f_{\mu,\nu_{\gamma n}}(\bar{x})$. It is easy to see that $y \sim \nu_{\gamma(n+1)}$ (i.e. a sample from the continous process (6.5) at time $t = (n+1)\gamma$) and $\bar{y} \sim \bar{\nu}_{n+1}^\gamma$ (i.e. a sample from (6.20)). Moreover, we introduce the approximation error $E((n+1)\gamma, n\gamma) := y - x + \gamma\nabla f_{\mu,\nu_{\gamma n}}(x)$ for which we know by Lemma 58 that $\mathcal{E}((n+1)\gamma, n\gamma) := \mathbb{E}[E((n+1)\gamma, n\gamma)^2]^{\frac{1}{2}}$ is upper-bounded by $\gamma^2 C$ for some positive constant C that depends only on T and the Lipschitz constant L. Denoting by

$a_n = W_2(\nu_{\gamma n}, \bar{\nu}_n^\gamma)$, one can therefore write:

$$\begin{aligned} a_{n+1} \leq & \mathbb{E}_\pi \left[\left\| x - \gamma \nabla f_{\mu,\nu_{\gamma n}}(x) - \bar{x} + \gamma \nabla f_{\mu,\nu_{\gamma n}}(\bar{x}) + E((n+1)\gamma, n\gamma) \right\|^2 \right]^{\frac{1}{2}} \\ \leq & \mathbb{E}_\pi \left[\|x - \bar{x}\|^2 \right]^{\frac{1}{2}} + \gamma \mathbb{E}_\pi \left[\left\| \nabla f_{\mu,\nu_{\gamma n}}(x) - \nabla f_{\mu,\nu_{\gamma n}}(\bar{x})) \right\|^2 \right]^{\frac{1}{2}} + \gamma^2 C \end{aligned}$$

Using that $\nabla f_{\mu,\nu_{\gamma n}}$ is $2L$-Lipschitz by Proposition 63 and recalling that $\mathbb{E}_\pi \left[\|x - \bar{x}\|^2 \right]^{\frac{1}{2}} = W_2(\nu_{\gamma n}, \bar{\nu}_n^\gamma)$, we get the recursive inequality $a_{n+1} \leq (1 + 2\gamma L) a_n + \gamma^2 C$. Finally, using Lemma 68 and recalling that $a_0 = 0$, since by definition $\bar{\nu}_0^\gamma = \nu_0^\gamma$, we conclude that $a_n \leq \gamma \frac{C}{2L}(e^{n\gamma 2L} - 1)$. □

Lemma 57. *For any $T > 0$ and n such that $n\gamma \leq T$*

$$W_2(\nu_n^\gamma, \bar{\nu}_n^\gamma) \leq \gamma \frac{C}{8L^2}(e^{4TL} - 1)^2$$

Proof. Consider now an optimal coupling π between $\bar{\nu}_n^\gamma$ and ν_n^γ. Similarly to Lemma 56, we denote by $(\bar{x}, x)$ a joint sample from π and $(\bar{y}, y)$ is obtained from $(\bar{x}, x)$ by applying the discrete updates : $\bar{y} = \bar{x} - \gamma \nabla f_{\mu,\nu_{\gamma n}}(\bar{x})$ and $y = x - \gamma \nabla f_{\mu,\nu_n^\gamma}(x)$. We again have that $y \sim \nu_{n+1}^\gamma$ (i.e. a sample from the time discretized process (6.8)) and $\bar{y} \sim \bar{\nu}_{n+1}^\gamma$ (i.e. a sample from (6.20)). Now, denoting by $b_n = W_2(\nu_n^\gamma, \bar{\nu}_n^\gamma)$, it is easy to see from the definition of $\bar{y}$ and y that we have:

$$\begin{aligned} b_{n+1} &\leq \mathbb{E}_\pi \left[\left\| x - \gamma \nabla f_{\mu,\nu_n^\gamma}(x) - \bar{x} + \gamma \nabla f_{\mu,\nu_{\gamma n}}(\bar{x}) \right\|^2 \right]^{\frac{1}{2}} \\ &\leq (1 + 2\gamma L) \mathbb{E}_\pi \left[\|x - \bar{x}\|^2 \right]^{\frac{1}{2}} + 2\gamma L W_2(\nu_n^\gamma, \nu_{\gamma n})) \\ &\leq (1 + 4\gamma L) b_n + \gamma L W_2(\bar{\nu}_n^\gamma, \nu_{\gamma n}) \end{aligned}$$

The second line is obtained recalling that $\nabla f_{\mu,\nu}(x)$ is $2L$-Lipschitz in both x and ν by Proposition 63. The third line follows by triangular inequality and using $\mathbb{E}_\pi \left[\|x - \bar{x}\|^2 \right]^{\frac{1}{2}} = W_2(\nu_n^\gamma, \bar{\nu}_n^\gamma) = b_n$, since π is an optimal coupling between $\bar{\nu}_n^\gamma$ and ν_n^γ. By Lemma 56, we have $W_2(\bar{\nu}_n^\gamma, \nu_{\gamma n}) \leq \gamma \frac{C}{2L}(e^{2n\gamma L} - 1)$, hence, for any n such that $n\gamma \leq T$ we get the recursive inequality

$$b_{n+1} \leq (1 + 4\gamma L) b_n + (C/2L)\gamma^2 (e^{2TL} - 1).$$

Finally, using again Lemma 68, it follows that $b_n \leq \gamma \frac{C}{8L^2}(e^{4TL} - 1)^2$. □

Lemma 58. *[Taylor expansion] Consider the process $\dot{x}_t = -\nabla f_{\mu,\nu_t}(x_t)$, and denote by $\mathcal{E}(t,s) = \mathbb{E}[\|x_t - x_s + (t-s)\nabla f_{\mu,\nu_s}(x_s)\|^2]^{\frac{1}{2}}$ for $0 \leq s \leq t \leq T$. Then one has:*

$$\mathcal{E}(t,s) \leq 2L^2 r_0 e^{LT}(t-s)^2$$

with $r_0 = \mathbb{E}_{(x,z)\sim\nu_0\otimes\mu}[\|x - z\|]$

Proof. By definition of x_t and $\mathcal{E}(t,s)$ one can write:

$$\begin{aligned}
\mathcal{E}(t,s) &= \mathbb{E}\left[\left\|\int_s^t (\nabla f_{\mu,\nu_s}(x_s) - \nabla f_{\mu,\nu_u}(x_u))\,\mathrm{d}u\right\|^2\right]^{\frac{1}{2}} \\
&\leq \int_s^t \mathbb{E}\left[\|(\nabla f_{\mu,\nu_s}(x_s) - \nabla f_{\mu,\nu_u}(x_u))\|^2\right]^{\frac{1}{2}} \mathrm{d}u \\
&\leq 2L\int_s^t \mathbb{E}\left[(\|x_s - x_u\| + W_2(\nu_s,\nu_u))^2\right]^{\frac{1}{2}} \mathrm{d}u \leq 4L\int_s^t \mathbb{E}\left[\|x_s - x_u\|^2\right]^{\frac{1}{2}} \mathrm{d}u
\end{aligned}$$

Where we used an integral expression for x_t in the first line then applied a triangular inequality for the second line. The last line is obtained recalling that $\nabla f_{\mu,\nu}(x)$ is jointly $2L$-Lipschitz in x and ν by Proposition 63 and that $W_2(\nu_s,\nu_u) \leq \mathbb{E}\left[\|x_s - x_u\|^2\right]^{\frac{1}{2}}$. Now we use again an integral expression for x_u which further gives:

$$\begin{aligned}
\mathcal{E}(t,s) \leq& 4L\int_s^t \mathbb{E}\left[\left\|\int_s^u \nabla f_{\mu,\nu_l}(x_l)\,\mathrm{d}l\right\|^2\right]^{\frac{1}{2}} \mathrm{d}u \\
\leq& 4L\int_s^t\int_s^u \mathbb{E}\left[\|\mathbb{E}\left[\nabla_1 k(x_l,x_l') - \nabla_1 k(x_l,z)\right]\|^2\right]^{\frac{1}{2}} \mathrm{d}l\,\mathrm{d}u \\
\leq& 4L^2\int_s^t\int_s^u \mathbb{E}\left[\|x_l' - z\|\right]\mathrm{d}l\,\mathrm{d}u
\end{aligned}$$

Again, the second line is obtained using a triangular inequality and recalling the expression of $\nabla f_{\mu,\nu}(x)$ from Proposition 63. The last line uses that ∇k is L-Lipschitz by Assumption **(A)**. Now we need to make sure that $\|x_l' - z\|$ remains bounded at finite times. For this we will first show that $r_t = \mathbb{E}[\|x_t - z\|]$ satisfies an integro-

differential inequality:

$$\begin{aligned} r_t \leq & \mathbb{E}\left[\left\| x_0 - z - \int_0^t \nabla f_{\mu,\nu_s}(x_s)\,\mathrm{d}s \right\|\right] \\ \leq & r_0 + \int_0^t \mathbb{E}\left[\|\nabla_1 k(x_s, x_s') - \nabla_1 k(x_s, z)\|\right] \mathrm{d}s \leq r_0 + L\int_0^t r_s\,\mathrm{d}s \end{aligned}$$

Again, we used an integral expression for x_t in the first line, then a triangular inequality recalling the expression of $\nabla f_{\mu,\nu_s}$. The last line uses again that ∇k is L-Lipschitz. By Gronwall's lemma it is easy to see that $r_t \leq r_0 e^{Lt}$ at all times. Moreover, for all $t \leq T$ we have a fortiori that $r_t \leq r_0 e^{LT}$. Recalling back the upper-bound on $\mathcal{E}(t,s)$ we have finally:

$$\mathcal{E}(t,s) \leq 4L^2 r_0 e^{LT} \int_s^t \int_s^u \mathrm{d}l\,\mathrm{d}u = 2L^2 r_0 e^{LT}(t-s)^2$$

□

We show now that (6.8) decreases the functional $\mathcal{F}$. In all the proofs, the step-size γ is fixed.

Proof of Proposition 50. Consider a path between ν_n and ν_{n+1} of the form $\rho_t = (I - \gamma t \nabla f_{\mu,\nu_n})_{\#}\nu_n$. We know by Proposition 63 that $\nabla f_{\mu,\nu_n}$ is $2L$ Lipschitz, thus by Lemma 64 and using $\phi(x) = -\gamma \nabla f_{\mu,\nu_n}(x)$, $\psi(x) = x$ and $q = \nu_n$ it follows that $\mathcal{F}(\rho_t)$ is differentiable and hence absolutely continuous. Therefore one can write:

$$\mathcal{F}(\rho_1) - \mathcal{F}(\rho_0) = \dot{\mathcal{F}}(\rho_0) + \int_0^1 \dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_0) dt. \tag{6.21}$$

Moreover, Lemma 64 also allows to write:

$$\begin{aligned} \dot{\mathcal{F}}(\rho_0) &= -\gamma \int \|\nabla f_{\mu,\nu_n}(x)\|^2 d\nu_n(x); \\ |\dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_0)| &\leq 3Lt\gamma^2 \int \|\nabla f_{\mu,\nu_n}(X)\|^2 d\nu_n(X). \end{aligned}$$

where $t \leq 1$. Hence, the result follows directly by applying the above expression to (6.21). □

A .2 Convergence of the W_2 gradient flow of the MMD

A .2.1 Λ-displacement convexity of the MMD

We provide now a proof of Proposition 51:

Proof of Proposition 51. [Λ- displacement convexity of the MMD] To prove that $\nu \mapsto \mathcal{F}(\nu)$ is Λ-convex we need to compute the second time derivative $\ddot{\mathcal{F}}(\rho_t)$ where $(\rho_t)_{t\in[0,1]}$ is a displacement geodesic between two probability distributions ν_0 and ν_1 as defined in (2.1):

$$\rho_t = (s_t)_{\#}\pi^*, s_t(x,y) = (1-t)x + ty.$$

Such geodesic always exists and can be written as $\rho_t = (s_t)_{\#}\pi$ with $s_t = x + t(y-x)$ for all $t \in [0,1]$ and π is an optimal coupling between ν_0 and ν_1 (Santambrogio [2015], Theorem 5.27). We denote by V_t the corresponding velocity vector defined by the continuity equation:

$$\partial_t \rho_t + div(\rho_t V_t) = 0 \qquad \forall t \in [0,1].$$

Recall that $\mathcal{F}(\rho_t) = \frac{1}{2}\|f_{\mu,\rho_t}\|^2_{\mathcal{H}}$, with f_{μ,ρ_t} defined in (6.1). We start by computing the first derivative of $t \mapsto \mathcal{F}(\rho_t)$. Since Assumptions **(A)** and **(B)** hold, Lemma 65 applies for $\phi(x,y) = y - x$, $\psi(x,y) = x$ and $q = \pi$, thus we know that $\ddot{\mathcal{F}}(\rho_t)$ is well defined and given by:

$$\begin{aligned}\ddot{\mathcal{F}}(\rho_t) =& \mathbb{E}\left[(y-x)^T \nabla_1 \nabla_2 k(s_t(x,y), s_t(x',y'))(y'-x')\right] \\ &+ \mathbb{E}\left[(y-x)^T (H_1 k(s_t(x,y), s_t(x',y')) - H_1 k(s_t(x,y), z))(y-x)\right]\end{aligned} \tag{6.22}$$

Moreover, Assumption **(C)** also holds which means by Lemma 65 that the second term in (6.22) can be lower-bounded by $-\sqrt{2}\lambda d\mathcal{F}(\rho_t)\mathbb{E}[\|y-x\|^2]$ so that:

$$\ddot{\mathcal{F}}(\rho_t) = \mathbb{E}\left[(y-x)^T \nabla_1 \nabla_2 k(s_t(x,y), s_t(x',y'))(y'-x')\right] - \sqrt{2}\lambda d\mathcal{F}(\rho_t)\mathbb{E}[\|y-x\|^2]$$

Recall now that $(\rho_t)_{t\in[0,1]}$ is a constant speed geodesic with velocity vector $(V_t)_{t\in[0,1]}$

thus by a change of variable, one further has:

$$\ddot{\mathcal{F}}(\rho_t) \geq \int \left[V_t^T(x)\nabla_1\nabla_2 k(x,x')V_t(x')\right] \mathrm{d}\rho_t(x) - \sqrt{2}\lambda d\mathcal{F}(\rho_t) \int \|V_t(x)\|^2 \,\mathrm{d}\rho_t(x).$$

Now we can introduce the function $\Lambda(\rho, v) = \langle v, (C_\rho - \sqrt{2}\lambda d\mathcal{F}(\rho)^{\frac{1}{2}} I)v\rangle_{L_2(\rho)}$ which is defined for any pair (ρ, v) with $\rho \in \mathcal{P}_2(\mathcal{X})$ and v a square integrable vector field in $L_2(\rho)$ and where C_ρ is a non-negative operator given by $(C_\rho v)(x) = \int \nabla_x \nabla_{x'} k(x,x')v(x')d\rho(x')$ for any $x \in \mathcal{X}$. This allows to write $\ddot{\mathcal{F}}(\rho_t) \geq \Lambda(\rho_t, V_t)$. It is clear that $\Lambda(\rho, .)$ is a quadratic form on $L_2(\rho)$ and satisfies the requirement in Definition 3. Finally, using Lemma 66 and Definition 4 we conclude that $\mathcal{F}$ is Λ-convex. Moreover, by the reproducing property we also know that for all $\rho \in \mathcal{P}_2(\mathcal{X})$:

$$\mathbb{E}_\rho\left[v(x)^T\nabla_1\nabla_2 k(x,x')v(x')\right] = \mathbb{E}_\rho\left[\left\langle v(x)^T\nabla_1 k(x,.), v(x')^T\nabla_1 k(x',.)\right\rangle_{\mathcal{H}}\right].$$

By Bochner integrability of $v(x)^T\nabla_1 k(x,.)$ it is possible to exchange the order of the integral and the inner-product [Retherford, 1978, Theorem 6]. This leads to the expression $\|\mathbb{E}[v(x)^T\nabla_1 k(x,.)]\|^2_{\mathcal{H}}$. Hence $\Lambda(\rho, v)$ has a second expression of the form:

$$\Lambda(\rho, v) = \left\|\mathbb{E}_\rho\left[v(x)^T\nabla_1 k(x,.)\right]\right\|^2_{\mathcal{H}} - \sqrt{2}\lambda d\mathcal{F}(\rho)^{\frac{1}{2}}\mathbb{E}_\rho\left[\|v(x)\|^2\right].$$

□

We also provide a result showing Λ convexity for $\mathcal{F}$ only under Assumption **(A)**:

Lemma 59 (Λ-displacement convexity)**.** *Under Assumption (A), for any $\nu, \nu' \in \mathcal{P}_2(\mathcal{X})$ and any constant speed geodesic ρ_t from ν to ν', $\mathcal{F}$ satisfies for all $0 \leq t \leq 1$:*

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\nu) + t\mathcal{F}(\nu') + 3LW_2^2(\nu, \nu')$$

Proof. Let ρ_t be a constant speed geodesic of the form $\rho_t = s_t\#\pi$ where π is an optimal coupling between ν and ν' and $s_t(x,y) = x + t(y-x)$. Since Assumption **(A)**

holds, one can apply Lemma 64 with $\psi(x,y) = x$, $\phi(x,y) = y - x$ and $q = \pi$. Hence, one has that $\mathcal{F}(\rho_t)$ is differentiable and its differential satisfies:

$$|\dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_s)| \leq 3L|t-s| \int \|y-x\|^2 \,\mathrm{d}\pi(x,y)$$

This implies that $\dot{\mathcal{F}}(\rho_t)$ is Lipschitz continuous and therefore is differentiable for almost all $t \in [0,1]$ by Rademacher's theorem. Hence, $\ddot{\mathcal{F}}(\rho_t)$ is well defined for almost all $t \in [0,1]$. Moreover, from the above inequality it follows that $\ddot{\mathcal{F}}(\rho_t) \geq -3L \int \|y-x\|^2 \,\mathrm{d}\pi(x,y) = -3LW_2^2(\nu,\nu')$ for almost all $t \in [0,1]$. Using Lemma 66 it follows directly that $\mathcal{F}$ satisfies the desired inequality. □

A .2.2 Descent up to a barrier

To provide a proof of Theorem 52, we need the following preliminary results. Firstly, an upper-bound on a scalar product involving $\nabla f_{\mu,\nu}$ for any $\mu, \nu \in \mathcal{P}_2(\mathcal{X})$ in terms of the loss functional $\mathcal{F}$, is obtained using the Λ-displacement convexity of $\mathcal{F}$ in Lemma 60. Then, an EVI (Evolution Variational Inequality) is obtained in Proposition 61 on the gradient flow of $\mathcal{F}$ in W_2. The proof of the theorem is given afterwards.

Lemma 60. *Let ν be a distribution in $\mathcal{P}_2(\mathcal{X})$ and μ the target distribution such that $\mathcal{F}(\mu) = 0$. Let π be an optimal coupling between ν and μ, and $(\rho_t)_{t\in[0,1]}$ the displacement geodesic defined by (2.1)*

$$\rho_t = (s_t)_{\#}\pi^*, s_t(x,y) = (1-t)x + ty.$$

Let $(V_t)_{t\in[0,1]}$ be the velocity vector corresponding to ρ_t as defined by the continuity equation

$$\partial_t \rho_t + div(\rho_t V_t) = 0 \qquad \forall t \in [0,1].$$

Finally let $\nabla f_{\nu,\mu}(X)$ be the gradient of the unnormalised witness function between

μ and ν. The following inequality holds:

$$\int \nabla f_{\mu,\nu}(x).(y-x)d\pi(x,y) \leq \mathcal{F}(\mu) - \mathcal{F}(\nu) - \int_0^1 \Lambda(\rho_s, V_s)(1-s)ds$$

where Λ is defined Proposition 51.

Proof. Recall that for all $t \in [0,1]$, ρ_t is given by $\rho_t = (s_t)_{\#}\pi$ with $s_t = x + t(y-x)$. By Λ-convexity of $\mathcal{F}$ the following inequality holds:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\nu) + t\mathcal{F}(\mu) - \int_0^1 \Lambda(\rho_s, V_s)G(s,t)ds$$

Hence by bringing $\mathcal{F}(\nu)$ to the l.h.s and dividing by t and then taking its limit at 0 it follows that:

$$\dot{\mathcal{F}}(\rho_t)|_{t=0} \leq \mathcal{F}(\mu) - \mathcal{F}(\nu) - \int_0^1 \Lambda(\rho_s, V_s)(1-s)ds. \tag{6.23}$$

where $\dot{\mathcal{F}}(\rho_t) = d\mathcal{F}(\rho_t)/dt$ and since $\lim_{t\to 0} G(s,t) = (1-s)$. Moreover, under Assumption **(A)**, Lemma 64 applies for $\phi(x,y) = y - x$, $\psi(x,y) = x$ and $q = \pi$. It follows therefore that $\dot{\mathcal{F}}(\rho_t)$ is differentiable with time derivative given by: $\dot{\mathcal{F}}(\rho_t) = \int \nabla f_{\mu,\rho_t}(s_t(x,y)).(y-x)\,\mathrm{d}\pi(x,y)$. Hence at $t=0$ we get: $\dot{\mathcal{F}}(\rho_t)|_{t=0} = \int \nabla f_{\mu,\nu}(x).(y-x)\,\mathrm{d}\pi(x,y)$ which shows the desired result when used in (6.23). □

Proposition 61. *Consider the sequence of distributions ν_n obtained from (6.8). For $n \geq 0$, consider the scalar $K(\rho^n) := \int_0^1 \Lambda(\rho_s^n, V_s^n)(1-s)\,\mathrm{d}s$ where $(\rho_s^n)_{0\leq s\leq 1}$ is a constant speed displacement geodesic from ν_n to the optimal value μ with velocity vectors $(V_s^n)_{0\leq s\leq 1}$. If $\gamma \leq 1/L$, where L is the Lispchitz constant of ∇k in Assumption (A), then:*

$$2\gamma(\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\mu)) \leq W_2^2(\nu_n, \mu) - W_2^2(\nu_{n+1}, \mu) - 2\gamma K(\rho^n).$$

Proof. Let Π^n be the optimal coupling between ν_n and μ, then the optimal transport

between ν_n and μ is given by:

$$W_2^2(\mu, \nu_n) = \int \|X - Y\|^2 d\Pi^n(\nu_n, \mu)$$

Moreover, consider $Z = X - \gamma \nabla f_{\mu,\nu_n}(X)$ where (X, Y) are samples from π^n. It is easy to see that (Z, Y) is a coupling between ν_{n+1} and μ, therefore, by definition of the optimal transport map between ν_{n+1} and μ it follows that:

$$W_2^2(\nu_{n+1}, \mu) \leq \int \|X - \gamma \nabla f_{\mu,\nu_n}(X) - Y\|^2 d\pi^n(\nu_n, \mu) \tag{6.24}$$

By expanding the r.h.s in (6.24), the following inequality holds:

$$W_2^2(\nu_{n+1}, \mu) \leq W_2^2(\nu_n, \mu) - 2\gamma \int \langle \nabla f_{\mu,\nu_n}(X), X - Y \rangle d\pi^n(\nu_n, \mu) + \gamma^2 D(\nu_n)$$

where $D(\nu_n) = \int \|\nabla f_{\mu,\nu_n}(X)\|^2 d\nu_n$. By Lemma 60 it holds that:

$$-2\gamma \int \nabla f_{\mu,\nu_n}(X).(X - Y) d\pi(\nu, \mu) \leq -2\gamma \left(\mathcal{F}(\nu_n) - \mathcal{F}(\mu) + K(\rho^n)\right)$$

where $(\rho_t^n)_{0 \leq t \leq 1}$ is a constant-speed geodesic from ν_n to μ and $K(\rho^n) := \int_0^1 \Lambda(\rho_s^n, v_s^n)(1 - s) ds$. Note that when $K(\rho^n) \leq 0$ it falls back to the convex setting. Therefore, the following inequality holds:

$$W_2^2(\nu_{n+1}, \mu) \leq W_2^2(\nu_n, \mu) - 2\gamma \left(\mathcal{F}(\nu_n) - \mathcal{F}(\mu) + K(\rho^n)\right) + \gamma^2 D(\nu_n)$$

Now we introduce a term involving $\mathcal{F}(\nu_{n+1})$. The above inequality becomes:

$$\begin{aligned} W_2^2(\nu_{n+1}, \mu) \leq & W_2^2(\nu_n, \mu) - 2\gamma \left(\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\mu) + K(\rho^n)\right) \\ & + \gamma^2 D(\nu_n) - 2\gamma(\mathcal{F}(\nu_n) - \mathcal{F}(\nu_{n+1})) \end{aligned} \tag{6.25}$$

It is possible to upper-bound the last two terms on the r.h.s. by a negative quantity when the step-size is small enough. This is mainly a consequence of the smoothness of the functional $\mathcal{F}$ and the fact that ν_{n+1} is obtained by following the steepest

direction of $\mathcal{F}$ starting from ν_n. Proposition 50 makes this statement more precise and enables to get the following inequality:

$$\gamma^2 D(\nu_n) - 2\gamma(\mathcal{F}(\nu_n) - \mathcal{F}(\nu_{n+1}) \leq -\gamma^2(1 - 3\gamma L)D(\nu_n), \tag{6.26}$$

where L is the Lispchitz constant of ∇k. Combining (6.25) and (6.26) we finally get:

$$\begin{aligned} 2\gamma(\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\mu)) + \gamma^2(1 - 3\gamma L)D(\nu_n) \leq & W_2^2(\nu_n, \mu) - W_2^2(\nu_{n+1}, \mu) \\ & - 2\gamma K(\rho^n). \end{aligned}$$

and under the condition $\gamma \leq 1/(3L)$ we recover the desired result. □

We can now give the proof of the Theorem 52.

Proof of Theorem 52. Consider the Lyapunov function $L_j = j\gamma(\mathcal{F}(\nu_j) - \mathcal{F}(\mu)) + \frac{1}{2}W_2^2(\nu_j, \mu)$ for any iteration j. At iteration $j+1$, we have:

$$\begin{aligned} L_{j+1} =& j\gamma(\mathcal{F}(\nu_{j+1}) - \mathcal{F}(\mu)) + \gamma(\mathcal{F}(\nu_{j+1}) - \mathcal{F}(\mu)) + \frac{1}{2}W_2^2(\nu_{j+1}, \mu) \\ \leq & j\gamma(\mathcal{F}(\nu_{j+1}) - \mathcal{F}(\mu)) + \frac{1}{2}W_2^2(\nu_j, \mu) - \gamma K(\rho^j) \\ \leq & j\gamma(\mathcal{F}(\nu_j) - \mathcal{F}(\mu)) + \frac{1}{2}W_2^2(\nu_j, \mu) - \gamma K(\rho^j) \\ & - j\gamma^2(1 - \frac{3}{2}\gamma L) \int \|\nabla f_{\mu,\nu_j}(X)\|^2 d\nu_j \\ \leq & L_j - \gamma K(\rho^j). \end{aligned}$$

where we used Proposition 61 and Proposition 50 successively for the two first inequalities. We thus get by telescopic summation:

$$L_n \leq L_0 - \gamma \sum_{j=0}^{n-1} K(\rho^j)$$

Let us denote $\bar{K}$ the average value of $(K(\rho^j))_{0 \leq j \leq n}$ over iterations up to n. We can

now write the final result:

$$\mathcal{F}(\nu_n) - \mathcal{F}(\mu) \leq \frac{W_2^2(\nu_0, \mu)}{2\gamma n} - \bar{K}$$

□

A .2.3 Lojasiewicz type inequalities

Given a probability distribution ν, the *weighted Sobolev semi-norm* is defined for all squared integrable functions f in $L_2(\nu)$ as $\|f\|_{\dot{H}(\nu)} = \left(\int \|\nabla f(x)\|^2 \,\mathrm{d}\nu(x)\right)^{\frac{1}{2}}$ with the convention $\|f\|_{\dot{H}(\nu)} = +\infty$ if f does not have a square integrable gradient. The *Negative weighted Sobolev distance* $\|.\|_{\dot{H}^{-1}(\nu)}$ is then defined on distributions as the dual norm of $\|.\|_{\dot{H}(\nu)}$. For convenience, we recall the definition of $\|.\|_{\dot{H}^{-1}(\nu)}$:

Definition 8. *Let $\nu \in \mathcal{P}_2(\mathcal{X})$, with its corresponding weighted Sobolev semi-norm $\|.\|_{\dot{H}(\nu)}$. The weighted negative Sobolev distance $\|p - q\|_{\dot{H}^{-1}(\nu)}$ between any p and q in $\mathcal{P}_2(\mathcal{X})$ is defined as*

$$\|p - q\|_{\dot{H}^{-1}(\nu)} = \sup_{f \in L_2(\nu), \|f\|_{\dot{H}(\nu)} \leq 1} \left| \int f(x) \,\mathrm{d}p(x) - \int f(x) \,\mathrm{d}q(x) \right|$$

with possibly infinite values.

There are several possible choices for the set of test functions f. While it is often required that f vanishes at the boundary (see Mroueh et al. [2019]), we do not make such restriction and rather use the definition from Peyre [2018]. We refer to Shestakov and Shlapunov [2009] for more discussion on the relationship between different choices for the set of test functions.

We provide now a proof for Proposition 53.

Proof of Proposition 53. This proof follows simply from the definition of the negative Sobolev distance. Under Assumption **(A)**, the kernel has at most quadratic growth hence, for any $\mu, \nu \in \mathcal{P}_2(\mathcal{X})^2$, $f_{\mu,\nu} \in L_2(\nu)$. Consider $g = \|f_{\mu,\nu_t}\|^{-1}_{\dot{H}(\nu_t)} f_{\mu,\nu_t}$,

then $g \in L_2(\nu_t)$ and $\|g\|_{\dot{H}(\nu_t)} \leq 1$. Therefore, we directly have:

$$\left| \int g \, \mathrm{d}\nu_t - \int g \, \mathrm{d}\mu \right| \leq \|\nu_t - \mu\|_{\dot{H}^{-1}(\nu_t)} \tag{6.27}$$

Now, recall the definition of g, which implies that

$$\left| \int g \, \mathrm{d}\nu_t - \int g \, \mathrm{d}\mu \right| = \|\nabla f_{\mu,\nu_t}\|^{-1}_{L_2(\nu_t)} \left| \int f_{\mu,\nu_t} \, \mathrm{d}\nu_t - \int f_{\mu,\nu_t} \, \mathrm{d}\mu \right|. \tag{6.28}$$

Moreover, we have that $\int f_{\mu,\nu_t} \, \mathrm{d}\nu_t - \int f_{\mu,\nu_t} \, \mathrm{d}\mu = \|f_{\mu,\nu_t}\|^2_{\mathcal{H}}$, since f_{μ,ν_t} is the unnormalised witness function between ν_t and μ. Combining (6.27) and (6.28) we thus get the desired Lojasiewicz inequality on f_{μ,ν_t}:

$$\|f_{\mu,\nu_t}\|^2_{\mathcal{H}} \leq \|f_{\mu,\nu_t}\|_{\dot{H}(\nu_t)} \|\mu - \nu_t\|_{\dot{H}^{-1}(\nu_t)}$$

where $\|f_{\mu,\nu_t}\|_{\dot{H}(\nu_t)} = \|\nabla f_{\mu,\nu_t}\|_{L_2(\nu_t)}$ by definition. Then, using Proposition 48 and recalling by assumption that: $\|\mu - \nu_t\|^2_{\dot{H}^{-1}(\nu_t)} \leq C$, we have:

$$\dot{\mathcal{F}}(\nu_t) = -\|\nabla f_{\mu,\nu_t}\|^2_{L_2(\nu_t)} \leq -\frac{1}{C}\|f_{\mu,\nu_t}\|^4_{\mathcal{H}} = -\frac{4}{C}\mathcal{F}(\nu_t)^2 \tag{6.29}$$

It is clear that if $\mathcal{F}(\nu_0) > 0$ then $\mathcal{F}(\nu_t) > 0$ at all times by uniqueness of the solution. Hence, one can divide by $\mathcal{F}(\nu_t)^2$ and integrate the inequality from 0 to some time t. The desired inequality is obtained by simple calculations.

Then, using Proposition 50 and (6.29) where ν_t is replaced by ν_n it follows:

$$\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq -\gamma \left(1 - \frac{3}{2}L\gamma\right) \|\nabla f_{\mu,\nu_n}\|^2_{L_2(\nu_n)} \leq -\frac{4}{C}\gamma \left(1 - \frac{3}{2}\gamma L\right) \mathcal{F}(\nu_n)^2.$$

Dividing by both sides of the inequality by $\mathcal{F}(\nu_n)\mathcal{F}(\nu_{n+1})$ and recalling that $\mathcal{F}(\nu_{n+1}) \leq \mathcal{F}(\nu_n)$ it follows directly that:

$$\frac{1}{\mathcal{F}(\nu_n)} - \frac{1}{\mathcal{F}(\nu_{n+1})} \leq -\frac{4}{C}\gamma \left(1 - \frac{3}{2}\gamma L\right).$$

The proof is concluded by summing over n and rearranging the terms. □

A .2.4 Noisy Gradient flow of the MMD

Proof of Proposition 54. To simplify notations, we write $\mathcal{D}_{\beta_n}(\nu_n) = \int \|V(x + \beta_n u)\|^2 g(u)\,\mathrm{d}\nu_n\,\mathrm{d}u$ where $V := \nabla f_{\mu,\nu_n}$ and g is the density of a standard gaussian. The symbol $\otimes$ denotes the product of two independent probability distributions. Recall that a sample x_{n+1} from ν_{n+1} is obtained using $x_{n+1} = x_n - \gamma V(x_n + \beta_n u_n)$ where x_n is a sample from ν_n and u_n is a sample from a standard gaussian distribution that is independent from x_n. Moreover, by assumption β_n is a non-negative scalar satisfying:

$$8\lambda^2\beta_n^2\mathcal{F}(\nu_n) \leq \mathcal{D}_{\beta_n}(\nu_n) \tag{6.30}$$

Consider now the map $(x, u) \mapsto s_t(x) = x - \gamma t V(x + \beta_n u)$ for $0 \leq t \leq 1$, then ν_{n+1} is obtained as a push-forward of $\nu_n \otimes g$ by s_1: $\nu_{n+1} = (s_1)_\#(\nu_n \otimes g)$. Moreover, the curve $\rho_t = (s_t)_\#(\nu_n \otimes g)$ is a path from ν_n to ν_{n+1}. We know by Proposition 63 that $\nabla f_{\mu,\nu_n}$ is $2L$-Lipschitz, thus using $\phi(x, u) = -\gamma V(x + \beta_n u)$, $\psi(x, u) = x$ and $q = \nu_n \otimes g$ in Lemma 64 it follows that $\mathcal{F}(\rho_t)$ is differentiable in t with:

$$\dot{\mathcal{F}}(\rho_t) = \int \nabla f_{\mu,\rho_t}(s_t(x)).(-\gamma V(x + \beta_n u))g(u)\,\mathrm{d}\nu_n(x)\,\mathrm{d}u$$

Moreover, $\dot{\mathcal{F}}(\rho_0)$ is given by $\dot{\mathcal{F}}(\rho_0) = -\gamma \int V(x).V(x + \beta_n u)g(u)\,\mathrm{d}\nu_n(x)\,\mathrm{d}u$ and the following estimate holds:

$$|\dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_0)| \leq 3\gamma^2 Lt \int \|V(x + \beta_n u)\|^2 g(u)\,\mathrm{d}\nu_n(x)\,\mathrm{d}u = 3\gamma^2 Lt\mathcal{D}_{\beta_n}(\nu_n). \tag{6.31}$$

Using the absolute continuity of $\mathcal{F}(\rho_t)$, one has $\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) = \dot{\mathcal{F}}(\rho_0) + \int_0^1 \dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_0)\,\mathrm{d}t$. Combining with (6.31) and using the expression of $\dot{\mathcal{F}}(\rho_0)$, it follows that:

$$\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq -\gamma \int V(x).V(x + \beta_n u)g(u)\,\mathrm{d}\nu_n(x)\,\mathrm{d}u + \frac{3}{2}\gamma^2 L\mathcal{D}_{\beta_n}(\nu_n). \tag{6.32}$$

Adding and subtracting $\gamma\mathcal{D}_{\beta_n}(\nu_n)$ in (6.32) it follows directly that:

$$\begin{aligned}\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq & -\gamma(1-\frac{3}{2}\gamma L)\mathcal{D}_{\beta_n}(\nu_n) \\ & + \gamma\int (V(x+\beta_n u) - V(x)).V(x+\beta_n u)g(u)\,\mathrm{d}\nu_n(x)\,\mathrm{d}u\end{aligned} \tag{6.33}$$

We shall control now the last term in (6.33). Recall now that for all $1 \leq i \leq d$, $V_i(x) = \partial_i f_{\mu,\nu_n}(x) = \langle f_{\mu,\nu_n}, \partial_i k(x,.)\rangle$ where we used the reproducing property for the derivatives of f_{μ,ν_n} in $\mathcal{H}$ (see Section 1). Therefore, it follows by Cauchy-Schwartz in $\mathcal{H}$ and using Assumption (**D**):

$$\begin{aligned}\|V(x+\beta_n u) - V(x)\|^2 &\leq \|f_{\mu,\nu_n}\|_{\mathcal{H}}^2 \left(\sum_{i=1}^d \|\partial_i k(x+\beta_n u,.) - \partial_i k(x,.)\|_{\mathcal{H}}^2\right) \\ &\leq \lambda^2\beta_n^2\|f_{\mu,\nu_n}\|_{\mathcal{H}}^2\|u\|^2\end{aligned}$$

for all $x, u \in \mathcal{X}$. Now integrating both sides w.r.t. ν_n and g and recalling that g is a standard gaussian, we have:

$$\int \|V(x+\beta_n u) - V(x)\|^2 g(u)\,\mathrm{d}\nu_n(x)\,\mathrm{d}u \leq \lambda^2\beta_n^2\|f_{\mu,\nu_n}\|_{\mathcal{H}}^2$$

Getting back to (6.33) and applying Cauchy-Schwarz in $L_2(\nu_n \otimes g)$ it follows:

$$\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq -\gamma(1-\frac{3}{2}\gamma L)\mathcal{D}_{\beta_n}(\nu_n) + \gamma\lambda\beta_n\|f_{\mu,\nu_n}\|_{\mathcal{H}}\mathcal{D}_{\beta_n}^{\frac{1}{2}}(\nu_n)$$

It remains to notice that $\|f_{\mu,\nu_n}\|_{\mathcal{H}}^2 = 2\mathcal{F}(\nu_n)$ and that β_n satisfies (6.30) to get:

$$\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq -\frac{\gamma}{2}(1-\frac{3}{2}\gamma L)\mathcal{D}_{\beta_n}(\nu_n).$$

We introduce now $\Gamma = 4\gamma(1-\frac{3}{2}\gamma L)\lambda^2$ to simplify notation and prove the second inequality. Using (6.30) again in the above inequality we directly have: $\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq -\Gamma\beta_n^2\mathcal{F}(\nu_n)$. One can already deduce that $\Gamma\beta_n^2$ is necessarily smaller than

1. Hence, taking $\mathcal{F}(\nu_n)$ to the r.h. side and iterating over n it follows that:

$$\mathcal{F}(\nu_n) \leq \mathcal{F}(\nu_0) \prod_{i=0}^{n-1} (1 - \Gamma \beta_n^2)$$

Simply using that $1 - \Gamma\beta_n^2 \leq e^{-\Gamma\beta_n^2}$ leads to the desired upper-bound $\mathcal{F}(\nu_n) \leq \mathcal{F}(\nu_0) e^{-\Gamma \sum_{i=0}^{n-1} \beta_n^2}$. □

A .3 Asymptotic properties of the particle algorithms

Proof of Theorem 55. Let $(u_n^i)_{1\leq i\leq N}$ be i.i.d standard gaussian variables and $(x_0^i)_{1\leq i\leq N}$ i.i.d. samples from ν_0. We consider $(x_n^i)_{1\leq i\leq N}$ the particles obtained using the approximate scheme (6.17): $x_{n+1}^i = x_n^i - \gamma \nabla f_{\hat{\mu},\hat{\nu}_n}(x_n^i + \beta_n u_n^i)$ starting from $(x_0^i)_{1\leq i\leq N}$, where $\hat{\nu_n}$ is the empirical distribution of these N interacting particles. Similarly, we denote by $(\bar{x}_n^i)_{1\leq i\leq N}$ the particles obtained using the exact update equation (6.14): $\bar{x}_{n+1}^i = \bar{x}_n^i - \gamma \nabla f_{\mu,\nu_n}(\bar{x}_n^i + \beta_n u_n^i)$ also starting from $(x_0^i)_{1\leq i\leq N}$. By definition of ν_n we have that $(\bar{x}_n^i)_{1\leq i\leq N}$ are i.i.d. samples drawn from ν_n with empirical distribution denoted by $\bar{\nu}_n$. We will control the expected error c_n defined as $c_n^2 = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\|x_n^i - \bar{x}_n^i\|^2\right]$. By recursion, we have:

$$\begin{aligned}
c_{n+1} =& \frac{1}{\sqrt{N}} \left(\sum_{i=1}^N \mathbb{E}\left[\left\| x_n^i - \bar{x}_n^i - \gamma \left(\nabla f_{\hat{\mu},\hat{\nu}_n}(x_n^i + \beta_n u_n^i) - \nabla f_{\mu,\nu_n}(\bar{x}_n^i + \beta_n u_n^i) \right) \right\|^2 \right] \right)^{\frac{1}{2}} \\
\leq & c_n + \frac{\gamma}{\sqrt{N}} \left[\sum_{i=1}^N \mathcal{E}_i \right]^{\frac{1}{2}} + \frac{\gamma}{\sqrt{N}} \left[\sum_{i=1}^N \mathcal{G}_i \right]^{\frac{1}{2}} \\
& + \frac{\gamma}{\sqrt{N}} \left(\sum_{i=1}^N \mathbb{E}\left[\left\| \nabla f_{\mu,\hat{\nu}_n}\left(x_n^i + \beta_n u_n^i\right) - \nabla f_{\mu,\bar{\nu}_n}\left(\bar{x}_n^i + \beta_n u_n^i\right) \right\|^2 \right] \right)^{\frac{1}{2}} \\
\leq & c_n + 2\gamma L \left(c_n + \mathbb{E}\left[W_2(\hat{\nu}_n, \bar{\nu}_n)^2 \right]^{\frac{1}{2}} \right) + \frac{\gamma}{\sqrt{N}} \left[\sum_{i=1}^N \mathcal{E}_i \right]^{\frac{1}{2}} + \frac{\gamma}{\sqrt{N}} \left[\sum_{i=1}^N \mathcal{G}_i \right]^{\frac{1}{2}}
\end{aligned}$$

where the second line follows from a simple triangular inequality and the last line is obtained recalling that $\nabla f_{\mu,\nu}(x)$ is jointly $2L$ Lipschitz in x and ν by Proposition 63. Here, $\mathcal{E}_i$ represents the error between $\bar{\nu}_n$ and ν_n while $\mathcal{G}_i$ represents the error between

$\hat{\mu}$ and μ and are given by:

$$\mathcal{E}_i = \mathbb{E}\left[\left\|\nabla f_{\mu,\bar{\nu}_n}(\bar{x}_n^i + \beta_n u_n^i) - \nabla f_{\mu,\nu_n}(\bar{x}_n^i + \beta_n u_n^i)\right\|^2\right]$$
$$\mathcal{G}_i = \mathbb{E}\left[\left\|\nabla f_{\hat{\mu},\hat{\nu}_n}(x_n^i + \beta_n u_n^i) - \nabla f_{\mu,\hat{\nu}_n}(x_n^i + \beta_n u_n^i)\right\|^2\right]$$

We will first control the error term $\mathcal{E}_i$. To simplify notations, we write $y^i = \bar{x}_n^i + \beta_n u_n^i$. Recalling the expression of $\nabla f_{\mu,\nu}$ from Proposition 63 and expanding the squared norm in $\mathcal{E}_i$, it follows:

$$\begin{aligned}
\mathcal{E}_i &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{j=1}^N \nabla k(y^i, \bar{x}_n^j) - \int \nabla k(y^i, x) d\nu_n(x)\right\|^2\right] \\
&= \frac{1}{N^2}\sum_{j=1}^N \mathbb{E}\left[\left\|\nabla k(y^i, \bar{x}_n^j) - \int \nabla k(y^i, x) d\nu_n(x)\right\|^2\right] \\
&\leq \frac{L^2}{N^2}\sum_{j=1}^N \mathbb{E}\left[\left\|\bar{x}_n^j - \int x d\nu_n(x)\right\|^2\right] = \frac{L^2}{N} var(\nu_n).
\end{aligned}$$

The second line is obtained using the independence of the auxiliary samples $(\bar{x}_n^i)_{1\leq i\leq N}$ and recalling that they are distributed according to ν_n. The last line uses the fact that $\nabla k(y, x)$ is L-Lipshitz in x by Assumption (A). To control the variance $var(\nu_n)$ we use Lemma 62 which implies that $var(\nu_n)^{\frac{1}{2}} \leq (B + var(\nu_0)^{\frac{1}{2}})e^{LT}$ for all $n \leq \frac{2T}{\gamma}$. For $\mathcal{G}_i$, it is sufficient to expand again the squared norm and recall that $\nabla k(y, x)$ is L-Lipschitz in x which then implies that $\mathcal{G}_i \leq \frac{L^2}{M} var(\mu)$. Finally, one can observe that $\mathbb{E}[W_2^2(\hat{\nu}_n, \bar{\nu}_n)] \leq \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\|x_n^i - \bar{x}_n^i\|^2\right] = c_n^2$, hence c_n satisfies the recursion:

$$c_{n+1} \leq (1 + 4\gamma L)c_n + \frac{\gamma L}{\sqrt{N}}(B + var(\nu_0)^{\frac{1}{2}})e^{2LT} + \frac{\gamma L}{\sqrt{M}} var(\mu).$$

Using Lemma 68 to solve the above inequality, it follows that:

$$c_n \leq \frac{1}{4}\left(\frac{1}{\sqrt{N}}(B + var(\nu_0)^{\frac{1}{2}})e^{2LT} + \frac{1}{\sqrt{M}} var(\mu)\right)(e^{4LT} - 1)$$

□

Lemma 62. *Consider an initial distribution ν_0 with finite variance, a sequence $(\beta_n)_{n\geq 0}$ of non-negative numbers bounded by $B < \infty$ and define the sequence of probability distributions ν_n of the process (6.14):*

$$x_{n+1} = x_n - \gamma \nabla f_{\mu,\nu_n}(x_n + \beta_n u_n) \qquad x_0 \sim \nu_0$$

where $(u_n)_{n\geq 0}$ are standard gaussian variables. Under Assumption (A), the variance of ν_n satisfies for all $T > 0$ and $n \leq \frac{T}{\gamma}$ the following inequality:

$$var(\nu_n)^{\frac{1}{2}} \leq (B + var(\nu_0)^{\frac{1}{2}})e^{2TL}$$

Proof. Let g be the density of a standard gaussian. Denote by (x, u) and (x', u') two independent samples from $\nu_n \otimes g$. The idea is to find a recursion from $var(\nu_n)$ to $var(\nu_{n+1})$:

$$\begin{aligned}
var(\nu_{n+1})^{\frac{1}{2}} &= \left(\mathbb{E}\left[\|x - \mathbb{E}[x'] - \gamma \nabla f_{\mu,\nu_n}(x + \beta_n u) + \gamma \mathbb{E}[\nabla f_{\mu,\nu_n}(x' + \beta_n u')]\|^2\right]\right)^{\frac{1}{2}} \\
&\leq var(\nu_n)^{\frac{1}{2}} + \gamma \left(\mathbb{E}\left[\|\nabla f_{\mu,\nu_n}(x + \beta_n u) - \mathbb{E}[\nabla f_{\mu,\nu_n}(x' + \beta_n u')]\|^2\right]\right)^{\frac{1}{2}} \\
&\leq var(\nu_n)^{\frac{1}{2}} + 2\gamma L \mathbb{E}_{\substack{x,x'\sim\nu_n \\ u,u'\sim g}}\left[\|x + \beta_n u - x' + \beta_n u'\|^2\right]^{\frac{1}{2}} \\
&\leq var(\nu_n)^{\frac{1}{2}} + 2\gamma L(var(\nu_n)^{\frac{1}{2}} + \beta_n)
\end{aligned}$$

The second and last lines are obtained using a triangular inequality while the third line uses that $\nabla f_{\mu,\nu_n}(x)$ is $2L$-Lipschitz in x by Proposition 63. Recalling that β_n is bounded by B it is easy to conclude using Lemma 68. □

A .4 Auxiliary results

Proposition 63. *Under Assumption (A), the unnormalised witness function $f_{\mu,\nu}$ between any probability distributions μ and ν in $\mathcal{P}_2(\mathcal{X})$ is differentiable and satisfies:*

$$\nabla f_{\mu,\nu}(z) = \int \nabla_1 k(z, x)\, \mathrm{d}\mu(x) - \int \nabla_1 k(z, x)\, \mathrm{d}\nu(x) \qquad \forall z \in \mathcal{X} \qquad (6.34)$$

where $z \mapsto \nabla_1 k(x,z)$ *denotes the gradient of* $z \mapsto k(x,z)$ *for a fixed* $x \in \mathcal{X}$. *Moreover, the map* $(z,\mu,\nu) \mapsto f_{\mu,\nu}(z)$ *is Lipschitz with:*

$$\|\nabla f_{\mu,\nu}(z) - \nabla f_{\mu',\nu'}(z')\| \leq 2L(\|z-z'\| + W_2(\mu,\mu') + W_2(\nu,\nu'))$$

Finally, each component of $\nabla f_{\mu,\nu}$ *belongs to* $\mathcal{H}$.

Proof. The expression of the unnormalised witness function is given in (6.1). To establish (6.34), we simply need to apply the differentiation lemma [Klenke, 2008, Theorem 6.28]. By Assumption (A), it follows that $(x,z) \mapsto \nabla_1 k(z,x)$ has at most a linear growth. Hence on any bounded neighborhood of z, $x \mapsto \|\nabla_1 k(z,x)\|$ is upper-bounded by an integrable function w.r.t. μ and ν. Therefore, the differentiation lemma applies and $\nabla f_{\mu,\nu}(z)$ is differentiable with gradient given by (6.34).

To prove the second statement, we will consider two optimal couplings: π_1 with marginals μ and μ' and π_2 with marginals ν and ν'. We use (6.34) to write:

$$\begin{aligned}
&\|\nabla f_{\mu,\nu}(z) - \nabla f_{\mu',\nu'}(z')\| \\
=&\|\mathbb{E}_{\pi_1}\left[\nabla_1 k(z,x) - \nabla_1 k(z',x')\right] - \mathbb{E}_{\pi_2}\left[\nabla_1 k(z,y) - \nabla_1 k(z',y')\right]\| \\
\leq&\mathbb{E}_{\pi_1}\left[\|\nabla_1 k(z,x) - \nabla_1 k(z',x')\|\right] + \mathbb{E}_{\pi_2}\left[\|\nabla_1 k(z,y) - \nabla_1 k(z',y')\|\right] \\
\leq&L\left(\|z-z'\| + \mathbb{E}_{\pi_1}[\|x-x'\|] + \|z-z'\| + \mathbb{E}_{\pi_2}[\|y-y'\|]\right) \\
\leq&L(2\|z-z'\| + W_2(\mu,\mu') + W_2(\nu,\nu'))
\end{aligned}$$

The second line is obtained by convexity while the third one uses Assumption (A) and finally the last line relies on π_1 and π_2 being optimal. The desired bound is obtained by further upper-bounding the last two terms by twice their amount. □

Lemma 64. *Let* U *be an open set,* q *a probability distribution in* $\mathcal{P}_2(\mathcal{X}\times\mathcal{U})$ *and* ψ *and* ϕ *two measurable maps from* $\mathcal{X}\times\mathcal{U}$ *to* $\mathcal{X}$ *which are square-integrable w.r.t* q. *Consider the path* ρ_t *from* $(\psi)_{\#}q$ *and* $(\psi+\phi)_{\#}q$ *given by:* $\rho_t = (\psi + t\phi)_{\#}q \quad \forall t \in [0,1]$. *Under Assumption (A),* $\mathcal{F}(\rho_t)$ *is differentiable in* t *with*

$$\dot{\mathcal{F}}(\rho_t) = \int \nabla f_{\mu,\rho_t}(\psi(x,u) + t\phi(x,u))\phi(x,u)\,\mathrm{d}q(x,u)$$

where f_{μ,ρ_t} is the unnormalised witness function between μ and ρ_t as defined in (6.1). Moreover:

$$\left|\dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_s)\right| \leq 3L\,|t-s| \int \|\phi(x,u)\|^2\, dq(x,u)$$

Proof. For simplicity, we write f_t instead of f_{μ,ρ_t} and denote by $s_t(x,u) = \psi(x,u) + t\phi(x,u)$ The function $h : t \mapsto k(s_t(x,u), s_t(x',u')) - k(s_t(x,u),z) - k(s_t(x',u'),z)$ is differentiable for all (x,u),(x',u') in $\mathcal{X}\times\mathcal{U}$ and $z \in \mathcal{X}$. Moreover, by Assumption **(A)**, a simple computation shows that for all $0 \leq t \leq 1$:

$$\begin{aligned}\left|\frac{\mathrm{d}h}{\mathrm{d}t}\right| \leq & L\left[(\|z-\phi(x,u)\| + \|\psi(x,u)\|)\,\|\phi(x',u')\|\right] \\ & + L\left[(\|z-\phi(x',u')\| + \|\psi(x',u')\|)\,\|\phi(x,u)\|\right]\end{aligned}$$

The right hand side of the above inequality is integrable when z, (x,u) and (x',u') are independent and such that $z \sim \mu$ and both (x,u) and (x',u') are distributed according to q. Therefore, by the differentiation lemma [Klenke, 2008, Theorem 6.28] it follows that $\mathcal{F}(\rho_t)$ is differentiable and:

$$\dot{\mathcal{F}}(\rho_t) = \mathbb{E}\left[(\nabla_1 k(s_t(x,u), s_t(x',u')) - \nabla_1 k(s_t(x,u),z)).\phi(x,u)\right].$$

By Proposition 63, we directly get $\dot{\mathcal{F}}(\rho_t) = \int \nabla f_{\mu,\rho_t}(\psi(x,u)+t\phi(x,u))\phi(x,u)\,\mathrm{d}q(x,u)$. We shall control now the difference $|\dot{F}(\rho_t) - \dot{\mathcal{F}}(\rho_{t'})|$ for $0 \leq t,t' \leq 1$. Using Assumption **(A)** and recalling that $s_t(x,u) - s_{t'}(x,u) = (t-t')\phi(x,u)$ a simple computation shows:

$$\begin{aligned}\left|\dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_{t'})\right| &\leq L\,|t-t'|\,\mathbb{E}\left[(2\|\phi(x,u)\| + \|\phi(x',u')\|)\,\|\phi(x,u)\|\right] \\ &\leq L|t-t'|(2\mathbb{E}\left[\|\phi(x,u)\|^2\right] + \mathbb{E}\left[\|\phi(x,u)\|\right]^2) \\ &\leq 3L|t-t'| \int \|\phi(x,u)\|^2\,\mathrm{d}q(x,u).\end{aligned}$$

which gives the desired upper-bound. □

We denote by $(x,y) \mapsto H_1 k(x,y)$ the Hessian of $x \mapsto k(x,y)$ for all $y \in \mathcal{X}$

and by $(x, y) \mapsto \nabla_1 \nabla_2 k(x, y)$ the upper cross-diagonal block of the hessian of $(x, y) \mapsto k(x, y)$.

Lemma 65. *Let q be a probability distribution in $\mathcal{P}_2(\mathcal{X} \times \mathcal{X})$ and ψ and ϕ two measurable maps from $\mathcal{X} \times \mathcal{X}$ to $\mathcal{X}$ which are square-integrable w.r.t q. Consider the path ρ_t from $(\psi)_\# q$ and $(\psi + \phi)_\# q$ given by: $\rho_t = (\psi + t\phi)_\# q \quad \forall t \in [0, 1]$. Under Assumptions (A) and (B), $\mathcal{F}(\rho_t)$ is twice differentiable in t with*

$$\begin{aligned}\ddot{\mathcal{F}}(\rho_t) =& \mathbb{E}\left[\phi(x, y)^T \nabla_1 \nabla_2 k(s_t(x, y), s_t(x', y'))\phi(x', y')\right] \\ &+ \mathbb{E}\left[\phi(x, y)^T (H_1 k(s_t(x, y), y'_t) - H_1 k(s_t(x, y), z))\phi(x, y)\right]\end{aligned}$$

where (x, y) and (x', y') are independent samples from q, z is a sample from μ and $s_t(x, y) = \psi(x, y) + t\phi(x, y)$. Moreover, if Assumption (C) also holds then:

$$\begin{aligned}\ddot{\mathcal{F}}(\rho_t) \geq& \mathbb{E}\left[\phi(x, y)^T \nabla_1 \nabla_2 k(s_t(x, y), s_t(x', y'))\phi(x', y')\right] \\ &- \sqrt{2}\lambda d \mathcal{F}(\rho_t)^{\frac{1}{2}} \mathbb{E}[\|\phi(x, y)\|^2]\end{aligned}$$

where we recall that $\mathcal{X} \subset \mathbb{R}^d$.

Proof. The first part is similar to Lemma 64. In fact we already know by Lemma 64 that $\dot{\mathcal{F}}(\rho_t)$ exists and is given by:

$$\dot{\mathcal{F}}(\rho_t) = \mathbb{E}\left[(\nabla_1 k(s_t(x, y), s_t(x', y')) - \nabla_1 k(s_t(x, y), z)).\phi(x, y)\right]$$

Define now the function $\Xi : t \mapsto (\nabla_1 k(s_t(x, y), s_t(x', y')) - \nabla_1 k(s_t(x, y), z)).\phi(x, y)$ which is differentiable for all (x, y),(x', y') in $\mathcal{X} \times \mathcal{X}$ and $z \in \mathcal{X}$ by Assumption (B). Moreover, its time derivative is given by:

$$\begin{aligned}\dot{\Xi} =& \phi(x', y')^T \nabla_2 \nabla_1 k(s_t(x, y), s_t(x', y'))\phi(x, y) \\ &+ \phi(x, y)^T (H_1 k(s_t(x, y), s_t(x', y')) - H_1 k(s_t(x, y), z))\phi(x, y)\end{aligned}$$

By Assumption (A) it follows in particular that $\nabla_2 \nabla_1 k$ and $H_1 k$ are bounded hence $|\dot{\Xi}|$ is upper-bounded by $(\|\phi(x, y)\| + \|\phi(x', u')\|)\|\phi(x, y)\|$ which is integrable.

Therefore, by the differentiation lemma [Klenke, 2008, Theorem 6.28] it follows that $\dot{\mathcal{F}}(\rho_t)$ is differentiable and $\ddot{\mathcal{F}}(\rho_t) = \mathbb{E}\left[\dot{\Xi}\right]$. We prove now the second statement. Bu the reproducing property, it is easy to see that the last term in the expression of $\dot{\Xi}$ can be written as:

$$\langle \phi(x,y)^T H_1 k(s_t(x,y),.)\phi(x,y), k(s_t(x',y'),.) - k(z,.)\rangle_{\mathcal{H}}$$

Now, taking the expectation w.r.t x' ,y' and z which can be exchanged with the inner-product in $\mathcal{H}$ since $(x',y',z) \mapsto k(s_t(x',y'),.) - k(z,.)$ is Bochner integrable [Retherford, 1978, Definition 1, Theorem 6] and recalling that such integral is given by f_{μ,ρ_t} one gets the following expression:

$$\langle \phi(x,y)^T H_1 k(s_t(x,y),.)\phi(x,y), f_{\mu,\rho_t}\rangle_{\mathcal{H}}$$

Using Cauchy-Schwartz and Assumption (C) it follows that:

$$|\left\langle \phi(x,y)^T H_1 k(s_t(x,y),.)\phi(x,y), f_{\mu,\rho_t}\right\rangle_{\mathcal{H}}| \leq \lambda d \|\phi(x,y)\|^2 \|f_{\mu,\rho_t}\|$$

We conclude using the expression of $\ddot{\mathcal{F}}(\rho_t)$ and recalling that $\mathcal{F}(\rho_t) = \frac{1}{2}\|f_{\mu,\rho_t}\|^2$. □

Lemma 66. *Assume that for any geodesic $(\rho_t)_{t\in[0,1]}$ between ρ_0 and ρ_1 in $\mathcal{P}(\mathcal{X})$ with velocity vectors $(V_t)_{t\in[0,1]}$ the following holds:*

$$\ddot{\mathcal{F}}(\rho_t) \geq \Lambda(\rho_t, V_t)$$

for some admissible functional Λ as defined in Definition 3, then:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\rho_0) + t\mathcal{F}(\rho_1) - \int_0^1 \Lambda(\rho_s, V_s) G(s,t) ds$$

with $G(s,t) = s(1-t)\mathbb{1}\{s \leq t\} + t(1-s)\mathbb{1}\{s \geq t\}$ for $0 \leq s,t \leq 1$.

Proof. This is a direct consequence of the general identity (Villani [2009], Proposition 16.2). Indeed, for any continuous function ϕ on $[0,1]$ with second derivative $\ddot{\phi}$

that is bounded below in distribution sense the following identity holds:

$$\phi(t) = (1-t)\phi(0) + t\phi(1) - \int_0^1 \ddot{\phi}(s)G(s,t)ds.$$

This holds a fortiori for $\mathcal{F}(\rho_t)$ since $\mathcal{F}$ is smooth. By assumption, we have that $\ddot{\mathcal{F}}(\rho_t) \geq \Lambda(\rho_t, V_t)$, hence, it follows that:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\rho_0) + t\mathcal{F}(\rho_1) - \int_0^1 \Lambda(\rho_s, V_s)G(s,t)ds.$$

□

Lemma 67. *[Mixture convexity] The functional $\mathcal{F}$ is mixture convex: for any probability distributions ν_1 and ν_2 and scalar $1 \leq \lambda \leq 1$:*

$$\mathcal{F}(\lambda\nu_1 + (1-\lambda)\nu_2) \leq \lambda\mathcal{F}(\nu_1) + (1-\lambda)\mathcal{F}(\nu_2)$$

Proof. Let ν and ν' be two probability distributions and $0 \leq \lambda \leq 1$. Expanding the RKHS norm in $\mathcal{F}$ it follows directly that:

$$\mathcal{F}(\lambda\nu + (1-\lambda)\nu') - \lambda\mathcal{F}(\nu) - (1-\lambda)\mathcal{F}(\nu') = -\frac{1}{2}\lambda(1-\lambda)MMD(\nu,\nu')^2 \leq 0.$$

which concludes the proof. □

Lemma 68. *[Discrete Gronwall lemma] Let $a_{n+1} \leq (1+\gamma A)a_n + b$ with $\gamma > 0$, $A > 0$, $b > 0$ and $a_0 = 0$, then:*

$$a_n \leq \frac{b}{\gamma A}(e^{n\gamma A} - 1).$$

Proof. Using the recursion, it is easy to see that for any $n > 0$:

$$a_n \leq (1+\gamma A)^n a_0 + b\left(\sum_{i=0}^{n-1}(1+\gamma A)^k\right)$$

One concludes using the identity $\sum_{i=0}^{n-1}(1+\gamma A)^k = \frac{1}{\gamma A}((1+\gamma A)^n - 1)$ and recalling that $(1+\gamma A)^n \leq e^{n\gamma A}$. □

B A simple example when Lojasiewicz holds

Consider a gaussian target distribution $\mu(x) = \mathcal{N}(a, \Sigma)$ and initial distribution $\nu_0 = \mathcal{N}(a_0, \Sigma_0)$. In this case it is sufficient to use a kernel that captures the first and second moments of the distribution. We simply consider a kernel of the form $k(x, y) = (x^\top y)^2 + x^\top y$. In this case, it is easy to see by simple computations that the following equation holds:

$$\dot{X}_t = -(\Sigma_t - \Sigma + a_t a_t^\top - aa^\top)X_t - (a_t - a), \qquad \forall t \geq 0 \tag{6.35}$$

Where a_t and Σ_t are the mean and covariance matrix of ν_t and satisfy the equations:

$$\dot{\Sigma}_t = -(S_t\Sigma_t + \Sigma_t S_t)$$
$$\dot{a}_t = -S_t a_t - (a_t - a).$$

Where we introduced $S_t = \Sigma_t - \Sigma + a_t a_t^\top - aa^\top$ for simplicity. (6.35) implies that ν_t is in fact a gaussian distribution since X_t is obtained by summing gaussian increments. The same conclusion can be reached by solving the corresponding continuity equation. Thus we will be only interested in the behavior of a_t and Σ_t. First we can express the squared MMD in terms of those parameters:

$$MMD^2(\mu, \nu_t) = \|S_t\|^2 + \|a_t - a\|^2. \tag{6.36}$$

Since a_t and Σ_t are obtained from the gradient flow of the MMD, it follows that $\|a_t - a\|^2$ and $\|S_t\|^2$ remain bounded. Moreover, the Negative Sobolev distance is obtained by solving a finite dimensional quadratic problem and can be simply written as:

$$D(\mu, \nu_t) = tr(Q_t\Sigma_t Q_t) + \|a_t - a\|^2$$

where Q_t is the unique solution of the Lyapounov equation:

$$\Sigma_t Q_t + Q_t \Sigma_t = \Sigma_t - \Sigma + (a_t - a)(a_t - a)^\top := G_t. \tag{6.37}$$

We first consider the one dimensional case, for which (6.37) has a particularly simple solution and allows to provide a closed form expression for the negative Sobolev distance:

$$Q_t = \frac{G_t}{2\Sigma_t}, \qquad D(\mu, \nu_t) = \frac{G_t^2}{4\Sigma_t} + (a_t - a)^2.$$

Recalling (6.36) and that $MMD^2(\mu, \nu_t)$ is bounded at all times by definition of ν_t, it follows that both G_t and $a_t - a$ are also bounded. Hence, it is easy to see that $D(\mu, \nu_t)$ will remain bounded iff Σ_t remains bounded away from 0. This analysis generalizes the higher dimensions using [Behr et al., 2018, Lemma 3.2 (iii)] which provides an expression for Q_t in terms of G_t and the singular value decomposition of $\Sigma_t = U_t D_t U_t^\top$:

$$Q_t = U_t \left(\left(\frac{1}{(D_t)_i + (D_t)_j} \right) \odot U_t^\top G_t U_t \right) U_t^\top.$$

Here, $\odot$ denotes the Hadamard product of matrices. It is easy to see from this expression that $D(\mu, \nu_t)$ will be bounded if all singular values $((D_t)_i)_{1 \leq i \leq d}$ of Σ_t remain bounded away from 0.

C Connection to Neural Networks optimization

In this sub-section we establish a formal connection between the MMD gradient flow defined in (6.5) and neural networks optimization. Such connection holds in the limit of infinitely many neurons and is based on the formulation in Rotskoff and Vanden-Eijnden [2018]. To remain consistent with the rest of this chapter, the parameters of a network will be denoted by $x \in \mathcal{X}$ while the input and outputs will be denoted as z and y. Given a neural network or any parametric function $(z, x) \mapsto \psi(z, x)$ with parameter $x \in \mathcal{X}$ and input data z we consider the supervised

learning problem:

$$\min_{(x_1,\dots,x_m)\in\mathcal{X}} \frac{1}{2}\mathbb{E}_{(y,z)\sim p}\left[\left\|y-\frac{1}{m}\sum_{i=1}^{m}\psi(z,x_i)\right\|^2\right] \tag{6.38}$$

where $(y,z)\sim p$ are samples from the data distribution and the regression function is an average of m different networks. The formulation in (6.38) includes any type of networks. Indeed, the averaged function can itself be seen as one network with augmented parameters $(x_1,...,x_m)$ and any network can be written as an average of sub-networks with potentially shared weights. In the limit $m\to\infty$, the average can be seen as an expectation over the parameters under some probability distribution ν. This leads to an expected network $\Psi(z,\nu)=\int\psi(z,x)\,\mathrm{d}\nu(x)$ and the optimization problem in (6.38) can be lifted to an optimization problem in $\mathcal{P}_2(\mathcal{X})$ the space of probability distributions:

$$\min_{\nu\in\mathcal{P}_2(\mathcal{X})}\mathcal{L}(\nu):=\frac{1}{2}\mathbb{E}_{(y,z)\sim p}\left[\left\|y-\int\psi(z,x)\,\mathrm{d}\nu(x)\right\|^2\right] \tag{6.39}$$

For convenience, we consider $\bar{\mathcal{L}}(\nu)$ the function obtained by subtracting the variance of y from $\mathcal{L}(\nu)$, i.e.: $\bar{\mathcal{L}}(\nu)=\mathcal{L}(\nu)-var(y)$. When the model is well specified, there exists $\mu\in\mathcal{P}_2(\mathcal{X})$ such that $\mathbb{E}_{y\sim\mathbb{P}(.|z)}[y]=\int\psi(z,x)\,\mathrm{d}\mu(x)$. In that case, the cost function $\bar{\mathcal{L}}$ matches the functional $\mathcal{F}$ defined in (6.3) for a particular choice of the kernel k. More generally, as soon as a global minimizer for (6.39) exists, Proposition 69 relates the two losses $\bar{\mathcal{L}}$ and $\mathcal{F}$.

Proposition 69. *Assuming a global minimizer of (6.39) is achieved by some $\mu\in\mathcal{P}_2(\mathcal{X})$, the following inequality holds for any $\nu\in\mathcal{P}_2(\mathcal{X})$:*

$$\left(\bar{\mathcal{L}}(\mu)^{\frac{1}{2}}+\mathcal{F}^{\frac{1}{2}}(\nu)\right)^2\geq\bar{\mathcal{L}}(\nu)\geq\mathcal{F}(\nu)+\bar{\mathcal{L}}(\mu) \tag{6.40}$$

where $\mathcal{F}(\nu)$ is defined by (6.3) with a kernel k constructed from the data as an

expected product of networks:

$$k(x, x') = \mathbb{E}_{z\sim\mathbb{P}}\left[\psi(z, x)^T\psi(z, x')\right] \tag{6.41}$$

Moreover, $\bar{\mathcal{L}} = \mathcal{F}$ iif $\bar{\mathcal{L}}(\mu) = 0$, which means that the model is well-specified.

The framing (6.40) implies that optimizing $\mathcal{F}$ can decrease $\mathcal{L}$ and vice-versa. Moreover, in the well specified case, optimizing $\mathcal{F}$ is equivalent to optimizing $\mathcal{L}$. Hence one can use the gradient flow of the MMD defined in (6.5) to solve (6.39). One particular setting when (6.39) is well-specified is the student-teacher problem as in Chizat and Bach [2018b]. In this case, a teacher network of the form $\Psi_T(z, \mu)$ produces a deterministic output $y = \Psi_T(z, \mu)$ given an input z while a student network $\Psi_S(z, \nu)$ tries to learn the mapping $z \mapsto \Psi_T(z, \mu)$ by minimizing (6.39). In practice μ and ν are given as empirical distributions on some particles $\Xi = (\Xi^1, ..., \Xi^M)$ and $X = (x^1, ..., x^N)$ with $\mu = \frac{1}{M}\sum_{j=1}^M \delta_{\Xi^j}$ and $\nu = \frac{1}{N}\sum_{i=1}^N \delta_{x^i}$. The particles $(x^i)_{1\leq i\leq N}$ are then optimized using gradient descent starting from an initial configuration $(x_0^i)_{1\leq i\leq N}$. This leads to the update equation:

$$x_{n+1}^i = x_n^i - \gamma\mathbb{E}_{z\sim p}\left[\left(\frac{1}{N}\sum_{j=1}^N \psi(z, x_n^j) - \frac{1}{M}\sum_{j=1}^M \psi(z, \Xi^j)\right)\nabla_{x_n^i}\psi(z, x_n^i)\right], \tag{6.42}$$

where $(x_n^i)_{1\leq i\leq N}$ are the particles at iteration n with empirical distribution ν_n. Here, the gradient is rescaled by the number of particles N. Re-arranging terms and recalling that $k(x, x') = \mathbb{E}_{z\sim p}[\psi(z, x)^T\psi(z, x')]$, equation (6.42) becomes:

$$x_{n+1}^i = x_n^i - \gamma\nabla f_{\mu,\nu_n}(x_n^i).$$

with $\nabla f_{\mu,\nu_n}(x_n^i) = \left(\frac{1}{N}\sum_{j=1}^N \nabla_2 k(x_n^j, x_n^i) - \frac{1}{M}\sum_{j=1}^M \nabla_2 k(\Xi^j, x_n^i)\right)$. The above equation is a discretized version of the gradient flow of the MMD defined in (6.5). Such discretization is obtained from (6.17) by setting the noise level β_n to 0. Hence, in the limit when $N \to \infty$ and $\gamma \to 0$, one recovers the gradient flow defined in (6.9). In general the kernel k is intractable and can be approximated using n_b samples

$(z_1, ..., z_{n_b})$ from the data distribution: $\hat{k}(x, x') = \frac{1}{n_b} \sum_{b=1}^{n_b} \psi(z_b, x)^T \psi(z_b, x')$. This finally leads to an approximate update:

$$x_{n+1}^i = x_n^i - \gamma \nabla \hat{f}_{\mu,\nu_n}(x_n^i).$$

where $\nabla \hat{f}_{\mu,\nu_n}$ is given by:

$$\nabla \hat{f}_{\mu,\nu_n}(x_n^i) = \frac{1}{n_b} \sum_{b=1}^{n_b} \left(\frac{1}{N} \sum_{j=1}^{N} \psi(z_b, x_n^j) - \frac{1}{M} \sum_{j=1}^{M} \psi(z_b, \Xi^j) \right) \nabla_{x_n^i} \psi(z_b, x_n^i)).$$

We provide now a proof for Proposition 69:

Proof of Proposition 69. Let $\Psi(z, \nu) = \int \psi(z, x) \, \mathrm{d}\nu(x)$. By (6.41), we have: $k(x, x') = \int_z \psi(z, x)^T \psi(z, x') \, \mathrm{d}s(z)$ where s denotes the distribution of z. It is easy to see that $\mathcal{F}(\nu) = \frac{1}{2} \int \|\Psi(z, \nu) - \Psi(z, \mu)\|^2 \, \mathrm{d}s(z)$. Indeed expanding the square in the l.h.s and exchanging the order of integrations w.r.t p and $(\mu \otimes \nu)$ one gets $\mathcal{F}(\nu)$. Now, introducing $\Psi(z, \mu)$ in the expression of $\mathcal{L}(\nu)$, it follows by a simple calculation that:

$$\mathcal{L}(\nu) = \mathcal{L}(\mu) + \mathcal{F}(\nu) + \int \langle \Psi(z, \mu) - m(z), \Psi(z, \nu) - \Psi(z, \mu) \rangle \, \mathrm{d}p(z) \quad (6.43)$$

where $m(z)$ is the conditional mean of y, i.e.: $m(z) = \int y \, \mathrm{d}p(y|z)$. On the other hand we have that $2\mathcal{L}(\mu) = var(y) + \int \|\Psi(z, \mu) - m(z)\|^2 \, \mathrm{d}p(z)$, so that $\int \|\Psi(z, \mu) - m(z)\|^2 \, \mathrm{d}p(z) = 2\bar{\mathcal{L}}(\mu)$. Hence, using Cauchy-Schwartz for the last term in (6.43), one gets the upper-bound:

$$\mathcal{L}(\nu) \leq \mathcal{L}(\mu) + \mathcal{F}(\nu) + 2\bar{\mathcal{L}}(\mu)^{\frac{1}{2}} \mathcal{F}(\nu)^{\frac{1}{2}}.$$

This in turn gives an upper-bound on $\bar{\mathcal{L}}(\nu)$ after subtracting $var(y)/2$ on both sides of the inequality. To get the lower bound on $\bar{\mathcal{L}}$ one needs to use the global optimality condition of μ for $\mathcal{L}$ from [Chizat and Bach, 2018a, Proposition 3.1]. Indeed, for

any $0 < \epsilon \leq 1$ it is easy to see that:

$$\epsilon^{-1}(\mathcal{L}(\mu+\epsilon(\nu-\mu))-\mathcal{L}(\mu)) = \int \langle \Psi(z,\mu) - m(z), \Psi(z,\nu) - \Psi(z,\mu)\rangle \, \mathrm{d}p(z)+o(\epsilon).$$

Taking the limit $\epsilon \to 0$ and recalling that the l.h.s is always non-negative by optimality of μ, it follows that $\int \langle \Psi(z,\mu) - m(z), \Psi(z,\nu) - \Psi(z,\mu)\rangle \, \mathrm{d}p(z)$ must also be non-negative. Therefore, from (6.43) one gets that $\mathcal{L}(\nu) \geq \mathcal{L}(\mu) + \mathcal{F}(\nu)$. The final bound is obtained by subtracting $var(y)/2$ again from both sides of the inequality. □

Algorithm 4 Noisy gradient flow of the MMD

Input N, n_{iter}, β_0, γ
Output $(x^i_{n_{iter}})_{1\leq i\leq N}$
Initialize N particles from initial distribution ν_0 : $x^i_0 \overset{\text{i.i.d}}{\sim} \nu_0$
Initialize the noise level: $\beta = \beta_0$
for $n = 0, \ldots, n_{iter}$ **do**
  Sample M points from the target μ: $\{y^1, ..., y^M\}$.
  Sample N gaussians : $\{u^1_n, ..., u^N_n\}$
  for $i = 1, \ldots, N$ **do**
    Compute the noisy values: $\widetilde{x}^i_n = x^i_n + \beta_n u^i_n$
    Evaluate vector field: $\nabla f_{\hat{\mu},\hat{\nu}_n}(\widetilde{x}^i_n) = \frac{1}{N}\sum_{j=1}^{N} \nabla_2 k(x^j_n, \widetilde{x}^i_n) - \frac{1}{M}\sum_{m=1}^{M} \nabla_2 k(y^m, \widetilde{x}^i_n)$
    Update the particles: $x^i_{n+1} = x^i_n - \gamma \nabla f_{\hat{\mu},\hat{\nu}_n}(\widetilde{x}^i_n)$
  end for
  Update the noise level using an update rule h: $\beta_{n+1} = h(\beta_n, n)$.
end for

D Connection to Sobolev descent Mroueh et al. [2019]: The equilibrium condition

We discuss here the equilibrium condition (6.10) and relate it to [Mroueh et al., 2019, Assumption A]. Recall that (6.10) is given by: $\int \|\nabla f_{\mu,\nu^*}(x)\|^2 \, \mathrm{d}\nu^*(x) = 0$. Under some mild assumptions on the kernel which are states in [Mroueh et al., 2019,

Algorithm 5 Noisy gradient flow of the MMD for student-teacher learning

Input N, n_{iter}, β_0, γ, n_b, $\Xi = (\Xi^j)_{1\leq j\leq M}$.
Output $(x^i_{n_{iter}})_{1\leq i\leq N}$.
Initialize N particles from initial distribution ν_0 : $x_0^i \overset{\text{i.i.d}}{\sim} \nu_0$.
Initialize the noise level: $\beta = \beta_0$.
for $n = 0, ..., n_{iter}$ **do**
  Sample minibatch of n_b data points: $\{z^1, ..., z^{n_b}\}$.
  for $b = 1, ..., n_b$ **do**
    Compute teacher's output: $y_T^b = \frac{1}{M}\sum_{j=1}^M \psi(z^b, \Xi^j)$.
    Compute students's output: $y_S^b = \frac{1}{N}\sum_{i=1}^N \psi(z^b, x_n^i)$.
  end for
  Sample N gaussians : $\{u_n^1, ..., u_n^N\}$.
  for $i = 1, ..., N$ **do**
    Compute noisy particles: $\widetilde{x}_n^i = x_n^i + \beta_n u_n^i$
    Evaluate vector field: $\nabla \hat{f}_{\nu_\Xi, \nu_{X_n}}(\widetilde{x}_n^i) = \frac{1}{n_b}\sum_{b=1}^{n_b}(y_S^b - y_T^b)\nabla_{x_n^i}\psi(z^b, \widetilde{x}_n^i)$
    Update particle i: $x_{n+1}^i = x_n^i - \gamma \nabla \hat{f}_{\nu_\Xi, \nu_{X_n}}(\widetilde{x}_n^i)$
  end for
  Update the noise level using an update rule h: $\beta_{n+1} = h(\beta_n, n)$.
end for

Appendix C.1] it is possible to write (6.10) as:

$$\int \|\nabla f_{\mu,\nu^*}(x)\|^2 \, \mathrm{d}\nu^*(x) = \langle f_{\mu,\nu^*}, D_{\nu^*} f_{\mu,\nu^*}\rangle_{\mathcal{H}} = 0$$

where D_{ν^*} is a Hilbert-Schmidt operator given by:

$$D_{\nu^*} = \int \sum_{i=1}^d \partial_i k(x, .) \otimes \partial_i k(x, .) \, \mathrm{d}\nu^*(x)$$

Hence (6.10) is equivalent to say that f_{μ,ν^*} belongs to the null space of D_{ν^*}. In [Mroueh et al., 2019, Theorem 2], a similar equilibrium condition is derived by considering the time derivative of the MMD along the KSD gradient flow:

$$\frac{1}{2}\frac{d}{dt} MMD^2(\mu, \nu_t) = -\lambda \langle f_{\mu,\nu_t}, (\frac{1}{\lambda} I - (D_{\nu_t} + \lambda I)^{-1}) f_{\mu,\nu_t}\rangle_{\mathcal{H}}$$

The r.h.s is shown to be always negative and thus the MMD decreases in time. Hence, as t approaches ∞, the r.h.s tends to 0 since the MMD converges to some limit value l. This provides the equilibrium condition:

$$\lambda \langle f_{\mu,\nu^*}, (\frac{1}{\lambda} I - (D_{\nu^*} + \lambda I)^{-1}) f_{\mu,\nu *} \rangle_{\mathcal{H}} = 0$$

It is further shown in [Mroueh et al., 2019, Lemma 2] that the above equation is also equivalent to having f_{μ,ν^*} in the null space of D_{ν^*} in the case when D_{ν^*} has finite dimensions. We generalize this statement to infinite dimension in Proposition 70. In [Mroueh et al., 2019, Assumption A], it is simply assumed that if $f_{\mu,\nu^*} \neq 0$ then $D_{\nu^*} f_{\mu,\nu^*} \neq 0$ which exactly amounts to assuming that local optima which are not global don't exist.

Proposition 70.

$$\langle f_{\mu,\nu^*}, (\frac{1}{\lambda} I - (D_{\nu^*} + \lambda I)^{-1}) f_{\mu,\nu *} \rangle_{\mathcal{H}} = 0 \iff f_{\mu,\nu^*} \in null(D_{\nu^*})$$

Proof. This follows simply by recalling D_{ν^*} is a symmetric non-negative Hilbert-Schmidt operator it has therefore an eigen-decomposition of the form:

$$D_{\nu^*} = \sum_{i=1}^{\infty} \lambda_i e_i \otimes e_i$$

where e_i is an ortho-norrmal basis of $\mathcal{H}$ and λ_i are non-negative. Moreover, f_{μ,ν^*} can be decomposed in $(e_i)_{1 \leq i}$ in the form:

$$f_{\mu,\nu^*} = \sum_{i=0}^{\infty} \alpha_i e_i$$

where α_i is a squared integrable sequence. It follows that $\langle f_{\mu,\nu^*}, (\frac{1}{\lambda} I - (D_{\nu^*} + \lambda I)^{-1}) f_{\mu,\nu *} \rangle_{\mathcal{H}}$ can be written as:

$$\langle f_{\mu,\nu^*}, (\frac{1}{\lambda} I - (D_{\nu^*} + \lambda I)^{-1}) f_{\mu,\nu *} \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \lambda} \alpha_i^2$$

Hence, if $f_{\mu,\nu^*} \in null(D_{\nu^*})$ then $\langle f_{\mu,\nu^*}, D_{\nu^*} f_{\mu,\nu^*} \rangle_{\mathcal{H}} = 0$, so that $\sum_{i=1}^{\infty} \lambda_i \alpha_i^2 = 0$. Since λ_i are non-negative, this implies that $\lambda_i \alpha_i^2 = 0$ for all i. Therefore, it must be that $\langle f_{\mu,\nu^*}, (\frac{1}{\lambda} I - (D_{\nu^*} + \lambda I)^{-1}) f_{\mu,\nu*} \rangle_{\mathcal{H}} = 0$. Similarly, if $\langle f_{\mu,\nu^*}, (\frac{1}{\lambda} I - (D_{\nu^*} + \lambda I)^{-1}) f_{\mu,\nu*} \rangle_{\mathcal{H}} = 0$ then $\frac{\lambda_i \alpha_i^2}{\lambda_i + \lambda} = 0$ hence $\langle f_{\mu,\nu*}, D_{\nu^*} f_{\mu,\nu*} \rangle_{\mathcal{H}} = 0$. This means that $f_{\mu,\nu*}$ belongs to $null(D_{\nu^*})$. □

E Connection to the birth-death dynamics Rotskoff et al. [2019]

The Wasserstein gradient flow of $\mathcal{F}$ can be seen as the continuous-time limit of the so called minimizing movement scheme Ambrosio et al. [2008]. Such proximal scheme is defined using an initial distribution ν_0, a step-size τ, and an iterative update equation:

$$\nu_{n+1} \in \arg\min_{\nu} \mathcal{F}(\nu) + \frac{1}{2\tau} W_2^2(\nu, \nu_n). \tag{6.44}$$

In Ambrosio et al. [2008], it is shown that the continuity equation $\partial_t \nu_t = div(\nu_t \nabla f_{\mu,\nu_t})$ can be obtained as the limit when $\tau \to 0$ of (6.44) using suitable interpolations between the elements ν_n. In Rotskoff et al. [2019], a different transport equation that includes a birth-death term is considered:

$$\partial_t \nu_t = \beta div(\nu_t \nabla f_{\mu,\nu_t}) + \alpha(f_{\mu,\nu_t} - \int f_{\mu,\nu_t}(x)\, \mathrm{d}\nu_t(x))\nu_t \tag{6.45}$$

When $\beta = 0$ and $\alpha = 1$, it is shown formally in Rotskoff et al. [2019] that the above dynamics corresponds to the limit of a proximal scheme using the KL instead of the Wasserstein distance. For general β and α, (6.45) corresponds to the limit of a different proximal scheme where $W_2^2(\nu, \nu_n)$ is replaced by the Wasserstein-Fisher-Rao distance $d^2_{\alpha,\beta}(\nu, \nu_n)$ (see Chizat et al. [2015], Liero et al. [2016], Kondratyev et al. [2016]). $d^2_{\alpha,\beta}(\nu, \nu_n)$ is an interpolation between the squared Wasserstein distance ($\beta = 1$ and $\alpha = 0$) and the squared Fisher-Rao distance as defined in [Chizat et al., 2015, Definition 6] ($\beta = 0$ and $\alpha = 1$). Such scheme is consistent with the one proposed in Rotskoff et al. [2019] and which uses the KL. In fact, as we will

show later, both the KL and the Fisher-Rao distance have the same local behavior therefore both proximal schemes are expected to be equivalent in the limit when $\tau \to 0$.

Under (6.45), the time evolution of $\mathcal{F}$ is given by [Rotskoff et al., 2019, Proposition 3.1]:

$$\dot{\mathcal{F}}(\nu_t) = -\beta \int \|\nabla f_{\mu,\nu_t}\|^2 \, \mathrm{d}\nu_t(x) - \alpha \int \left| f_{\mu,\nu_t}(x) - \int f_{\mu,\nu_t}(x') \, \mathrm{d}\nu_t(x') \right|^2 \mathrm{d}\nu_t(x)$$

We would like to apply the same approach as in Section 3 .2 to provide a condition on the convergence of (6.45). Hence we first introduce an analogue to the Negative Sobolev distance in Definition 7 by duality:

$$D_\nu(p,q) = \sup_{\substack{g \in L_2(\nu) \\ \beta\|\nabla g\|^2_{L_2(\nu)} + \alpha\|g - \bar{g}\|^2_{L_2(\nu)} \leq 1}} \left| \int g(x) \, \mathrm{d}p(x) - \int g(x) \, \mathrm{d}q(x) \right|$$

where $\bar{g}$ is simply the expectation of g under ν. Such quantity defines a distance, since it is the dual of a semi-norm. Now using the particular structure of the MMD, we recall that $f_{\mu,\nu} \in L_2(\nu)$ and that $\beta\|\nabla f\|^2_{L_2(\nu)} + \alpha\|f - \bar{f}\|^2_{L_2(\nu)} < \infty$. Hence for a particular g of the form:

$$g = \frac{f_{\mu,\nu}}{\left(\beta\|\nabla f_{\mu,\nu}\|^2_{L_2(\nu)} + \alpha\|f_{\mu,\nu} - \bar{f}_{\mu,\nu}\|^2_{L_2(\nu)}\right)^{\frac{1}{2}}}$$

the following inequality holds:

$$D_\nu(\mu,\nu) \geq \frac{\left| \int f_{\mu,\nu} \, \mathrm{d}\nu(x) - \int f_{\mu,\nu} \, \mathrm{d}\mu(x) \right|}{\left(\beta\|\nabla f_{\mu,\nu}\|^2_{L_2(\nu)} + \alpha\|f_{\mu,\nu} - \bar{f}_{\mu,\nu}\|^2_{L_2(\nu)}\right)^{\frac{1}{2}}}.$$

But since $f_{\mu,\nu}$ is the unnormalised witness function between μ and ν we have that $2\mathcal{F}(\nu) = \left| \int f_{\mu,\nu} \, \mathrm{d}\nu(x) - \int f_{\mu,\nu} \, \mathrm{d}\mu(x) \right|$. Hence one can write that:

$$D^2_\nu(\mu,\nu) \left(\beta\|\nabla f_{\mu,\nu}\|^2_{L_2(\nu)} + \alpha\|f_{\mu,\nu} - \bar{f}_{\mu,\nu}\|^2_{L_2(\nu)}\right) \geq 4\mathcal{F}^2(\nu)$$

Now provided that $D_\nu^2(\mu, \nu_t)$ remains bounded at all time t by some constant $C > 0$ one can easily deduce a rate of convergence for $\mathcal{F}(\nu_t)$ just as in Proposition 53. In fact, in the case when $\beta = 1$ and $\alpha = 0$ one recovers Proposition 53. Another interesting case is when $\beta = 0$ and $\alpha = 1$. In this case, $D_\nu(p, q)$ is defined for p and q such that the difference $p - q$ is absolutely continuous w.r.t. ν. Moreover, $D_\nu(p, q)$ has the simple expression:

$$D_\nu(p, q) = \int \left(\frac{p - q}{\nu}(x)\right)^2 \mathrm{d}\nu(x)$$

where $\frac{p-q}{\nu}$ denotes the radon nikodym density of $p - q$ w.r.t. ν. More importantly, $D_\nu^2(\mu, \nu)$ is exactly equal to $\chi^2(\mu\|\nu)^{\frac{1}{2}}$. As we will show now, $(\chi^2)^{\frac{1}{2}}$ turns out to be a linearization of $\sqrt{2}KL^{\frac{1}{2}}$ and the Fisher-Rao distance.

Linearization of the KL and the Fisher-Rao distance. We first show the result for the KL. Given a probability distribution ν' that is absolutely continuous w.r.t to ν and for $0 < \epsilon < 1$ denote by $G(\epsilon) := KL(\nu\|(\nu + \epsilon(\nu' - \nu))$. It can be shown that $G(\epsilon) = \frac{1}{2}\chi^2(\nu'\|\nu)\epsilon^2 + o(\epsilon^2)$. To see this, one needs to perform a second order Taylor expansion of $G(\epsilon)$ at $\epsilon = 0$. Exchanging the derivatives and the integral, $\dot{G}(\epsilon)$ and $\ddot{G}(\epsilon)$ are both given by:

$$\dot{G}(\epsilon) = -\int \frac{\mu - \nu}{\nu + \epsilon(\mu - \nu)} \mathrm{d}\nu$$
$$\ddot{G}(\epsilon) = \int \frac{(\nu - \mu)^2}{(\nu + \epsilon(\mu - \nu))^2} \mathrm{d}\nu$$

Hence, we have for $\epsilon = 0$: $\dot{G}(0) = 0$ and $\ddot{G}(0) = \chi^2(\mu\|\nu)$. Therefore, it follows: $G(\epsilon) = \frac{1}{2}\chi^2(\mu\|\nu)\epsilon^2 + o(\epsilon^2)$, which means that

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[2KL\left(\nu\|\nu + \epsilon(\nu' - \nu)\right)\right]^{\frac{1}{2}} = \chi^2(\nu'\|\nu)^{\frac{1}{2}}.$$

The same approach can be used for the Fisher-Rao distance $d_{0,1}(\nu, \nu')$. From [Chizat et al., 2015, Theorem 3.1] we have that:

$$d_{0,1}^2(\nu, \nu') = 2 \int (\sqrt{\nu(x)} - \sqrt{\nu'(x)})^2 \mathrm{d}x$$

where ν and ν' are assumed to have a density w.r.t. Lebesgue measure. Using the exact same approach as for the KL one easily show that $\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[2d_{0,1}^2 \left(\nu \| \nu + \epsilon(\nu' - \nu)\right)\right]^{\frac{1}{2}} = \chi^2(\nu' \| \nu)^{\frac{1}{2}}$.

Linearization of the W_2. Similarly, it can be shown that the *Negative weighted Sobolev distance* is a linearization of the W_2 under suitable conditions. We recall here [Villani, 2003, Theorem 7.26] which relates the two quantities:

Theorem 71. *Let $\nu \in \mathcal{P}(\mathcal{X})$ be a probability measure with finite second moment, absolutely continuous w.r.t the Lebesgue measure and let $h \in L^\infty(\mathcal{X})$ with $\int h(x) \, \mathrm{d}\nu(x) = 0$. Then*

$$\|h\|_{\dot{H}^{-1}(\nu)} \leq \liminf_{\epsilon \to 0} \frac{1}{\epsilon} W_2(\nu, (1 + \epsilon h)\nu).$$

Theorem 71 implies that for any probability distribution ν' that has a bounded density w.r.t. to ν one has:

$$\|\nu' - \nu\|_{\dot{H}^{-1}(\nu)} \leq \liminf_{\epsilon \to 0} \frac{1}{\epsilon} W_2(\nu, \nu + \epsilon(\nu' - \nu)).$$

To get the converse inequality, one needs to assume that the support of ν is $\mathcal{X}$. Proposition 72 provides such inequality and uses techniques from Peyre [2018].

Proposition 72. *Let $\nu \in \mathcal{P}(\mathcal{X})$ be a probability measure with finite second moment, absolutely continuous w.r.t the Lebesgue measure with support equal to $\mathcal{X}$ and let $h \in L^\infty(\mathcal{X})$ with $\int h(x) \, \mathrm{d}\nu(x) = 0$ and $1 + h \geq 0$. Then*

$$\limsup_{\epsilon \to 0} \frac{1}{\epsilon} W_2(\nu, (1 + \epsilon h)\nu) \leq \|h\|_{\dot{H}^{-1}(\nu)}$$

Proof. Consider the elliptic equation: $\nu h + div(\nu \nabla F) = 0$ with Neumann boundary condition on $\partial \mathcal{X}$. Such equation admits a unique solution F in $\dot{H}(\nu)$ up to a constant since ν is supported on all of $\mathcal{X}$ (see [Otto and Villani, 2000, Section 7 (Linearizations)]). Moreover, we have that $\int F(x)h(x) \, \mathrm{d}\nu(x) = \int \|\nabla F(x)\|^2 \, \mathrm{d}\nu(x)$ which implies that $\|h\|_{\dot{H}^{-1}(\nu)} \geq \|F\|_{\dot{H}(\nu)}$. Now consider the path: $s_u = (1 + u\epsilon h)\nu$ for $u \in [0, 1]$. s_u is a probability distribution for all $u \in [0, 1]$ with $s_0 = \nu$ and

$s_1 = (1 + \epsilon h)\nu$. It is easy to see that s_u satisfies the continuity equation:

$$\partial_u s_u + div(s_u V_u) = 0$$

with $V_u = \frac{\epsilon \nabla F}{1+u\epsilon h}$. Indeed, for any smooth test function f one has:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}u}\int f(x)\,\mathrm{d}s_u(x) &= \epsilon \int f(x)h(x)\,\mathrm{d}\nu(x)\\ &= \epsilon \int \nabla f(x).\nabla F(x)\,\mathrm{d}\nu(x)\\ &= \int \nabla f(x).V_u(x)\,\mathrm{d}s_u(x).\end{aligned}$$

We used the definition of F for the second equality and that ν admits a density w.r.t. to s_u provided that ϵ is small enough. Such density is given by $1/(1 + u\epsilon h)$ and is positive and bounded when $\epsilon \leq \frac{1}{2\|h\|_\infty}$. Now, using the Benamou-Brenier formula for $W_2(\nu, (1 + \epsilon h)\nu)$ one has in particular that:

$$W_2(\nu, (1 + \epsilon h)\nu) \leq \int \|V_u\|_{L^2(s_u)}\,\mathrm{d}u$$

Using the expressions of V_u and s_u, one gets by simple computation:

$$\begin{aligned}W_2(\nu, (1 + \epsilon h)\nu) \leq& \epsilon \int \left(\int \frac{\|\nabla F(x)\|^2}{1 - u\epsilon + u\epsilon(h + 1)}\,\mathrm{d}\nu(x)\right)^{\frac{1}{2}}\mathrm{d}u\\ &\leq \epsilon \left(\int \|\nabla F(x)\|^2\,\mathrm{d}\nu(x)\right)^{\frac{1}{2}} \int_0^1 (1 - u\epsilon)^{-\frac{1}{2}}\,\mathrm{d}u.\end{aligned}$$

Finally, $\epsilon \int_0^1 (1 - u\epsilon)^{-\frac{1}{2}}\,\mathrm{d}u = 2(1 - \sqrt{1 - \epsilon}) \to 1$ when $\epsilon \to 0$, hence:

$$\limsup_{\epsilon \to 0} W_2(\nu, (1 + \epsilon h)) \leq \|F\|_{\dot{H}(\nu)} \leq \|h\|_{\dot{H}^{-1}(\nu)}.$$

□

Theorem 71 and Proposition 72 allow to conclude that $\lim_{\epsilon\to 0} \frac{1}{\epsilon} W_2(\nu, \nu + \epsilon(\nu' - \nu)) = \|\nu - \nu'\|_{\dot{H}^{-1}(\nu)}$ for any ν' that has a bounded density w.r.t. ν.

By analogy, one could wonder if D is also a linearization of the the Wasserstein-

Fisher-Rao distance. We leave such question for future work.

Chapter 7

Scalable Wasserstein natural gradient

Many machine learning problems can be expressed as the optimization of some cost functional over a parametric family of probability distributions. It is often beneficial to solve such optimization problems using natural gradient methods. These methods are invariant to the parametrization of the family, and thus can yield more effective optimization. Unfortunately, computing the natural gradient is challenging as it requires inverting a high dimensional matrix at each iteration. We propose a general framework to approximate the natural gradient for the Wasserstein metric, by leveraging a dual formulation of the metric restricted to a Reproducing Kernel Hilbert Space. Our approach leads to an estimator for gradient direction that can trade-off accuracy and computational cost, with theoretical guarantees. We verify its accuracy on simple examples, and show the advantage of using such an estimator in classification tasks on `Cifar10` and `Cifar100` empirically.

1 Introduction

The success of machine learning algorithms relies on the quality of an underlying optimization method. Many of the current state-of-the-art methods rely on variants of Stochastic Gradient Descent (SGD) such as AdaGrad [Duchi et al., 2011], RMSProp [Hinton et al., 2012], and Adam [Kingma and Ba, 2015]. While generally effective, the performance of such methods remains sensitive to the curvature of the

optimization objective. When the Hessian matrix of the objective at the optimum has a large condition number, the problem is said to have a pathological curvature [Martens, 2010, Sutskever et al., 2013]. In this case, the first-order optimization methods tend to have poor performance. Using adaptive step sizes can help when the principal directions of curvature are aligned with the coordinates of the vector parameters. Otherwise, an additional rotation of the basis is needed to achieve this alignment. One strategy is to find an alternative parametrization of the same model that has a better-behaved curvature and is thus easier to optimize with standard first-order optimization methods. Designing good network architectures [Simonyan and Zisserman, 2014, He et al., 2015] along with normalization techniques [LeCun et al., 2012, Ioffe and Szegedy, 2015, Salimans and Kingma, 2016] is often critical for the success of such optimization methods.

The natural gradient method [Amari, 1998] takes a related but different perspective. Rather than re-parametrizing the model, the natural gradient method tries to make the optimizer itself invariant to re-parameterizations by directly operating on the manifold of probability distributions. This requires endowing the parameter space with a suitable notion of proximity formalized by a metric. An important metric in this context is the Fisher information metric [Fisher and Russell, 1922, Rao, 1992], which induces the Fisher-Rao natural gradient [Amari, 1985]. Another important metric in probability space is the Wasserstein metric [Villani, 2009, Otto, 2001], which induces the Wasserstein natural gradient [Li and Montufar, 2018a,b, Li, 2018]; see similar formulations in Gaussian families [Malagò et al., 2018, Modin, 2017]. In spite of their numerous theoretical advantages, applying natural gradient methods is challenging in practice. Indeed, each parameter update requires inverting the metric tensor. This becomes infeasible for current deep learning models, which typically have millions of parameters. This has motivated research into finding efficient algorithms to estimate the natural gradient [Martens and Grosse, 2015, Grosse and Martens, 2016, George et al., 2018, Heskes, 2000, Bernacchia et al., 2018]. Such algorithms often address the case of the Fisher metric and either exploit a particular structure of the parametric family or rely on a low rank decomposition

of the information matrix. Recently, Li et al. [2019] proposed to estimate the metric based on a dual formulation and used this estimate in a proximal method. While this avoids explicitly computing the natural gradient, the proximal method also introduces an additional optimization problem to be solved at each update of the model's parameters. The quality of the solver will thus depend on the accuracy of this additional optimization.

In this work, we use the dual formulation of the metric to directly obtain a closed form expression of the natural gradient as a solution to a convex functional optimization problem. We focus on the Wasserstein metric as it has the advantage of being well defined even when the model doesn't admit a density. The expression remains valid for general metrics including the Fisher-Rao metric. We leverage recent work on Kernel methods [Sriperumbudur et al., 2017, Sutherland et al., 2018, Mroueh et al., 2019] to compute an estimate of the natural gradient by restricting the functional space appearing in the dual formulation to a Reproducing Kernel Hilbert Space. We demonstrate empirically the accuracy of our estimator on toy examples, and show how it can be effectively used to approximate the trajectory of the natural gradient descent algorithm. We also analyze the effect of the dimensionality of the model on the accuracy of the proposed estimator. Finally, we illustrate the benefits of our proposed estimator for solving classification problems when the model has an ill-conditioned parametrization.

This chapter is organized as follows. In Section 2 , after a brief description of natural gradients, we discuss Legendre duality of metrics, and provide details on the Wasserstein natural gradient. In Section 3 , we present our kernel estimator of the natural gradient. In Section 4 we present experiments to evaluate the accuracy of the proposed estimator and demonstrate its effectiveness in supervised learning tasks.

2 Natural Gradient Descent

We first briefly recall the natural gradient descent method in Section 2 .1, and its relation to metrics on probability distribution spaces in Section 2 .2. We next present Legendre dual formulations for metrics in Section 2 .3 where we highlight the

Fisher-Rao and Wasserstein metrics as important examples.

2 .1 General Formulation

It is often possible to formulate learning problems as the minimization of some cost functional $\rho \mapsto \mathcal{F}(\rho)$ over probability distributions ρ from a parametric model $\mathcal{P}_\Theta$. The set $\mathcal{P}_\Theta$ contains probability distributions defined on an open sample space $\Omega \subset \mathbb{R}^d$ and parametrized by some vector $\theta \in \Theta$, where Θ is an open subset of $\mathbb{R}^q$. The learning problem can thus be formalized as finding an optimal value θ^* that locally minimizes a loss function $\mathcal{L}(\theta) := \mathcal{F}(\rho_\theta)$ defined over the parameter space Θ. One convenient way to solve this problem approximately is by gradient descent, which uses the *Euclidean gradient* of $\mathcal{L}$ w.r.t. the parameter vector θ to produce a sequence of updates θ_t according to the following rule:

$$\theta_{t+1} = \theta_t - \gamma_t \nabla \mathcal{L}(\theta_t).$$

Here the step-size γ_t is a positive real number. The *Euclidean gradient* can be viewed as the direction in parameter space that leads to the highest decrease of some *linear model* $\mathcal{M}_t$ of the cost function $\mathcal{L}$ per unit of change of the parameter. More precisely, the *Euclidean gradient* is obtained as the solution of the optimization problem:

$$\nabla \mathcal{L}(\theta_t) = - \operatorname*{argmin}_{u \in \mathbb{R}^q} \ \mathcal{M}_t(u) + \frac{1}{2}\|u\|^2. \tag{7.1}$$

The linear model $\mathcal{M}_t$ is an approximation of the cost function $\mathcal{L}$ in the neighborhood of θ_t and is simply obtained by a first order expansion: $\mathcal{M}_t(u) = \mathcal{L}(\theta_t) + \nabla \mathcal{L}(\theta_t)^\top u$. The quadratic term $\|u\|^2$ penalizes the change in the parameter and ensures that the solution remains in the neighborhood where the linear model is still a good approximation of the cost function.

This particular choice of quadratic term is what defines the *Euclidean gradient* descent algorithm, which can often be efficiently implemented for neural network models using *back-propagation*. The performance of this algorithm is highly dependent on the parametrization of the model $\mathcal{P}_\Theta$, however [Martens, 2010, Sutskever et al., 2013]. To obtain an algorithm that is robust to parametrization, one can take

advantage of the structure of the cost function $\mathcal{L}(\theta)$ which is obtained as the composition of the functional $\mathcal{F}$ and the model $\theta \mapsto \rho_\theta$ and define a *generalized natural gradient* [Amari and Cichocki, 2010]. We first provide a conceptual description of the general approach to obtain such gradient. The starting point is to choose a divergence D between probability distributions and use it as a new penalization term:

$$-\operatorname*{argmin}_{u\in\mathbb{R}^q} \mathcal{M}_t(u) + \frac{1}{2}D(\rho_{\theta_t}, \rho_{\theta_t+u}). \tag{7.2}$$

Here, changes in the model are penalized directly in probability space rather than parameter space as in (7.1). In the limit of small u, the penalization term can be replaced by a quadratic term $u^\top G_D(\theta)u$ where $G_D(\theta)$ contains second order information about the model as measured by D. This leads to the following expression for the *generalized natural gradient* $\nabla^D\mathcal{L}(\theta_t)$ where the dependence in D is made explicit:

$$\nabla^D\mathcal{L}(\theta_t) := -\operatorname*{argmin}_{u\in\mathbb{R}^q} \mathcal{M}_t(u) + \frac{1}{2}u^\top G_D(\theta_t)u. \tag{7.3}$$

From (7.3), it is possible to express the *generalized natural gradient* by means of the *Euclidean gradient*: $\nabla^D\mathcal{L}(\theta_t) = G_D(\theta_t)^{-1}\nabla\mathcal{L}(\theta_t)$. The parameter updates are then obtained by the new update rule:

$$\theta_{t+1} = \theta_t - \gamma_t G_D(\theta_t)^{-1}\nabla\mathcal{L}(\theta_t). \tag{7.4}$$

Equation (7.4) leads to a descent algorithm which is invariant to parametrization in the continuous-time limit:

Proposition 73. *Let Ψ be an invertible and smoothly differentiable reparametrization $\psi = \Psi(\theta)$ and denote by $\bar{\mathcal{L}}(\psi) := \mathcal{L}(\Psi^{-1}(\psi))$. Consider the continuous-time natural gradient flows:*

$$\dot{\theta}_s = -\nabla^D_\theta\mathcal{L}(\theta_s), \qquad \dot{\psi}_s = -\nabla^D_\psi\bar{\mathcal{L}}(\psi_s), \qquad \psi_0 = \Psi(\theta_0)$$

Then ψ_s and θ_s are related by the equation $\psi_s = \Psi(\theta_s)$ at all times $s \geq 0$.

This result implies that an ill-conditioned parametrization of the model has little effect on the optimization when (7.4) is used. It is a consequence of the transformation properties of the natural gradient by change of parametrization: $\nabla^D_\psi \bar{\mathcal{L}}(\psi) = \nabla_\theta \Psi(\theta) \nabla^D_\theta \mathcal{L}(\theta)$ which holds in general for any covariant gradient. We provide a proof of Proposition 73 in Section B .1 in the particular the case when D is either Kullback-Leibler divergence F, or the squared Wasserstein-2 distance W using notions introduced later in Section 2 .3 and refer to Ollivier et al. [2011] for a detailed discussion.

The approach based on (7.2) for defining the generalized natural gradient is purely conceptual and can be formalized using the notion of metric tensor from differential geometry which allows for more generality. In Section 2 .2, we provide such formal definition in the case when D is either the Kullback-Leibler divergence F, or the squared Wasserstein-2 distance W.

2 .2 Information matrix via differential geometry

When D is the Kullback-Leibler divergence or relative entropy F, then (7.3) defines the *Fisher-Rao natural gradient* $\nabla^F \mathcal{L}(\theta)$ [Amari, 1985] and $G_F(\theta)$ is called the *Fisher information matrix*. $G_F(\theta)$ is well defined when the probability distributions in $\mathcal{P}_\Theta$ all have positive densities, and when some additional differentiability and integrability assumptions on ρ_θ are satisfied. In fact, it has an interpretation in Riemannian geometry as the pull-back of a metric tensor g^F defined over the set of probability distributions with positive densities and known as the *Fisher-Rao metric* (see Definition 5 in Section 3 ; see also Holbrook et al. [2017]):

Definition 9 (Fisher information matrix)**.** *Assume $\theta \mapsto \rho_\theta(x)$ is differentiable for all x on Ω and that $\int \frac{\|\nabla \rho_\theta(x)\|^2}{\rho_\theta(x)} \mathrm{d}x < \infty$. Then the Fisher information matrix is defined as the pull-back of the Fisher-Rao metric g^F:*

$$G_F(\theta)_{ij} = g^F_{\rho_\theta}(\partial_i \rho_\theta, \partial_j \rho_\theta) := \int f_i(x) f_j(x) \rho_\theta(x) dx,$$

where the functions f_i on Ω are given by: $f_i = \frac{\partial_i \rho_\theta}{\rho_\theta}$.

Definition 9 directly introduces G_F using the *Fisher-Rao metric* tensor which captures the infinitesimal behavior of the KL. This approach can be extended to any metric tensor g defined on a suitable space of probability distributions containing $\mathcal{P}_\Theta$. In particular, when D is the Wasserstein-2, the *Wasserstein information matrix* is obtained directly by means of the Wasserstein-2 metric tensor g^W [Otto and Villani, 2000, Lafferty and Wasserman, 2008] as proposed in Li and Montufar [2018a], Chen and Li [2018]:

Definition 10 (Wasserstein information matrix)**.** *The Wasserstein information matrix (WIM) is defined as the pull-back of the Wasserstein 2 metric g^W:*

$$G_W(\theta)_{ij} = g^W_{\rho_\theta}(\partial_i \rho_\theta, \partial_j \rho_\theta) := \int \phi_i(x)^\top \phi_j(x) \, \mathrm{d}\rho_\theta(x),$$

where ϕ_i are vector valued functions on $\Omega \subset \mathbb{R}^d$ that are solutions to the partial differential equations with Neumann boundary condition:

$$\partial_i \rho_\theta = -div(\rho_\theta \phi_i), \qquad \forall 1 \leq i \leq q.$$

Moreover, ϕ_i are required to be in the closure of the set of gradients of smooth and compactly supported functions in $L_2(\rho_\theta)^d$. In particular, when ρ_θ has a density, $\phi_i = \nabla_x f_i$, for some real valued function f_i on Ω.

The partial derivatives $\partial_i \rho_\theta$ should be understood in distribution sense, as discussed in more detail in Section 2 .3. This allows to define the *Wasserstein natural gradient* even when the model ρ_θ does not admit a density. Moreover, it allows for more generality than the conceptual approach based on (7.2) which would require performing a first order expansion of the Wasserstein distance in terms of its *linearized version* known as the *Negative Sobolev distance*. We provide more discussion of those two approaches and their differences in Section C . From now on, we will focus on the above two cases of the natural gradient $\nabla^D \mathcal{L}(\theta)$, namely $\nabla^F \mathcal{L}(\theta)$ and $\nabla^W \mathcal{L}(\theta)$. When the dimension of the parameter space is high, directly using equation (7.4) becomes impractical as it requires storing and inverting the matrix $G(\theta)$. In Section 2 .3 we will see how equation (7.3) can be exploited along

with Legendre duality to get an expression for the natural gradient that can be efficiently approximated using kernel methods.

2 .3 Legendre Duality for Metrics

In this section we provide an expression for the *natural gradient* defined in (7.3) as the solution of a saddle-point optimization problem. It exploits Legendre duality for metrics to express the quadratic term $u^\top G(\theta)u$ as a solution to a functional optimization problem over $C_c^\infty(\Omega)$, the set of smooth and compactly supported functions on Ω. The starting point is to extend the notion of gradient $\nabla\rho_\theta$ which appears in Definitions 9 and 10 to the distributional sense of Definition 11 below.

Definition 11. *Given a parametric family $\mathcal{P}_\Theta$ of probability distributions, we say that ρ_θ admits a distributional gradient at point θ if there exists a linear continuous map $\nabla\rho_\theta : C_c^\infty(\Omega) \to \mathbb{R}^q$ such that for any $f \in C_c^\infty(\Omega)$ and $u \in \mathbb{R}^q$:*

$$\int f(x)\,\mathrm{d}\rho_{\theta+\epsilon u}(x) - \int f(x)\,\mathrm{d}\rho_\theta(x) = \epsilon\nabla\rho_\theta(f)^\top u + \epsilon\delta(\epsilon, f, u),$$

where $\delta(\epsilon, f, u)$ depends on f and u and converges to 0 as ϵ approaches 0. $\nabla\rho_\theta$ is called the distributional gradient of ρ_θ at point θ.

When the distributions in $\mathcal{P}_\Theta$ have a density, written $x \mapsto \rho_\theta(x)$ by abuse of notation, that is differentiable w.r.t. θ and with a jointly continuous gradient in θ and x then $\nabla\rho_\theta(f)$ is simply given by $\int f(x)\nabla_\theta\rho_\theta(x)\,\mathrm{d}x$ as shown in Proposition 84 of Section B .1. In this case, the *Fisher-Rao natural gradient* admits a formulation as a saddle point solution involving $\nabla\rho_\theta$ and provided in Proposition 74 with a proof in Section B .1.

Proposition 74. *Under the same assumptions as in Definition 9, the Fisher information matrix admits the dual formulation:*

$$\frac{1}{2}u^\top G_F(\theta)u := \sup_{\substack{f\in C_c^\infty(\Omega)\\ \int f(x)\,\mathrm{d}\rho_\theta(x)=0}} \nabla\rho_\theta(f)^\top u - \frac{1}{2}\int f(x)^2\,\mathrm{d}\rho_\theta(x)\,\mathrm{d}x. \tag{7.5}$$

Moreover, defining $\mathcal{U}_\theta(f) = \nabla\mathcal{L}(\theta) + \nabla\rho_\theta(f)$, the Fisher-Rao natural gradient

satisfies:

$$\nabla^F \mathcal{L}(\theta) = -\operatorname*{argmin}_{u \in \mathbb{R}^q} \sup_{\substack{f \in C_c^\infty(\Omega) \\ \int f(x)\,\mathrm{d}\rho_\theta(x) = 0}} \mathcal{U}_\theta(f)^\top u - \frac{1}{2}\int f(x)^2\,\mathrm{d}\rho_\theta(x)\,\mathrm{d}x,$$

Another important case is when $\mathcal{P}_\Theta$ is defined as an implicit model. In this case, any sample x from a distribution ρ_θ in $\mathcal{P}_\Theta$ is obtained as $x = h_\theta(z)$, where z is a sample from a fixed latent distribution ν defined over a latent space $\mathcal{Z}$ and $(\theta, z) \mapsto h_\theta(z)$ is a deterministic function with values in Ω. This can be written in a more compact way as the push-forward of ν by the function h_θ:

$$\mathcal{P}_\Theta := \{\rho_\theta := (h_\theta)_\# \nu \mid \theta \in \Omega\}. \tag{7.6}$$

A different expression for $\nabla\rho_\theta$ is obtained in the case of implicit models when $\theta \mapsto h_\theta(z)$ is differentiable for ν-almost all z and ∇h_θ is square integrable under ν:

$$\nabla\rho_\theta(f) = \int \nabla h_\theta(z)^\top \nabla_x f(h_\theta(z))\,\mathrm{d}\nu(z). \tag{7.7}$$

Equation (7.7) is also known as the re-parametrization trick [Kingma et al., 2015] and allows to derive a dual formulation of the *Wasserstein natural gradient* in the case of implicit models. Proposition 75 below provides such formulation under mild assumptions stated in Section A .2 along with a proof in Section B .1.

Proposition 75. *Assume $\mathcal{P}_\Theta$ is defined by (7.6) such that $\nabla\rho_\theta$ is given by (7.7). Under Assumptions **(B)** and **(C)**, the Wasserstein information matrix satisfies:*

$$\frac{1}{2}u^\top G_W(\theta) u = \sup_{f \in C_c^\infty(\Omega)} \nabla\rho_\theta(f)^\top u - \frac{1}{2}\int \|\nabla_x f(x)\|^2\,\mathrm{d}\rho_\theta(x) \tag{7.8}$$

and the Wasserstein natural gradient satisfies:

$$\nabla^W \mathcal{L}(\theta) = -\operatorname*{argmin}_{u \in \mathbb{R}^q} \sup_{f \in C_c^\infty(\Omega)} \mathcal{A}_\theta(f, u). \tag{7.9}$$

where $\mathcal{A}_\theta(f,u)$ is a function from $\mathcal{C}_c^\infty(\Omega)\times\mathbb{R}^q$ defined as:

$$\mathcal{A}_\theta(f,u) := \mathcal{U}_\theta(f)^\top u - \frac{1}{2}\int \|\nabla_x f(x)\|^2 \,\mathrm{d}\rho_\theta(x),$$

with U_θ defined as in Proposition 74.

The similarity between the variational formulations provided in Propositions 74 and 75 is worth noting. A first difference however, is that Proposition 75 doesn't require the test functions f to have 0 mean under ρ_θ. This is due to the form of the objective in (7.8) which only depends on the gradient of f. More importantly, while (7.8) is well defined, the expression in (7.5) can be infinite when $\nabla\rho_\theta$ is given by (7.7). Indeed, if the ρ_θ doesn't admit a density, it is always possible to find an admissible function $f \in C_c^\infty(\Omega)$ with bounded second moment under ρ_θ but for which $\nabla\rho_\theta(f)$ is arbitrarily large. This is avoided in (7.8) since the quadratic term directly penalizes the gradient of functions instead. For similar reasons, the dual formulation of the Sobolev distance considered in Mroueh et al. [2019] can also be infinite in the case of implicit models as discussed in Section C although formally similar to (7.8). Nevertheless, a similar estimator as in Mroueh et al. [2019] can be considered using kernel methods which is the object of Section 3 .

3 Kernelized Wasserstein Natural Gradient

In this section we propose an estimator for the Wasserstein natural gradient using kernel methods and exploiting the formulation in (7.9). We restrict to the case of the Wasserstein natural gradient (WNG), denoted by $\nabla^W\mathcal{L}(\theta)$, as it is well defined for implicit models, but a similar approach can be used for the Fisher-Rao natural gradient in the case of models with densities. We first start by presenting the *kernelized Wasserstein natural gradient* (KWNG) in Section 3 .1, then we introduce an efficient estimator for KWNG in Section 3 .2. In Section 3 .4 we provide statistical guarantees and discuss practical considerations in Section 3 .3.

3 .1 General Formulation and Minimax Theorem

Consider a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ which is a Hilbert space endowed with an inner product $\langle ., . \rangle_{\mathcal{H}}$ along with its norm $\|.\|_{\mathcal{H}}$. $\mathcal{H}$ has the additional property that there exists a symmetric positive semi-definite kernel $k : \Omega \times \Omega \mapsto \mathbb{R}$ such that $k(x, .) \in \mathcal{H}$ for all $x \in \Omega$ and satisfying the *Reproducing property* for all functions f in $\mathcal{H}$:

$$f(x) = \langle f, k(x, .) \rangle_{\mathcal{H}}, \qquad \forall x \in \Omega. \tag{7.10}$$

The above property is central in all kernel methods as it allows to obtain closed form expressions for some class of functional optimization problems. In order to take advantage of such property for estimating the natural gradient, we consider a new saddle problem obtained by restricting (7.9) to functions in the RKHS $\mathcal{H}$ and adding some regularization terms:

$$\widetilde{\nabla}^W \mathcal{L}(\theta) := -\arg\min_{u \in \mathbb{R}^q} \sup_{f \in \mathcal{H}} \mathcal{A}_\theta(f, u) + \frac{1}{2}(\epsilon u^\top D(\theta) u - \lambda \|f\|^2_{\mathcal{H}}). \tag{7.11}$$

The *kernelized Wasserstein natural gradient* is obtained by solving (7.11) and is denoted by $\widetilde{\nabla}^W \mathcal{L}(\theta)$. Here, ϵ is a positive real numbers, λ is non-negative while $D(\theta)$ is a diagonal matrix in $\mathbb{R}^q$ with positive diagonal elements whose choice will be discussed in Section 3 .3. The first regularization term makes the problem strongly convex in u, while the second term makes the problem strongly concave in f when $\lambda > 0$. When $\lambda = 0$, the problem is still concave in f. This allows to use a version of the minimax theorem [Ekeland and Témam, 1999, Proposition 2.3, Chapter VI] to exchange the order of the supremum and minimum which also holds true when $\lambda = 0$. A new expression for the kernelized natural gradient is therefore obtained:

Proposition 76. *Assume that $\epsilon > 0$ and $\lambda > 0$, then the kernelized natural gradient is given by:*

$$\widetilde{\nabla}^W \mathcal{L}(\theta) = \frac{1}{\epsilon} D(\theta)^{-1} \mathcal{U}_\theta(f^*),$$

where f^* is the unique solution to the quadratic optimization problem:

$$\inf_{f\in\mathcal{H}} \mathcal{J}(f) := \int \|\nabla_x f(x)\|^2 \,\mathrm{d}\rho_\theta(x) + \frac{1}{\epsilon}\mathcal{U}_\theta(f)^\top D(\theta)^{-1}\mathcal{U}_\theta(f) + \lambda\|f\|^2_{\mathcal{H}}. \quad (7.12)$$

When $\lambda = 0$, f^* might not be well defined, still, we have: $\widetilde{\nabla}^W\mathcal{L}(\theta) = \lim_{j\to\infty}\frac{1}{\epsilon}D(\theta)^{-1}\mathcal{U}_\theta(f_j)$ for any limiting sequence of (7.12).

Proposition 76 allows to compute the kernelized natural gradient directly, provided that the functional optimization (7.12) can be solved. This circumvents the direct computation and inversion of the metric as suggested by (7.11). In Section 3 .2, we propose a method to efficiently compute an approximate solution to (7.12) using Nyström projections. We also show in Section 3 .4 that restricting the space of functions to $\mathcal{H}$ can still lead to a good approximation of the WNG provided that $\mathcal{H}$ enjoys some denseness properties.

3 .2 Nyström Methods for the Kerenalized Natural Gradient

We are interested now in finding an approximate solution to (7.12) which will allow to compute an estimator for the WNG using Proposition 76. Here we consider N samples $(Z_n)_{1\leq n\leq N}$ from the latent distribution ν which are used to produce N samples $(X_n)_{1\leq n\leq N}$ from ρ_θ using the map h_θ, i.e., $X_n = h_\theta(Z_n)$. We also assume we have access to an estimate of the *Euclidean gradient* $\nabla\mathcal{L}(\theta)$ which is denoted by $\widehat{\nabla\mathcal{L}(\theta)}$. This allows to compute an empirical version of the cost function in (7.12),

$$\hat{\mathcal{J}}(f) := \frac{1}{N}\sum_{n=1}^{N}\|\nabla_x f(X_n)\|^2 + \frac{1}{\epsilon}\widehat{\mathcal{U}_\theta}(f)^\top D(\theta)^{-1}\widehat{\mathcal{U}_\theta}(f) + \lambda\|f\|^2_{\mathcal{H}}, \quad (7.13)$$

where $\widehat{\mathcal{U}_\theta}(f)$ is given by: $\widehat{\mathcal{U}_\theta}(f) = \widehat{\nabla\mathcal{L}(\theta)} + \frac{1}{N}\sum_{n=1}^{N}\nabla h_\theta(Z_n))^\top\nabla_x f(X_n)$. (7.13) has a similar structure as the empirical version of the kernel Sobolev distance introduced in Mroueh et al. [2019], it is also similar to another functional arising in the context of *score* estimation for *infinite dimensional exponential families* as in Chapter 3 and [Sriperumbudur et al., 2017, Sutherland et al., 2018]. It can be shown using the generalized Representer Theorem [Schölkopf et al., 2001] that the optimal function minimizing (7.13) is a linear combination of functions of the form

$x \mapsto \partial_i k(X_n, x)$ with $1 \leq n \leq N$ and $1 \leq i \leq d$ and $\partial_i k(y, x)$ denotes the partial derivative of k w.r.t. y_i. This requires to solve a system of size $Nd \times Nd$ which can be prohibitive when both N and d are large. Nyström methods provide a way to improve such computational cost by further restricting the optimal solution to belong to a finite dimensional subspace $\mathcal{H}_M$ of $\mathcal{H}$ called the *Nyström subspace*. In the context of *score* estimation, Sutherland et al. [2018] proposed to use a subspace formed by linear combinations of the *basis functions* $x \mapsto \partial_i k(Y_m, x)$:

$$span\left\{x \mapsto \partial_i k(Y_m, x) \mid 1 \leq m \leq M; \quad 1 \leq i \leq d\right\}, \tag{7.14}$$

where $(Y_m)_{1 \leq m \leq M}$ are *basis points* drawn uniformly from $(X_n)_{1 \leq n \leq N}$ with $M \leq N$. This further reduces the computational cost when $M \ll N$ but still has a cubic dependence in the dimension d since all partial derivatives of the kernel are considered to construct (7.14). Here, we propose to randomly sample one component of $(\partial_i k(Y_m, .))_{1 \leq i \leq d}$ for each basis point Y_m. Hence, we consider M indices $(i_m)_{1 \leq m \leq M}$ uniformly drawn form $\{1, \ldots, d\}$ and define the *Nyström subspace* $\mathcal{H}_M$ to be:

$$\mathcal{H}_M := span\left\{x \mapsto \partial_{i_m} k(Y_m, x) | 1 \leq m \leq M\right\}.$$

An estimator for the *kernelized Wasserstein natural gradient* (KWNG) is then given by:

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon} D(\theta)^{-1} \widehat{\mathcal{U}_\theta}(\hat{f}^*), \qquad \hat{f}^* := \operatorname*{argmin}_{f \in \mathcal{H}_M} \hat{\mathcal{J}}(f). \tag{7.15}$$

By definition of the Nyström subspace $\mathcal{H}_M$, the optimal solution $\hat{f}^*$ is necessarily of the form: $\hat{f}^*(x) = \sum_{m=1}^M \alpha_m \partial_{i_m} k(Y_m, x)$, where the coefficients $(\alpha_m)_{1 \leq m \leq M}$ are obtained by solving a finite dimensional quadratic optimization problem. Proposition 77 provides a closed form expression for (7.17) in terms of the derivatives of the kernel collected in three matrices T, C and K. The matrices T and C belong to

$\mathbb{R}^{M\times Nd}$ and $\mathbb{R}^{M\times M}$ and are given by:

$$C_{m,(n,i)} = \partial_{i_m}\partial_{i+d}k(Y_m, X_n), \qquad K_{m,m'} = \partial_{i_m}\partial_{i_{m'}+d}k(Y_m, Y_{m'}) \tag{7.16}$$
$$T := \nabla\tau(\theta),$$

where T is expressed as the Jacobian of $\theta \mapsto \tau(\theta) \in \mathbb{R}^M$, i.e., $T := \nabla\tau(\theta)$, with

$$(\tau(\theta))_m = \frac{1}{N}\sum_{n=1}^{N} \partial_{i_m} k(Y_m, h_\theta(Z_n)).$$

Proposition 77. *The estimator in (7.15) is given by:*

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon}\left(D(\theta)^{-1} - \Lambda\right)\widehat{\nabla\mathcal{L}(\theta)}, \tag{7.17}$$

with S being a symmetric positive matrix expressed in terms of the matrices C, K and T defined in (7.16):

$$\Lambda := D(\theta)^{-1}T^\top \left(TD(\theta)^{-1}T^\top + \lambda\epsilon K + \frac{\epsilon}{N}CC^\top\right)^\dagger TD(\theta)^{-1}.$$

In (7.16), we used the notation $\partial_{i+d}k(y, x)$ for the partial derivative of k w.r.t. x_i. A proof of Proposition 77 is provided in Section B .2 and relies on the reproducing property (7.10) and its generalization for partial derivatives of functions. The estimator in Proposition 77 is in fact a low rank approximation of the natural gradient obtained from the dual representation of the metric (7.9). While low-rank approximations for the Fisher-Rao natural gradient were considered in the context of variational inference and for a Gaussian variational posterior [Mishkin et al., 2018], (7.17) can be applied as a plug-in estimator for any family $\mathcal{P}_\Theta$ obtained as an implicit model. We next discuss a numerically stable expression of (7.17), its computational cost and the choice of the damping term in Section 3 .3. We then provide asymptotic rates of convergence for (7.17) in Section 3 .4.

3 .3 Practical Considerations

Numerically stable expression. When $\lambda = 0$, the estimator in (7.17) has an additional structure which can be exploited to get more accurate solutions. By the chain rule, the matrix T admits a second expression of the form $T = CB$ where B is the Jacobian matrix of $(h_\theta(Z_n))_{1\leq n\leq N}$. Although this expression is impractical to compute in general, it suggests that C can be 'simplified'. This simplification can be achieved in practice by computing the SVD of $CC^\top = USU^T$ and pre-multiplying T by $S^\dagger U^T$. The resulting expression is given in Proposition 78 and falls into the category of *Ridgless estimators* (Liang and Rakhlin [2019]).

Proposition 78. *Consider an SVD decomposition of CC^T of the form $CC^\top = USU^T$, then (7.17) is equal to:*

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon}\left(D(\theta)^{-1} - \widetilde{\Lambda}\right)\widehat{\nabla \mathcal{L}(\theta)}, \tag{7.18}$$

where $\widetilde{\Lambda}$ is defined as:

$$\widetilde{\Lambda} := D(\theta)^{-1}\widetilde{T}^\top\left(\widetilde{T}D(\theta)^{-1}\widetilde{T}^\top + \frac{\epsilon}{N}P\right)^\dagger \widetilde{T}D(\theta)^{-1}$$

with $P := S^\dagger S$ and $\widetilde{T} := S^\dagger U^T T$.

Choice of damping term. So far, we only required $D(\theta)$ to be a diagonal matrix with positive coefficients. While a natural choice would be the identity matrix, this doesn't necessarily represent the best choice. As discussed by [Martens and Sutskever, 2012, Section 8.2], using the identity breaks the self-rescaling properties enjoyed by the natural gradient. Instead, we consider a scale-sensitive choice by setting $(D(\theta))_i = \|\widetilde{T}_{\cdot,i}\|$ where $\widetilde{T}$ is defined in Proposition 78. When the sample-size is limited, as it is often the case when N is the size of a mini-batch, larger values for ϵ might be required. That is to prevent the KWNG from over-estimating the step-size in low curvature directions. Indeed, these directions are rescaled by the inverse of the smallest eigenvalues of the information matrix which are harder to estimate accurately. To adjust ϵ dynamically during training, we use a variant of the

Levenberg-Marquardt heuristic as in Martens and Sutskever [2012] which seems to perform well in practice; see Section 4 .

Computational cost. The number of basis points M controls the computational cost of both (7.17) and (7.18) which is dominated by the cost of computing T and C, solving an $M \times M$ linear system and performing an SVD of CC^T in the case of (7.18). This gives an overall cost of $O(dNM^2 + qM^2 + M^3)$. In practice, M can be chosen to be small ($M \leq 20$) while N corresponds to the number of samples in a mini-batch. Hence, in a typical deep learning model, most of the computational cost is due to computing T as the typical number of parameters q is of the order of millions. In fact, T can be computed using automatic differentiation and would require performing M backward passes on the model to compute the gradient for each component of τ. Overall, the proposed estimator can be efficiently implemented and used for typical deep learning problems as shown in Section 4 .

Choice of the kernel. We found that using either a gaussian kernel or a rational quadratic kernel to work well in practice. We also propose a simple heuristic to adapt the bandwidth of those kernels to the data by setting it to $\sigma = \sigma_0 \sigma_{N,M}$, where $\sigma_{N,M}$ is equal to the average square distance between samples $(X_n)_{1 \leq n \leq N}$ and the basis points $(Y_m)_{1 \leq m \leq M}$ and σ_0 is fixed a priori. Another choice is the median heuristic Garreau et al. [2018].

3 .4 Theory

In this section we are interested in the behavior of the estimator in the limit of large N and M and when $\lambda > 0$; we leave the case when $\lambda = 0$ for future work. We work under Assumptions **(A)** to **(G)** in Section A .2 which state that Ω is a non-empty subset, k is continuously twice differentiable with bounded second derivatives, $\nabla h_\theta(z)$ has at most a linear growth in z and ν satisfies some standard moments conditions. Finally, we assume that the estimator of the euclidean gradient $\widehat{\nabla \mathcal{L}(\theta)}$ satisfies Chebychev's concentration inequality which is often the case in Machine learning problem as discussed in Remark 2 of Section A .2. We distinguish two cases: the *well-specified* case and the *miss-specified* case. In the *well-specified* case, the vector valued functions $(\phi_i)_{1 \leq i \leq q}$ involved in Definition 10 are assumed to be

gradients of functions in $\mathcal{H}$ and their smoothness is controlled by some parameter $\alpha \geq 0$ with worst case being $\alpha = 0$. Under this assumption, we obtain smoothness dependent convergence rates as shown in Theorem 86 of Section B .3 using techniques from Rudi et al. [2015], Sutherland et al. [2018]. Here, we will only focus on the *miss-specified* which relies on a weaker assumption:

Assumption 3. *There exists two constants $C > 0$ and $c \geq 0$ such that for all $\kappa > 0$ and all $1 \leq i \leq q$, there is a function f_i^κ satisfying:*

$$\|\phi_i - \nabla f_i^\kappa\|_{L_2(\rho_\theta)} \leq C\kappa, \qquad \|f_i^\kappa\|_{\mathcal{H}} \leq C\kappa^{-c}. \tag{7.19}$$

The left inequality in (7.19) represents the accuracy of the approximation of ϕ_i by gradients of functions in $\mathcal{H}$ while the right inequality represents the complexity of such approximation. Thus, the parameter c characterizes the difficulty of the problem: a higher value of c means that a more accurate approximation of ϕ_i comes at a higher cost in terms of its complexity. Theorem 79 provides convergences rates for the estimator in Proposition 77 under Assumption 3:

Theorem 79. *Let δ be such that $0 \leq \delta \leq 1$ and $b := \frac{1}{2+c}$. Under Assumption 3 and Assumptions (A) to (G) listed in Section A .2, for N large enough, $M \sim (dN^{\frac{1}{2b+1}} \log(N))$, $\lambda \sim N^{\frac{1}{2b+1}}$ and $\epsilon \lesssim N^{-\frac{b}{2b+1}}$, it holds with probability at least $1 - \delta$ that:*

$$\|\widehat{\nabla^W \mathcal{L}(\theta)} - \nabla^W \mathcal{L}(\theta)\|^2 = \mathcal{O}\Big(N^{-\frac{2}{4+c}}\Big).$$

A proof of Theorem 79 is provided in Section B .3. In the best case where $c = 0$, we recover a convergence rate of $\frac{1}{\sqrt{N}}$ as in the *well specified* case for the worst smoothness parameter value $\alpha = 0$. Hence, Theorem 79 is a consistent extension of the *well-specified* case. For harder problems where $c > 0$ more basis points are needed, with M required to be of order $dN \log(N)$ in the limit when $c \to \infty$ in which case the Nyström approximation loses its computational advantage.

4 Experiments

This section presents an empirical evaluation of (KWNG) based on (7.18).

4 .1 Synthetic Models

4 .1.1 Consistency of the estimator

Experimental setting To empirically assess the accuracy of KWNG, we consider three choices for the parametric model $\mathcal{P}_\Theta$: the multivariate normal model, the multivariate log-normal model and uniform distributions on hyper-spheres. All have the advantage that the WNG can be computed in closed form [Chen and Li, 2018, Malagò et al., 2018]. While the first models admit a density, the third one doesn't, hence the Fisher natural gradient is not defined in this case. While this choice of models is essential to obtain closed form expressions for WNG, the proposed estimator is agnostic to such choice of family. We also assume we have access to the exact Euclidean Gradient (EG) which is used to compute both of WNG and KWNG.

Results Figure 7.1 shows the evolution of the the relative error w.r.t. the sample-size N, the number of basis points M and the dimension d in the case of the hyper-sphere model. As expected from the consistency results provided in Section 3 .4, the relative error decreases as the samples size N increases. The behavior in the number of basis points M shows a clear threshold beyond which the estimator becomes consistent and where increasing M doesn't decrease the relative error anymore. This threshold increases with the dimension d as discussed in Section 3 .4. In practice, using the rule $M = \left\lfloor d\sqrt{N} \right\rfloor$ seems to be a good heuristic as shown in Figure 7.1 (a). All these observations persist in the case of the normal and log-normal model as shown in Figure 7.2. In addition, we report in Figure 7.3 the sensitivity to the choice of the bandwidth σ which shows a robustness of the estimator to a wide choice of σ.

4 .1.2 Optimization trajectory

We also compare the optimization trajectory obtained using KWNG with the trajectories of both the exact WNG and EG in a simple setting: $\mathcal{P}_\Theta$ is the multivariate normal family and the loss function $\mathcal{L}(\theta)$ is the squared Wasserstein 2 distance between ρ_θ and a fixed target distribution ρ_{θ^*}.

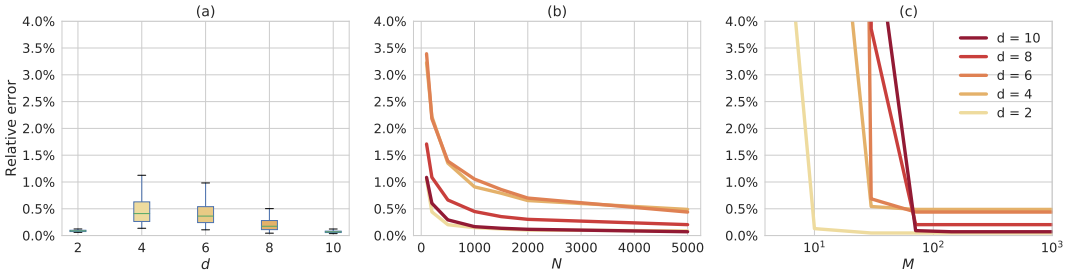


Figure 7.1: Relative error of KWNG averaged over 100 runs for varying dimension form $d = 1$ (yellow) to $d = 10$ (dark red) for the hyper-sphere model. (a): box-plot of the relative error as d increases while $N = 5000$ and $M = \left\lfloor d\sqrt{N} \right\rfloor$. (b) Relative error as the sample size N increases and $M = \left\lfloor d\sqrt{N} \right\rfloor$. (c): Relative error as M increases and $N = 5000$. A gaussian kernel is used with a fixed bandwidth $\sigma = 1$.

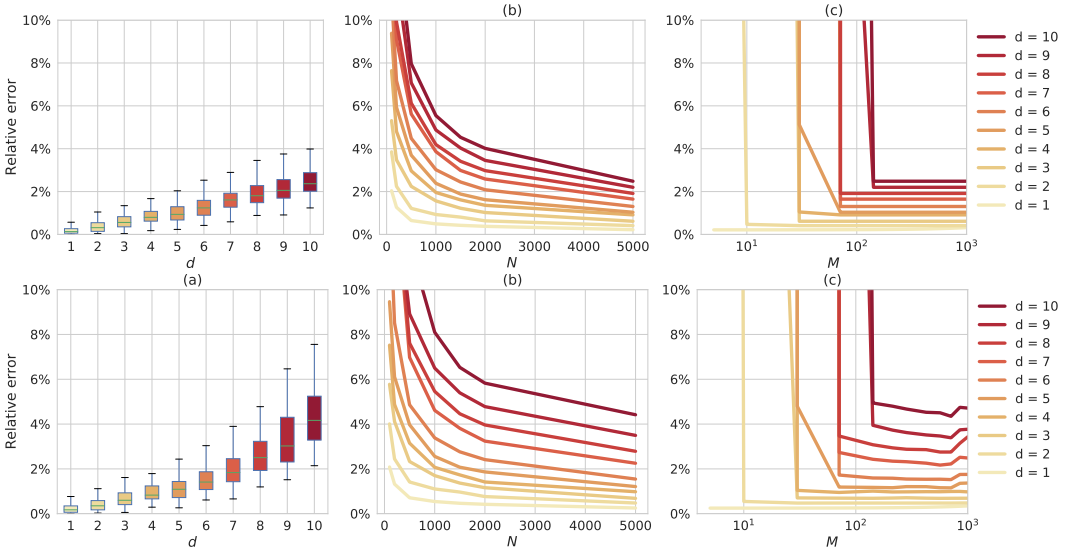


Figure 7.2: Evolution of the relative error of KWNG averaged over 100 runs for varying dimension form $d = 1$ (yellow) to $d = 10$ (dark red). For each run, a random value for the parameter θ and for the Euclidean gradient $\nabla\mathcal{L}(\theta)$ is sampled from a centered Gaussian with variance 0.1. In all cases, $\lambda = 0$ and $\epsilon = 10^{-5}$. Top row: multivariate normal model, bottom row: multivariate log-normal. Left (a): box-plot of the relative error as d increases with $N = 5000$ and the number of basis points is set to $M = \left\lfloor d\sqrt{N} \right\rfloor$. (b) Relative error as the sample size N increases and the number of basis points is set to $M = \left\lfloor d\sqrt{N} \right\rfloor$. Right (c): Relative error as M increases and N fixed to 5000.

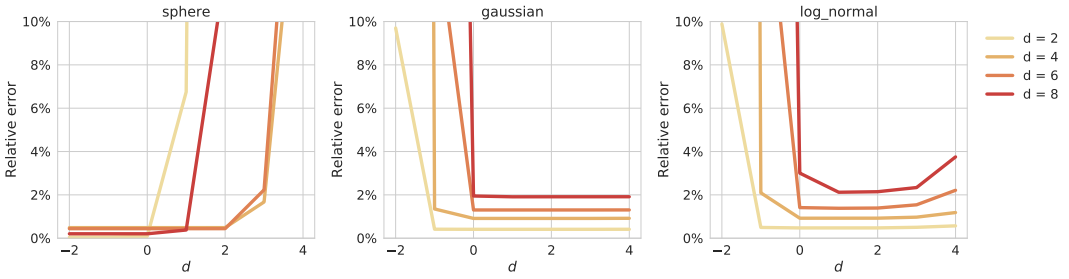


Figure 7.3: Relative error of the KWNG for varying bandwidth of the kernel. Results are averaged over 100 runs for varying dimension form $d = 1$ (yellow) to $d = 10$ (dark red). For each run, a random value for the parameter θ and for the Euclidean gradient $\nabla\mathcal{L}(\theta)$ is sampled from a centered Gaussian with variance 0.1. In all cases, $\lambda = \epsilon = 10^{-10}$. The sample size is fixed to $N = 5000$ and the number of basis points is set to $M = \left\lfloor d\sqrt{N} \right\rfloor$. Left: uniform distributions on a hyper-sphere, middle: multivariate normal, and right: multivariate log-normal.

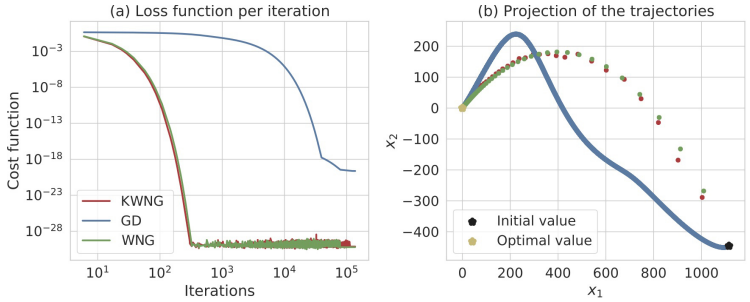


Figure 7.4: Left (a): Training error per iteration for KWNG, WNG, and EG. Right (b): projection of the sequence of updates obtained using KWNG, WNG and EG along the first two PCA directions of the WNG trajectory. The dimension of the sample space is fixed to $d = 10$. Exact valued for the gradient are used for EG and WNG. For KWNG, $N = 128$ samples and $M = 100$ basis points are used. The regularization parameters are set to: $\lambda = 0$ and $\epsilon = 10^{-10}$. An optimal step-size γ_t is used: $\gamma_t = 0.1$ for both KWNG and WNG while $\gamma_t = 0.0001$ for EG.

Figure 7.4 (a), shows the evolution of the loss function at every iteration. There is a clear advantage of using the WNG over EG as larger step-sizes are allowed leading to faster convergence. Moreover, KWNG maintains this properties while being agnostic to the choice of the model. Figure 7.4 (b) shows the projected dynamics of the three methods along the two PCA directions of the WNG trajectory with highest variance. The dynamics of WNG seems to be well approximated by the one obtained using KWNG.

4 .2 Approximate Invariance to Parametrization

4 .2.1 Experimental setting.

We illustrate now the approximate invariance to parametrization of the KWNG and show its benefits for training deep neural networks when the model is ill-conditioned. We consider a classification task on two datasets `Cifar10` and `Cifar100` with a Residual Network He et al. [2015].

To use the KWNG estimator, we view the input RGB image as a latent variable z with probability distribution ν and the output logits of the network $x := h_\theta(z)$ as a sample from the model distribution $\rho_\theta \in \mathcal{P}_\Theta$ where θ denotes the weights of the network. The loss function $\mathcal{L}$ is given by:

$$\mathcal{L}(\theta) := \int y(z)^\top \log(SM(Uh_\theta(z)))\, \mathrm{d}\nu(z),$$

where SM is the Softmax function, $y(z)$ denotes the one-hot vector representing the class of the image z and U is a fixed invertible diagonal matrix which controls how well the model is conditioned.

We consider two cases, the *Well-conditioned* case (WC) in which U is the identity and the *Ill-conditioned* case (IC) where U is chosen to have a condition number equal to 10^7.

We compare the performance of the proposed method with several variants of SGD: plain SGD, SGD + Momentum, and SGD + Momentum + Weight decay. We also compare with Adam Kingma and Ba [2015], KFAC optimizer [Martens and Grosse, 2015, Grosse and Martens, 2016] and eKFAC [George et al., 2018] which implements a fast approximation of the empirical Fisher Natural Gradient. We emphasize that gradient clipping by norm was used for all experiments and was crucial for a stable optimization using KWNG.

4 .2.2 Experimental details

Architecture. We use a residual network with one convolutional layer followed by 8 residual blocks and a final fully connected layer. Each residual block consists of two 3×3 convolutional layers each and ReLU nonlinearity. We use batch normalization

	Kernel size	Output shape
z		$32 \times 32 \times 3$
Conv	3×3	64
Residual block	$[3 \times 3] \times 2$	64
Residual block	$[3 \times 3] \times 2$	128
Residual block	$[3 \times 3] \times 2$	256
Residual block	$[3 \times 3] \times 2$	512
Linear	-	Number of classes

Table 7.1: Network architecture.

for all methods. Details of the intermediate output shapes and kernel size are provided in Table 7.1.

Hyper-parameters. For all methods, we used a batch-size of 128. The optimal step-size γ was selected in $\{10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ for each method. In the case of SGD with momentum, we used a momentum parameter of 0.9 and a weight decay of either 0 or 5×10^{-4}. For KFAC and EKFAC, we used a damping coefficient of 10^{-3} and a frequency of reparametrization of 100 updates. For KWGN we set $M = 5$ and $\lambda = 0$ while the initial value for ϵ is set to $\epsilon = 10^{-5}$ and is adjusted using an adaptive scheme based on the Levenberg-Marquardt dynamics as in [Martens and Grosse, 2015, Section 6.5]. More precisely, we use the following update equation for ϵ after every 5 iterations of the optimizer:

$$\begin{aligned} \epsilon &\leftarrow \omega\epsilon, && \text{if } r > \frac{3}{4} \\ \epsilon &\leftarrow \omega^{-1}\epsilon, && \text{if } r < \frac{1}{4}. \end{aligned}$$

Here, r is the reduction ratio:

$$r = \max_{t_{k-1} \leq t \leq t_k} \left(2 \frac{\mathcal{L}(\theta_t)) - \mathcal{L}(\theta_{t+1})}{\nabla^W \mathcal{L}(\theta)^\top \nabla \mathcal{L}(\theta)^\top} \right)$$

where $(t_k)_k$ are the times when the updates occur. and ω is the decay constant chosen to $\omega = 0.85$.

4 .2.3 Results.

Cifar10 Figure 7.5 shows the training and test accuracy at each epoch on `Cifar10` in both (WC) and (IC) cases. While all methods achieve a similar test accuracy in the (WC) case on both datasets, methods based on the Euclidean gradient seem to suffer a drastic drop in performance in the (IC) case. This doesn't happen for KWNG (red line) which achieves a similar test accuracy as in (WC) case. Moreover, a speed-up in convergence in number of iterations can be obtained by increasing the number of basis points M (brown line). The time cost is also in favor of KWNG as shown in Figure 7.6.

Cifar100 On `Cifar100`, KWNG is also less affected by the ill-conditioning, albeit to a lower extent. Indeed, the larger number of classes in `Cifar100` makes the estimation of KWNG harder as discussed in Section 4 .1. In this case, increasing the batch-size can substantially improve the training accuracy (pink line). Moreover, methods that are used to improve optimization using the Euclidean gradient can also be used for KWNG. For instance, using Momentum leads to an improved performance in the (WC) case (grey line).

Effect of the damping Interestingly, KFAC seems to also suffer a drop in performance in the (IC) case. This might result from the use of an isotropic damping term $D(\theta) = I$ which would be harmful in this case. We also observe a drop in performance when a different choice of damping is used for KWNG. More importantly, using only a diagonal pre-conditioning of the gradient doesn't match the performance of KWNG as shown in Figure 7.7.

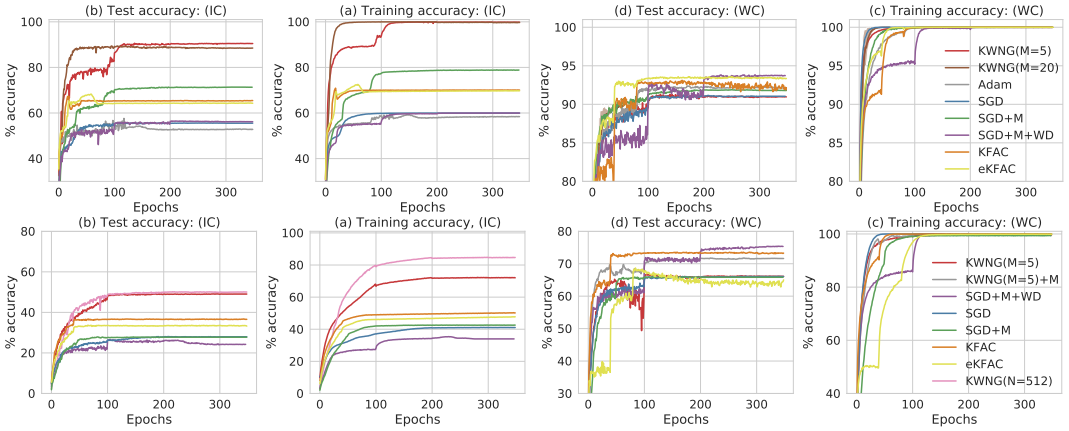


Figure 7.5: Test accuracy and Training accuracy for classification on `Cifar10` (top) and `Cifar100` (bottom) in both the ill-conditioned case (left side) and well-conditioned case (right side) for different optimization methods. on `Cifar10` Results are averaged over 5 independent runs except for KFAC and eKFAC.

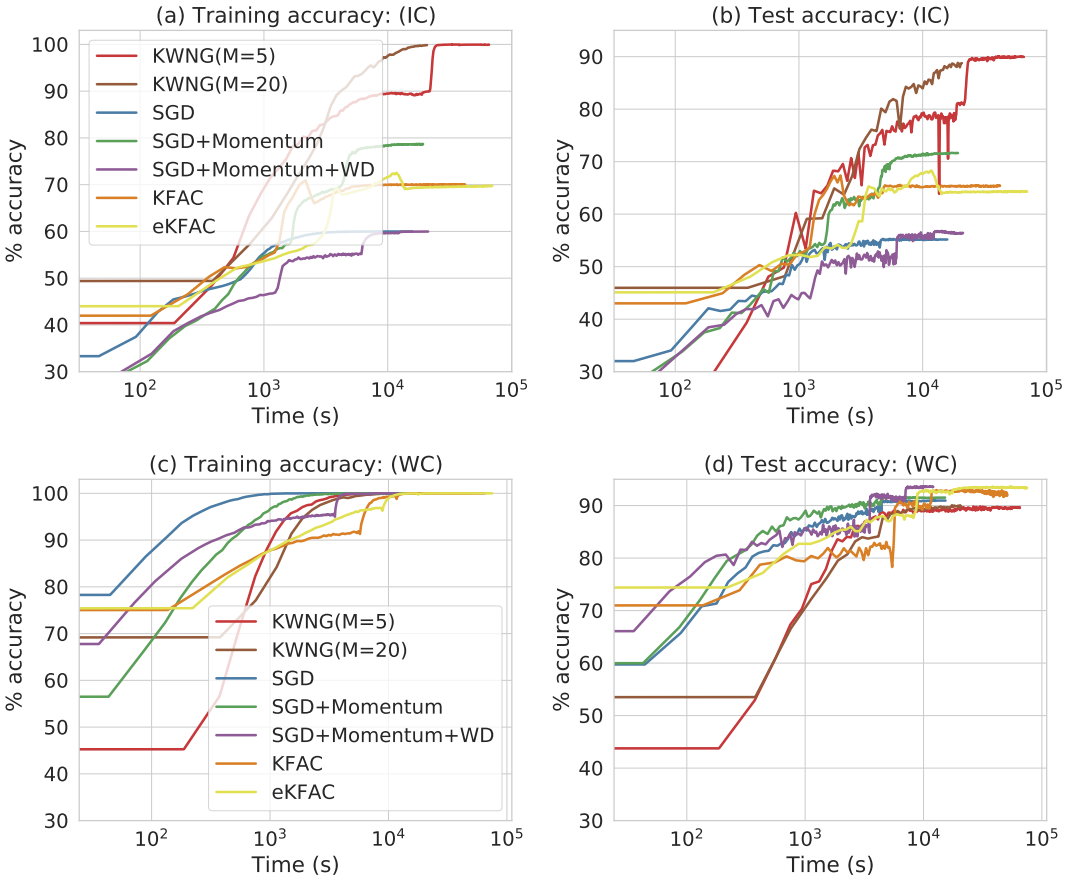


Figure 7.6: Training accuracy (left) and test accuracy (right) as a function of time for classification on `Cifar10` in both the ill-conditioned case (top) and well-conditioned case (bottom) for different optimization methods.

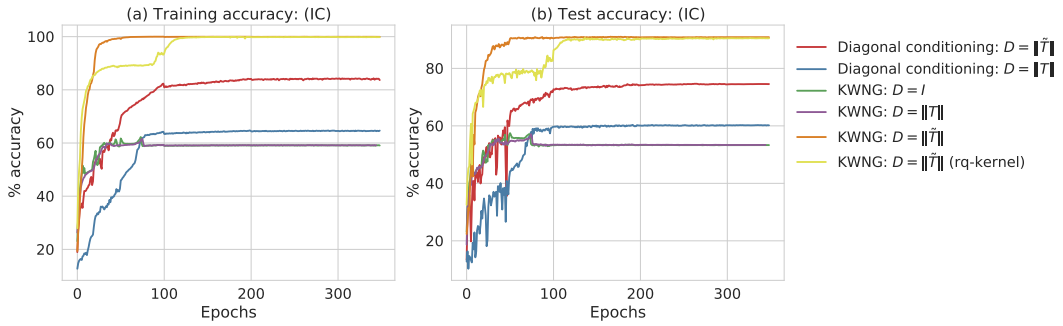


Figure 7.7: KWNG vs Diagonal conditioning in the ill-conditioned case on Cifar10. In red and blue, the euclidean gradient is preconditioned using a diagonal matrix D either given by $D_i = \|T_{.,i}\|$ or $D_i = \|\widetilde{T}_{.,i}\|$, where T and $\widetilde{T}$ are defined in Propositions 77 and 78. The rest of the traces are obtained using the stable version of KWNG in Proposition 78 with different choices for the damping term $D = I$, $D = \|T_{.,i}\|$ and $\|\widetilde{T}_{.,i}\|$. All use a gaussian kernel except the yellow traces which uses a rational quadratic kernel.

Supplementary

A Preliminaries

A .1 Notation

We recall that Ω is an open subset of $\mathbb{R}^d$ while Θ is an open subset of parameters in $\mathbb{R}^q$. Let $\mathcal{Z} \subset \mathbb{R}^p$ be a latent space endowed with a probability distribution ν over $\mathcal{Z}$. Additionally, $(\theta, z) \mapsto h_\theta(z) \in \Omega$ is a function defined over $\Theta \times \mathcal{Z}$. We consider a parametric set of probability distributions $\mathcal{P}_\Theta$ over Ω defined as the implicit model:

$$\mathcal{P}_\Theta := \{\rho_\theta := (h_\theta)_\# \nu \quad ; \quad \theta \in \Theta\},$$

where by definition, $\rho_\theta = (h_\theta)_\# \nu$ means that any sample x from ρ_θ can be written as $x = h_\theta(z)$ where z is a sample from ν. We will write B to denote the jacobian of h_θ w.r.t. θ viewed as a linear map from $\mathbb{R}^q$ to $L_2(\nu)^d$ without explicit reference to θ:

$$Bu(z) = \nabla h_\theta(z).u; \qquad \forall u \in \mathbb{R}^q.$$

As in the main text, $\mathcal{L} : \Theta \to \mathbb{R}$ is a loss functions which is assumed to be of the form $\mathcal{L} = \mathcal{F}(\rho_\theta)$, with $\mathcal{F}$ being a real valued functional over the set of probability distributions. $\nabla\mathcal{L}(\theta)$ denotes the euclidean gradient of $\mathcal{L}$ w.r.t θ while $\widehat{\nabla\mathcal{L}(\theta)}$ is an estimator of $\nabla\mathcal{L}(\theta)$ using N samples from ρ_θ.

We also consider a Reproducing Kernel Hilbert Space $\mathcal{H}$ of functions defined over Ω with inner product $\langle ., . \rangle_\mathcal{H}$ and norm $\|.\|_\mathcal{H}$ and with a kernel $k : \Omega \times \Omega \to \mathbb{R}$. The reproducing property for the derivatives [Steinwart and Christmann, 2008, Lemma 4.34] will be important: $\partial_i f(x) = \langle f, \partial_i k(x, .) \rangle_\mathcal{H}$ for all $x \in \Omega$. It holds as

long as k is differentiable.

$C_b^\infty(\Omega)$ denotes the space of smooth bounded real valued functions on Ω, and $C_c^\infty(\Omega) \subset C_b^\infty(\Omega)$ denotes the subset of compactly supported functions. For any measured space $\mathcal{Z}$ with probability distribution ν, we denote by $L_2(\nu)$ the space of real valued and square integrable functions under ν and by $L_2(\nu)^d$ the space of square integrable vector valued functions under ν and with values in $\mathbb{R}^d$.

A .2 Assumptions

We make the following set of assumptions:

(A) Ω is a non-empty open subset of $\mathbb{R}^d$.

(B) There exists positive constants ζ and σ such that $\int \|z\|^p \, \mathrm{d}\nu(z) \leq \frac{1}{2} p! \zeta^{p-2} \sigma^2$ for any $p \geq 2$.

(C) For all $\theta \in \Theta$ there exists $C(\theta)$ such that $\|\nabla_\theta h_\theta(z)\| \leq C(\theta)(1 + \|z\|)$ for all $z \in \mathcal{Z}$.

(D) k is twice continuously differentiable on $\Omega \times \Omega$.

(E) For all $\theta \in \Theta$ it holds that $\int \partial_i \partial_{i+d} k(x, x) \, \mathrm{d}p_\theta(x) < \infty$ for all $1 \leq i \leq d$.

(F) The following quantity is finite: $\kappa^2 = \sup_{\substack{x \in \Omega \\ 1 \leq i \leq q}} \partial_i \partial_{i+q} k(x, x)$.

(G) For all $0 \leq \delta \leq 1$, it holds with probability at least $1 - \delta$ that $\|\widehat{\nabla \mathcal{L}(\theta)} - \nabla \mathcal{L}(\theta)\| \lesssim N^{-\frac{1}{2}}$.

Remark 2. *Assumption (G) holds if for instance $\widehat{\nabla \mathcal{L}(\theta)}$ can be written as an empirical mean of i.i.d. terms with finite variance:*

$$\widehat{\nabla \mathcal{L}(\theta)} = \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta l(h_\theta(Z_i))$$

where Z_i are i.i.d. samples from the latent distribution ν where $\int \nabla_\theta l(h_\theta(z)) \, \mathrm{d}\nu(z) = \mathcal{L}(\theta)$. This is often the case in the problems considered in machine-learning. In, this

case, the sum of variances of the vector $\widehat{\nabla \mathcal{L}(\theta)}$ *along its coordinates satisfies:*

$$\int \|\widehat{\nabla \mathcal{L}(\theta)} - \nabla \mathcal{L}(\theta)\|^2 \, \mathrm{d}\nu(z) = \frac{1}{N} \int \|\nabla_\theta l(h_\theta(z))\|^2 \, \mathrm{d}\nu(z) := \frac{1}{N}\sigma^2$$

One can then conclude using Cauchy-Schwarz inequality followed by Chebychev's inequality that with probability $1 - \delta$*:*

$$\|\widehat{\nabla \mathcal{L}(\theta)} - \nabla \mathcal{L}(\theta)\| \leq \frac{\sigma}{\sqrt{\delta N}}$$

Moreover, Assumption **(C)** *is often satisfied when the implicit model is chosen to be a deep networks with ReLU non-linearity.*

A .3 Operators definition

Differential operators. We introduce the linear L operator and its adjoint $L^\top$:

$$\begin{aligned} L :& \mathcal{H} \to L_2(\nu)^d & L^\top :& L_2(\nu)^d \to \mathcal{H} \\ & f \mapsto (\partial_i f \circ h_\theta)_{1 \leq i \leq d} & & v \mapsto \int \sum_{i=1}^{d} \partial_i k(h_\theta(z), .) v_i(z) \, \mathrm{d}\nu(z) \end{aligned}$$

This allows to obtain the linear operator A defined in Assumption 4 in the main text by composition $A := L^\top L$. We recall here another expression for A in terms of outer product $\otimes$ and its regularized version for a given $\lambda > 0$,

$$A = \int \sum_{i=1}^{d} \partial_i k(h_\theta(z), .) \otimes \partial_i k(h_\theta(z), .) \, \mathrm{d}\nu(z) \qquad A_\lambda := A + \lambda I.$$

It is easy to see that A is a symmetric positive operator. Moreover, it was established in Sriperumbudur et al. [2017] that A is also a compact operator under Assumption **(E)**.

Assume now we have access to N samples $(Z_n)_{1 \leq n \leq N}$ as in the main text. We define the following objects:

$$\hat{A} := \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{d} \partial_i k(h_\theta(Z_n), .) \otimes \partial_i k(h_\theta(Z_n), .), \qquad \hat{A}_\lambda := \hat{A} + \lambda I.$$

Furthermore, if v is a continuous function in $L_2(\nu)^d$, then we can also consider an empirical estimator for $L^\top v$:

$$\widehat{L^\top v} := \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{d}\partial_i k(h_\theta(Z_n),.)v_i(Z_n).$$

Subsampling operators. We consider the operator Q_M defined from $\mathcal{H}$ to $\mathbb{R}^M$ by:

$$Q_M := \frac{\sqrt{q}}{\sqrt{M}}\sum_{m=1}^{M} e_m \otimes \partial_{i_m} k(Y_m,.) \tag{7.20}$$

where $(e_m)_{1\leq m\leq M}$ is an orthonormal basis of $\mathbb{R}^M$. Q_M admits a singular value decomposition of the form $Q_M = U\Sigma V^\top$, with $VV^\top := P_M$ being the orthogonal projection operator on the Nyström subspace $\mathcal{H}_M$. Similarly to Rudi et al. [2015], Sutherland et al. [2018], we define the projected inverse function $\mathcal{G}_M(C)$ as:

$$\mathcal{G}_M(C) = V(V^\top CV)^{-1}V^\top.$$

We recall here some properties of $\mathcal{G}_M$ from [Sutherland et al., 2018, Lemma 1]:

Lemma 80. *Let $A : \mathcal{H} \to \mathcal{H}$ be a positive operator, and define $A_\lambda = A + \lambda I$ for any $\lambda > 0$. The following holds:*

1. $\mathcal{G}_M(A)P_M = \mathcal{G}_M(A)$
2. $P_M\mathcal{G}_M(A) = \mathcal{G}_M(A)$
3. $\mathcal{G}_M(A_\lambda)A_\lambda P_M = P_M$
4. $\mathcal{G}_M(A_\lambda) = (P_M A P_M + \lambda I)^{-1}P_M$
5. $\|A_\lambda^{\frac{1}{2}}\mathcal{G}_M(A_\lambda)A_\lambda^{\frac{1}{2}}\|$

Estimators of the Wasserstein information matrix. Here we would like to express the estimator in Proposition 77 in terms of the operators introduced previously. We have the following proposition:

Proposition 81. *The estimator defined in Proposition 77 admits the following representation:*

$$\widehat{\nabla^W \mathcal{L}(\theta)} = (\epsilon D(\theta) + G_{M,N})^{-1} \widehat{\nabla \mathcal{L}(\theta)}$$

where $G_{M,N}$ is given by:

$$G_{M,N} := (\widehat{L^\top B})^\top \mathcal{G}_M(\hat{A}_\lambda) \widehat{L^\top B}.$$

Proof. This a direct consequence of the minimax theorem [Ekeland and Témam, 1999, Proposition 2.3, Chapter VI] and applying [Sutherland et al., 2018, Lemma 3]. □

The matrix $G_{M,N}$ is in fact an estimator of the Wasserstein information matrix defined in Definition 10. We will also need to consider the following population version of $G_{M,N}$ defined as :

$$G_M := (L^\top B)^\top \mathcal{G}_M(A_\lambda) L^\top B \tag{7.21}$$

B Proofs

B .1 Preliminary results

Here we provide a proof of the invariance properties of the Fisher and Wasserstein natural gradient descent in the continuous-time limit as stated in Proposition 73. Consider an invertible and smoothly differentiable re-parametrization Ψ, satisfying $\psi = \Psi(\theta)$. Denote by $\bar{\rho}_\psi = \rho_{\Psi^{-1}(\psi)}$ the re-parametrized model and $\bar{G}_W(\psi)$ and $\bar{G}_F(\psi)$ their corresponding Wasserstein and Fisher information matrices whenever they are well defined.

Proof of Proposition 73. Here we only consider the case when $\nabla^D \mathcal{L}(\theta)$ is either given by the Fisher natural gradient $\nabla^F \mathcal{L}(\theta)$ or the Wasserstein Natural gradient $\nabla^W \mathcal{L}(\theta)$. We will first define $\widetilde{\psi}_s := \Psi(\theta_s)$ and show that in fact $\widetilde{\psi}_s = \psi_s$ at all times

$s > 0$. First, let's differentiate $\widetilde{\psi}_s$ in time:

$$\dot{\widetilde{\psi}}_s = -\nabla_\theta \Psi(\theta_s)^\top G_D(\theta_s)^{-1} \nabla_\theta \mathcal{L}(\theta_s)$$

By the chain rule, we have that $\nabla_\theta \mathcal{L}(\theta_s) = \nabla_\theta \Psi(\theta_s) \nabla_\psi \bar{\mathcal{L}}(\widetilde{\psi}_s)$, hence:

$$\dot{\widetilde{\psi}}_s = -\nabla_\theta \Psi(\theta_s)^\top G_D(\theta_s)^{-1} \nabla_\theta \Psi(\theta_s) \nabla_\psi \bar{\mathcal{L}}(\widetilde{\psi}_s).$$

It is easy to see that $\nabla_\theta \Psi^{-1}(\widetilde{\psi}_s) = (\nabla_\psi \Psi^{-1}(\psi_s))^{-1}$ by definition of Ψ and $\widetilde{\psi}_s$, hence by Lemma 82 one can conclude that:

$$\dot{\widetilde{\psi}}_s = -G_D(\widetilde{\psi}_s)^{-1} \nabla_\psi \bar{\mathcal{L}}(\widetilde{\psi}_s).$$

Hence, $\widetilde{\psi}_s$ satisfies the same differential equation as ψ_s. Now keeping in mind that $\psi_0 = \widetilde{\psi}_0 = \Psi(\theta_0)$, it follows that $\psi_0 = \widetilde{\psi}_0 = \Psi(\theta_0)$ by uniqueness of differential equations. □

Lemma 82. *Under conditions of Propositions 74 and 75, the informations matrices $\bar{G}_W(\psi)$ and $\bar{G}_F(\psi)$ are related to $G_W(\theta)$ and $G_F(\theta)$ by the relation:*

$$\bar{G}_W(\psi) = \nabla_\psi \Psi^{-1}(\psi)^\top G_W(\theta) \nabla_\psi \Psi^{-1}(\psi)$$
$$\bar{G}_F(\psi) = \nabla_\psi \Psi^{-1}(\psi)^\top G_F(\theta) \nabla_\psi \Psi^{-1}(\psi)$$

Proof. Let $v \in R^q$ and write $u = \nabla_\theta \Psi^{-1}(\psi) v$, then by the dual formulations of $G_W(\theta)$ and $G_F(\theta)$ in Proposition 74 we have that:

$$\begin{aligned} &\frac{1}{2} v^\top \nabla_\psi \Psi^{-1}(\psi)^\top G_F(\theta) \nabla_\psi \Psi^{-1}(\psi) v \\ = &\sup_{\substack{f \in C_c^\infty(\Omega) \\ \int f(x)\, \mathrm{d}\rho_\theta(x) = 0}} \nabla \rho_\theta(f)^\top \nabla_\theta \Psi^{-1}(\psi) v - \frac{1}{2} \int f(x)^2 \, \mathrm{d}\rho_\theta(x) \, \mathrm{d}x, \end{aligned}$$

Now recalling that $\nabla_\psi \bar{\rho}_\psi = \nabla_\theta \rho_\theta \nabla_\psi \Psi^{-1}(\psi)$ by Lemma 83, it follows that:

$$\begin{aligned}
&\frac{1}{2} v^\top \nabla_\psi \Psi^{-1}(\psi)^\top G_F(\theta) \nabla_\psi \Psi^{-1}(\psi) v \\
&= \sup_{\substack{f \in C_c^\infty(\Omega) \\ \int f(x)\, \mathrm{d}\rho_\theta(x) = 0}} \nabla_\psi \bar{\rho}_\psi(f)^\top v - \frac{1}{2} \int f(x)^2 \, \mathrm{d}\rho_\theta(x)\, \mathrm{d}x,
\end{aligned}$$

Using again Proposition 74 for the reparametrized model $\bar{\rho}_\psi$, we directly have that:

$$\frac{1}{2} v^\top G_F(\psi) v = \sup_{\substack{f \in C_c^\infty(\Omega) \\ \int f(x)\, \mathrm{d}\rho_\theta(x) = 0}} \nabla_\psi \bar{\rho}_\psi(f)^\top v - \frac{1}{2} \int f(x)^2 \, \mathrm{d}\rho_\theta(x)\, \mathrm{d}x,$$

The result follows by equating both expression. The same procedure can be applied for the case the Wasserstein information matrix using Proposition 75. □

Lemma 83. *The distributional gradients $\nabla_\psi \bar{\rho}_\psi$ and $\nabla_\theta \rho_\theta$ are related by the expression:*

$$\nabla_\psi \bar{\rho}_\psi = \nabla_\theta \rho_\theta \nabla_\psi \Psi^{-1}(\psi)$$

Proof. The proof follows by considering a fixed direction $u \in \mathbb{R}^q$ and a test function $f \in \mathcal{C}_c^\infty(\Omega)$ and the definition of distributional gradient in Definition 11:

$$\begin{aligned}
\nabla \bar{\rho}_\psi(f)^\top u &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int f(x)\, \mathrm{d}\bar{\rho}_{\psi + \epsilon u}(x) - \int f(x)\, \mathrm{d}\bar{\rho}_\psi(x) \\
&= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int f(x)\, \mathrm{d}\rho_{\Psi^{-1}(\psi + \epsilon u)}(x) - \int f(x)\, \mathrm{d}\rho_{\Psi^{-1}(\psi)}(x)
\end{aligned}$$

Now by differentiability of Ψ^{-1}, we have the following first order expansion:

$$\Psi^{-1}(\psi + \epsilon u) = \Psi^{-1}(\psi) + \epsilon \nabla \Psi^{-1}(\psi)^\top u + \epsilon \upsilon(\epsilon)$$

where $\upsilon(\epsilon)$ converges to 0 when $\epsilon \to 0$. Now using again the definition Definition 11

for ρ_{Ψ^ψ} one has:

$$\frac{1}{\epsilon}\int f(x)\,\mathrm{d}\left(\rho_{\Psi^{-1}(\psi+\epsilon u)}-\rho_{\Psi^{-1}(\psi)}\right)(x) = \nabla\rho_{\Psi^{-1}(\psi)}(f)^\top\nabla\Psi^{-1}(\psi)^\top u + \epsilon\upsilon(\epsilon)+\delta(\epsilon,f,(\nabla\Psi^{-1}(\psi)u+\upsilon(\epsilon)))$$

The last two terms converge to 0 as $\epsilon\to 0$, hence leading to the desired expression. □

Proposition 84. *When ρ_θ admits a density that is continuously differentiable w.r.t θ and such that $x\mapsto\nabla\rho_\theta(x)$ is continuous, then the distributional gradient is of the form:*

$$\nabla\rho_\theta(f)=\int f(x)\nabla\rho_\theta(x)\,\mathrm{d}x,\qquad \forall f\in\mathcal{C}_c^\infty(\Omega)$$

where $\nabla\rho_\theta(x)$ denotes the gradient of the density of $\rho_\theta(x)$ at x.

Proof. Let $\epsilon>0$ and $u\in\mathbb{R}^q$, we define the function $\nu(\epsilon,u,f)$ as follows:

$$\nu(\epsilon,u,f)=\int f(x)\left(\frac{1}{\epsilon}(\rho_{\theta+\epsilon u})-\rho_\theta-\nabla\rho_\theta^\top u\right)\mathrm{d}x$$

we just need to show that $\nu(\epsilon,u,f)\to 0$ as $\epsilon\to 0$. This follows form the differentiation lemma [Klenke, 2008, Theorem 6.28] applied to the function $(\theta,x)\mapsto f(x)\rho_\theta(x)$. Indeed, this function is integrable in x for any θ' in a neighborhood U of θ that is small enough, it is also differentiable on that neighborhood U and satisfies the domination inequality:

$$|f(x)\nabla\rho_\theta(x)^\top u|\le|f(x)|\sup_{x\in Supp(f),\theta\in U}\nabla\rho_\theta(x)^\top u|.$$

The inequality follows from continuity of $(\theta,x)\nabla\rho_\theta(x)$ and recalling that f is compactly supported. This concludes the proof. □

We fist provide a proof of the dual formulation for the Fisher information matrix.

Proof of Proposition 74 . Consider the optimization problem:

$$\sup_{\substack{f\in C_c^\infty(\Omega)\\ \int f(x)\,\mathrm{d}\rho_\theta(x)=0}} \left(\int f(x)\nabla\rho_\theta(x)\,\mathrm{d}x\right)^\top u - \frac{1}{2}\int f(x)^2\rho_\theta(x)\,\mathrm{d}x \tag{7.22}$$

Recalling that the set of smooth and compactly supported functions $C_c^\infty(\infty)$ is dense in $L_2(\rho_\theta)$ and that the objective function in (7.22) is continuous and coercive in f, it follows that (7.22) admits a unique solution f^* in $L_2(\rho_\theta)$ which satisfies the optimality condition:

$$\int f(x)(\nabla\rho_\theta(x))^\top u\,\mathrm{d}x = \int f(x)f^*(x)\rho_\theta(x)\,\mathrm{d}x \qquad \forall f\in L_2(\rho_\theta)$$

Hence, it is easy to see that $f^* = (\nabla\rho_\theta)^\top u/\rho_\theta$ and that the optimal value of (7.22) is given by:

$$\frac{1}{2}\int \frac{((\nabla\rho_\theta(x))^\top u)^2}{\rho_\theta(x)}\,\mathrm{d}x.$$

This is equal to $u^\top G_F(\theta)u$ by Definition 9. □

The next proposition ensures that the Wasserstein information matrix defined in Definition 10 is well-defined and has a dual formulation.

Proposition 85. *Consider the model defined in (7.6) and let $(e_s)_{1\le s\le q}$ be an orthonormal basis of $\mathbb{R}^q$. Under Assumptions (B) and (C), there exists an optimal solution $\Phi = (\phi_s)_{1\le s\le q}$ with ϕ_s in $L_2(\rho_\theta)^d$ satisfying the PDE:*

$$\partial_s\rho_\theta = -div(\rho_\theta\phi_s)$$

The elliptic equations also imply that $L^\top\nabla h_\theta = L^\top(\Phi\circ h_\theta)$. Moreover, the Wasserstein information matrix $G_W(\theta)$ on $\mathcal{P}_\Theta$ at point θ can be written as $G_W(\theta) = \Phi^\top\Phi$ where the inner-product is in $L_2(\rho_\theta)^d$ and satisfies:

$$\frac{1}{2}u^\top G_W(\theta)u = \sup_{\substack{f\in C_c^\infty(\Omega)\\ \int f(x)\,\mathrm{d}\rho_\theta(x)=0}} \nabla\rho_\theta(f)^\top u - \frac{1}{2}\int \|\nabla_x f(h_\theta(z))\|^2\,\mathrm{d}\nu(z).$$

for all $u\in\mathbb{R}^q$.

Proof. Let $(e_s)_{1\leq s\leq q}$ be an orthonormal basis of $\mathbb{R}^q$. For all $1 \leq s \leq q$, we will establish the existence of an optimal solution ϕ_s in $L_2(\rho_\theta)^d$ satisfying the PDE:

$$\partial_s \rho_\theta = -div(\rho_\theta \phi_s) \tag{7.23}$$

Consider the variational problem:

$$\sup_{\phi\in\mathcal{S}} \int \phi(h_\theta(z))^\top \partial_{\theta_s} h_\theta(z) - \frac{1}{2}\|\phi\|^2_{L_2(\rho_\theta)} \tag{7.24}$$

where $\mathcal{S}$ is a Hilbert space obtained as the closure in $L_2(\rho_\theta)^d$ of functions of the form $\phi = \nabla_x f$ with $f \in C_c^\infty(\Omega)$:

$$\mathcal{S} := \overline{\{\nabla_x f \quad | \quad f \in C_c^\infty(\Omega)\}}_{L_2(\rho_\theta^d)}.$$

We have by Assumption **(C)** that:

$$\int \phi(h_\theta(z))^\top \partial_{\theta_s} h_\theta(z)\, \mathrm{d}\nu(z) \leq\leq C(\theta)\sqrt{\int (1+\|z\|^2)\, \mathrm{d}\nu(z)} \int \|\phi\|_{L_2(\rho_\theta)}.$$

Moreover, by Assumption **(B)**, we know that $\sqrt{\int (1+\|z\|^2)\, \mathrm{d}\nu(z)} < \infty$. This implies that the objective in (7.24) is continuous in ϕ while also being convex.s It follows that (7.24) admits a unique solution $\phi_s^* \in \mathcal{S}$ which satisfies for all $\phi \in \mathcal{S}$:

$$\int \phi(h_\theta(z))^\top \phi_s^*(h_\theta(z))\, \mathrm{d}\nu(z) = \int \phi(h_\theta(z))^\top \partial_{\theta_s} h_\theta(z))\, \mathrm{d}\nu(z)$$

In particular, for any $f \in C_c^\infty(\Omega)$, it holds that:

$$\int \nabla_x f(h_\theta(z))^\top \phi_s^*(h_\theta(z))\, \mathrm{d}\nu(z) = \int \nabla_x f(h_\theta(z))^\top \partial_{\theta_s} h_\theta(z))\, \mathrm{d}\nu(z)$$

which is equivalent to (7.23) and implies directly that $L^T \nabla h_\theta = L^T \Phi \circ h_\theta$ where $\Phi := (\phi_s^*)_{1\leq s\leq q}$. The variational expression for $\frac{1}{2}u^\top G_W u$ follows by noting that

(7.24) admits the same optimal value as

$$\sup_{\substack{f\in C_c^\infty(\Omega)\\ \int f(x)\,\mathrm{d}\rho_\theta(x)=0}} \nabla\rho_\theta(f)^\top u - \frac{1}{2}\int \|\nabla_x f(h_\theta(z))\|^2\,\mathrm{d}\nu(z).$$

That is because $\mathcal{S}$ is by definition the closure in $L_2(\rho_\theta)^d$ of the set of gradients of smooth and compactly supported functions on Ω. □

Proof of Proposition 75. This is a consequence of Proposition 85. □

B .2 Expression of the Estimator

We provide here a proof of Proposition 77

Proof of Proposition 77. Here, to simplify notations, we simply write D instead of $D(\theta)$. First consider the following optimization problem:

$$\inf_{f\in\mathcal{H}_M} \frac{1}{N}\sum_{n=1}^N \|\nabla f(X_n)\|^2 + \lambda\|f\|_{\mathcal{H}}^2 + \frac{1}{\epsilon}\mathcal{R}(f)^\top D^{-1}\mathcal{R}(f) + \frac{2}{\epsilon}\mathcal{R}(f)^\top D^{-1}\widehat{\nabla\mathcal{L}(\theta)}$$

with $\mathcal{R}(f)$ given by $\mathcal{R}(f) = \frac{1}{N}\sum_{n=1}^N \nabla f(X_n)^\top B(Z_n)$. Now, recalling that any $f \in \mathcal{H}_M$ can be written as $f = \sum_{m=1}^M \alpha_m \partial_{i_m} k(Y_m, .)$, and using the reproducing property $\partial_i f(x) = \langle f, \partial_i k(x, .)\rangle_{\mathcal{H}}$ [Steinwart and Christmann, 2008, Lemma 4.34] , it is easy to see that:

$$\frac{1}{N}\sum_{n=1}^N \|\nabla f(X_n)\|^2 = \frac{1}{N}\sum_{\substack{1\le n\le N\\ 1\le i\le d}} \Big(\sum_{m=1}^M \alpha_m \partial_{i_m}\partial_{i+d} k(Y_m, X_n)\Big)^2.$$

$$\|f\|_{\mathcal{H}}^2 = \sum_{1\le m,m'\le M} \alpha_m\alpha_{m'}\partial_{i_m}\partial_{i_{m'}+d}k(Y_m, Y_{m'})$$

$$\mathcal{R}(f) = \frac{1}{N}\sum_{\substack{1\le n\le N\\ 1\le i\le d\\ 1\le m\le M}} \alpha_m \partial_{i_m}\partial_{i+d}k(Y_m, X_n)B_i(Z_n)$$

The above can be expressed in matrix form using the matrices defined in Proposi-

tion 77:

$$\frac{1}{N}\sum_{n=1}^{N}\|\nabla f(X_n)\|^2 = \alpha^\top CC^\top\alpha; \qquad \|f\|_{\mathcal{H}}^2 = \alpha^\top K\alpha; \qquad \mathcal{R}(f) = \alpha^\top CB.$$

Hence the optimal solution $\hat{f}^*$ is of the form $\hat{f}^* = \sum_{m=1}^{M}\alpha_m^*\partial_{i_m}k(Y_m,.)$, with α^* obtained as a solution to the finite dimensional problem in $\mathbb{R}^M$:

$$\min_{\alpha\in\mathbb{R}^M}\alpha^\top(\epsilon CC^\top + \epsilon\lambda K + CBD^{-1}B^\top C^\top)\alpha + 2\alpha^\top CBD^{-1}\widehat{\nabla\mathcal{L}(\theta)}$$

It is easy to see that α^* are given by:

$$\alpha^* = -(\epsilon CC^T + \epsilon\lambda K + CBD^{-1}B^TC^T)^\dagger CBD^{-1}\widehat{\nabla\mathcal{L}(\theta)}.$$

Now recall that the estimator in Proposition 77 is given by: $\widehat{\nabla^W\mathcal{L}(\theta)} = \frac{1}{\epsilon}D^{-1}\mathcal{U}_\theta(\hat{f}^*)$. Hence, $\frac{1}{\epsilon}D^{-1}(\widehat{\nabla\mathcal{L}(\theta)} - B^TC^T\alpha^*)$ The desired expression is obtained by noting that $CB = T$ using the chain rule.

□

B .3 Consistency Results

Well-specified case. Here, we assume that the vector valued functions $(\phi_i)_{1\le i\le q}$ involved in Definition 10 can be expressed as gradients of functions in $\mathcal{H}$. More precisely:

Assumption 4. *For all $1 \le i \le q$, there exits functions $f_i \in \mathcal{H}$ such that $\phi_i = \nabla f_i$. Additionally, f_i are of the form $f_i = A^\alpha v_i$ for some fixed $\alpha \ge 0$, with $v_i \in \mathcal{H}$ and A being the differential covariance operator defined on $\mathcal{H}$ by $A : f \mapsto \int\sum_{i=1}^{d}\partial_i k(h_\theta(z),.)\partial_i f(h_\theta(z))\,\mathrm{d}\nu(z)$.*

The parameter α characterizes the smoothness of f_i and therefore controls the statistical complexity of the estimation problem. Using a similar analysis as Sutherland et al. [2018] we obtain a convergence rate for the estimator in Proposition 77

the following convergence rates for the estimator in Proposition 77:

Theorem 86. *Let δ be such that $0 \leq \delta \leq 1$ and $b := \min(1, \alpha + \frac{1}{2})$. Under Assumption 4 and Assumptions (A) to (G) listed in Section A .2, for N large enough, $M \sim (dN^{\frac{1}{2b+1}} \log(N))$, $\lambda \sim N^{-\frac{1}{2b+1}}$ and $\epsilon \lesssim N^{-\frac{b}{2b+1}}$, it holds with probability at least $1 - \delta$ that:*

$$\|\widehat{\nabla^W \mathcal{L}(\theta)} - \nabla^W \mathcal{L}(\theta)\|^2 = \mathcal{O}\Big(N^{-\frac{2b}{2b+1}}\Big).$$

In the worst case where $\alpha = 0$, the proposed estimator needs at most $M \sim (d\sqrt{N}\log(N))$ to achieve a convergence rate of $N^{-\frac{1}{2}}$. The smoothest case requires only $M \sim (dN^{\frac{1}{3}}\log(N))$ to achieve a rate of $N^{-\frac{2}{3}}$. Thus, the proposed estimator enjoys the same statistical properties as the ones proposed by Sriperumbudur et al. [2017], Sutherland et al. [2018] while maintaining a computational advantage[1] tNow we provide a proof for Theorem 86 which relies on the same techniques used by Rudi et al. [2015], Sutherland et al. [2018].

Proof of Theorem 86 . The proof is a direct consequence of Proposition 87 under Assumption 4. □

Proof of Theorem 79. The proof is a direct consequence of Proposition 87 under Assumption 3. □

Proposition 87. *Under Assumptions (A) to (G) and for $0 \leq \delta \leq 1$ and N large enough, it holds with probability at least $1 - \delta$:*

$$\|\widehat{\nabla^W \mathcal{L}} - \nabla^W \mathcal{L}\| = \mathcal{O}(N^{-\frac{b}{2b+1}})$$

provided that $M \sim dN^{\frac{1}{2b+1}} \log N$, $\lambda \sim N^{\frac{1}{2b+1}}$ and $\epsilon \lesssim N^{-\frac{b}{2b+1}}$ where $b := \min(1, \alpha + \frac{1}{2})$ when Assumption 4 holds and $b = \frac{1}{2+c}$ when Assumption 3 holds instead.

Proof. Here for simplicity we assume that $D(\theta) = I$ without loss of generality and we omit the dependence in θ and write $\nabla^W \mathcal{L}$ and $\nabla \mathcal{L}$ instead of $\nabla^W \mathcal{L}(\theta)$ and $\nabla \mathcal{L}(\theta)$

[1] The estimator proposed by Sutherland et al. [2018] also requires M to grow linearly with the dimension d although such dependence doesn't appear explicitly in the statement of [Sutherland et al., 2018, Theorem 2].

and $\nabla^W \mathcal{L}(\theta)$. We also define $\hat{G}_\epsilon = \epsilon I + G_{M,N}$ and $G_\epsilon = \epsilon I + G_W$. By Proposition 81, we know that $\widehat{\nabla^W \mathcal{L}} = \hat{G}_\epsilon^{-1} \widehat{\nabla \mathcal{L}}$. We use the following decomposition:

$$\begin{aligned}\|\widehat{\nabla^W \mathcal{L}} - \nabla^W \mathcal{L}\| \leq & \|\hat{G}_\epsilon^{-1}(\widehat{\nabla \mathcal{L}} - \nabla \mathcal{L})\| + \epsilon \|\hat{G}_\epsilon^{-1} G_W^{-1} \nabla \mathcal{L}\| \\ & + \|\hat{G}_\epsilon^{-1}(G_{M,N} - G_W) G_W^{-1} \nabla \mathcal{L}\|\end{aligned}$$

To control the norm of $\hat{G}_\epsilon^{-1}$ we write $\hat{G}_\epsilon^{-1} = G_\epsilon^{-\frac{1}{2}}(H+I)^{-1} G_\epsilon^{-\frac{1}{2}}$, where H is given by $H := G_\epsilon^{-\frac{1}{2}}(G_{M,N} - G_W) G_\epsilon^{-\frac{1}{2}}$. Hence, provided that $\mu := \lambda_{\max}(H)$, the highest eigenvalue of H, is smaller than 1, it holds that:

$$\|(H+I)^{-1}\| \leq (1-\mu)^{-1}.$$

Moreover, since G_W is positive definite, its smallest eigenvalue η is strictly positive. Hence, $\|G_\epsilon^{-1}\| \leq (\eta + \epsilon)^{-1}$. Therefore, we have $\|\hat{G}_\epsilon^{-1}\| \leq (\eta+\epsilon)(1-\mu))^{-1}$, which implies:

$$\begin{aligned}\|\widehat{\nabla^W \mathcal{L}} - \nabla^W \mathcal{L}\| \leq & (\eta+\epsilon)^{-1} \left(\frac{\|\widehat{\nabla \mathcal{L}} - \nabla \mathcal{L}\|}{1-\mu} + \epsilon \eta^{-1} \|\nabla \mathcal{L}\| \right) \\ & + \eta^{-1}(\eta+\epsilon)^{-1} \|\nabla \mathcal{L}\| \|G_{M,N} - G_W\|\end{aligned}$$

Let $0 \leq \delta \leq 1$. We have by Assumption **(G)** that $\|\widehat{\nabla \mathcal{L}} - \nabla \mathcal{L}\| = \mathcal{O}(N^{-\frac{1}{2}})$ with probability at least $1-\delta$. Similarly, by Proposition 88 and for N large enough, we have with probability at least $1-\delta$ that $\|G_{M,N} - G_W\| = \mathcal{O}(N^{-\frac{b}{2b+1}})$ where b is defined in Proposition 88. Moreover, for N large enough, one can ensure that $\mu \leq \frac{1}{2}$ so that the following bound holds with probability at least $1-\delta$:

$$\|\widehat{\nabla^W \mathcal{L}} - \nabla^W \mathcal{L}\| \lesssim (\eta+\epsilon)^{-1} \left(2N^{-\frac{1}{2}} + \eta^{-1} \|\nabla \mathcal{L}\| (N^{-\frac{b}{2b+1}} + \epsilon) \right).$$

Thus by setting $\epsilon \lesssim N^{-\frac{b}{2b+1}}$ we get the desired convergence rate. □

Proposition 88. *For any $0 \leq \delta \leq 1$, we have with probability as least $1-\delta$ and for*

N large enough that:

$$\|G_{M,N} - G_W\| = \mathcal{O}(N^{-\frac{b}{2b+1}}).$$

provided that $M \sim dN^{\frac{1}{2b+1}} \log N$ *where* $b := \min(1, \alpha + \frac{1}{2})$ *when Assumption 4 holds and* $b = \frac{1}{2+c}$ *when Assumption 3 holds instead.*

Proof. To control the error $\|G_{M,N} - G_W\|$ we decompose it into an estimation error $\|G_{M,N} - G_M\|$ and approximation error $\|G_M - G_W\|$:

$$\|G_{M,N} - G_W\| \leq \|G_M - G_W\| + \|G_M - G_{M,N}\|$$

were G_M is defined in (7.21) and is obtained by taking the number of samples N to infinity while keeping the number of basis points M fixed.

The estimation error $\|G_M - G_{M,N}\|$ is controlled using Proposition 89 where, for any $0 \leq \delta \leq 1$, we have with probability at least $1-\delta$ and as long as $N \geq M(1, \lambda, \delta)$:

$$\|G_{M,N} - G_M\| \leq \frac{\|B\|}{\sqrt{N\lambda}}(a_{N,\delta} + \sqrt{2\gamma_1\kappa} + 2\gamma_1\frac{\lambda+\kappa}{\sqrt{N\lambda}}) + \frac{1}{N\lambda}a_{N,\delta}^2.$$

In the limit where $N \to \infty$ and $\lambda \to 0$, only the dominant terms in the above equation remain which leads to an error $\|G_{M,N} - G_M\| = \mathcal{O}((N\lambda)^{-\frac{1}{2}})$. Moreover, the condition on N can be expressed as $\lambda^{-1} \log \lambda^{-1} \lesssim N$.

To control the error approximation error $\|G_M - G_W\|$ we consider two cases: the *well-specified* case and the *miss-specified* case.

- *Well-specified* case. Here we work under Assumption 4 which allows to use Proposition 91. Hence, for any $0 \leq \delta \leq 1$ and if $M \geq M(d, \lambda, \delta)$, it holds with probability at least $1 - \delta$:

$$\|G_M - G_W\| \lesssim \lambda^{\min(1,\alpha+\frac{1}{2})}$$

- *Miss-specified* case. Here we work under Assumption 3 which allows to use Proposition 90. Hence, for any $0 \leq \delta \leq 1$ and if $M \geq M(d, \lambda, \delta)$, it holds

with probability at least $1-\delta$:

$$\|G_M - G_W\| \lesssim \lambda^{\frac{1}{2+c}}$$

Let's set $b := \min(1, \alpha + \frac{1}{2})$ for the well-specified case and $b = \frac{1}{2+c}$ for the miss-specified case. In the limit where $M \to \infty$ and $\lambda \to 0$ the condition on M becomes: $M \sim d\lambda^{-1}\log\lambda^{-1}$. Hence, when $M \sim d\lambda^{-1}\log\lambda^{-1}$ and $\lambda^{-1}\log\lambda^{-1} \lesssim N$ it holds with probability as least $1-\delta$ that

$$\|G_{M,N} - G_W\| = \mathcal{O}(\lambda^b + (\lambda N)^{-\frac{1}{2}}).$$

One can further choose λ of the form $\lambda = N^{-\theta}$. This implies a condition on M of the form $dN^\theta\log(N) \lesssim M$ and $N^\theta\log(N) \lesssim N$. After optimizing over θ to get the tightest bound, the optimal value is obtained when $\theta = 1/(2b+1)$ and the requirement on N is always satisfied once N is large enough. Moreover, one can choose $M \sim dN^{\frac{1}{2b+1}}\log N$ so that the requirement on M is satisfied for N large enough. In this case we get the following convergence rate:

$$\|G_{M,N} - G_W\| = \mathcal{O}(N^{-\frac{b}{2b+1}}).$$

□

Proposition 89. *For any $0 \leq \delta \leq 1$, provided that $N \geq M(1,\lambda,\delta)$, we have with probability as least $1-\delta$:*

$$\|G_{M,N} - G_M\| \leq \frac{\|B\|}{\sqrt{N\lambda}}(2a_{N,\delta} + \sqrt{2\gamma_1\kappa} + 2\gamma_1\frac{\lambda+\kappa}{\sqrt{N\lambda}}) + \frac{1}{N\lambda}a_{N,\delta}^2.$$

with:

$$a_{N,\delta} := \sqrt{2\sigma_1^2\log\frac{2}{\delta}} + \frac{2a\log\frac{2}{\delta}}{\sqrt{N}}$$

Proof. For simplicity, we define $E = \widehat{L^\top B} - L^\top B$. By definition of $G_{M,N}$ and G_M

we have the following decomposition:

$$G_{M,N} - G_M = \underbrace{E^\top \mathcal{G}_M(\hat{A}_\lambda)E}_{\mathfrak{E}_0} + \underbrace{E^\top \mathcal{G}_M(\hat{A}_\lambda)L^\top B}_{\mathfrak{E}_1} + \underbrace{B^\top L\mathcal{G}_M(\hat{A}_\lambda)E}_{\mathfrak{E}_2} - \underbrace{B^\top L\mathcal{G}_M(A_\lambda)P_M(\hat{A}-A)P_M\mathcal{G}_M(\hat{A}_\lambda)L^\top B}_{\mathfrak{E}_3}$$

The first three terms can be upper-bounded in the following way:

$$\begin{aligned}
\|\mathfrak{E}_0\| &= \|E^\top \hat{A}_\lambda^{-\frac{1}{2}} \hat{A}_\lambda^{\frac{1}{2}} \mathcal{G}_M(\hat{A}_\lambda)\hat{A}_\lambda^{\frac{1}{2}} \hat{A}_\lambda^{-\frac{1}{2}} E\| \\
&\leq \|E\|^2 \underbrace{\|\hat{A}_\lambda^{-1}\|}_{\leq 1/\lambda} \underbrace{\|\hat{A}_\lambda^{\frac{1}{2}} \mathcal{G}_M(\hat{A}_\lambda)\hat{A}_\lambda^{\frac{1}{2}}\|}_{\leq 1} \\
\|\mathfrak{E}_1\| = \|\mathfrak{E}_2\| &= \|E^\top A_\lambda^{-\frac{1}{2}} A_\lambda^{\frac{1}{2}} \mathcal{G}_M(A_\lambda)A_\lambda^{\frac{1}{2}} A_\lambda^{-\frac{1}{2}} L^\top B\| \\
&\leq \|B\|\|E\| \underbrace{\|\hat{A}_\lambda^{-\frac{1}{2}}\|}_{\leq 1/\sqrt{\lambda}} \underbrace{\|\hat{A}_\lambda^{\frac{1}{2}} \mathcal{G}_M(\hat{A}_\lambda)\hat{A}_\lambda^{\frac{1}{2}}\|}_{\leq 1} \underbrace{\|A_\lambda^{-\frac{1}{2}} L^\top\|}_{\leq 1} \|\hat{A}_\lambda^{-\frac{1}{2}} A_\lambda^{\frac{1}{2}}\|
\end{aligned}$$

For the last term $\mathfrak{E}_3$, we first recall that by definition of $\mathcal{G}_M(A_\lambda)$ we have:

$$\mathcal{G}_M(A_\lambda)P_M(\hat{A}-A)P_M\mathcal{G}_M(A_\lambda) = \mathcal{G}_M(A_\lambda)(\hat{A}-A)\mathcal{G}_M(A_\lambda).$$

Therefore, one can introduce the matrices $A_\lambda^{\frac{1}{2}}$, $\hat{A}_\lambda^{\frac{1}{2}}$ and their inverses in the expression of $\|\mathfrak{E}_3\|$ and write:

$$\begin{aligned}
&\|\mathfrak{E}_3\| \\
=&\|B^\top L\mathcal{G}_M(A_\lambda)(\hat{A}-A)\mathcal{G}_M(\hat{A}_\lambda)L^\top B\| \\
\leq&\|B\|^2 \underbrace{\|LA_\lambda^{-\frac{1}{2}}\|^2}_{\leq 1} \underbrace{\|A_\lambda^{\frac{1}{2}}\mathcal{G}_M(A_\lambda)A_\lambda^{\frac{1}{2}}\|}_{\leq 1} \|A_\lambda^{-\frac{1}{2}}(\hat{A}-A)A_\lambda^{-\frac{1}{2}}\| \underbrace{\|\hat{A}_\lambda^{\frac{1}{2}}\mathcal{G}_M(\hat{A}_\lambda)\hat{A}_\lambda^{\frac{1}{2}}\|}_{\leq 1} \|A_\lambda^{\frac{1}{2}}\hat{A}_\lambda^{\frac{1}{2}}\|^2 \\
\leq&\|B\|^2\|A_\lambda^{\frac{1}{2}}\hat{A}_\lambda^{\frac{1}{2}}\|^2\|A_\lambda^{-\frac{1}{2}}(\hat{A}-A)A_\lambda^{-\frac{1}{2}}\|
\end{aligned}$$

We recall now [Rudi et al., 2015, Proposition 7.] which allows to upper-bound $\|A_\lambda^{\frac{1}{2}}\hat{A}_\lambda^{\frac{1}{2}}\|$ by $(1-\eta)^{-\frac{1}{2}}$ where $\eta = \lambda_{\max}(A_\lambda^{\frac{1}{2}}(A-\hat{A})A_\lambda^{\frac{1}{2}})$ provided that $\eta < 1$. Moreover, [Rudi et al., 2015, Proposition 8.] allows to control both η and $\|A_\lambda^{-\frac{1}{2}}(\hat{A}-A)A_\lambda^{-\frac{1}{2}}\|$ under Assumption **(F)**. Indeed, for any $0 \leq \delta \leq 1$ and provided that

$0 < \lambda \leq \|A\|$ it holds with probability $1-\delta$ that:

$$\|A_\lambda^{-\frac{1}{2}}(\hat{A}-A)A_\lambda^{-\frac{1}{2}}\| \leq 2\gamma_1\frac{1+\kappa/\lambda}{3N}+\sqrt{\frac{2\gamma_1\kappa}{N\lambda}}; \qquad \eta \leq \frac{2\gamma_2}{3N}+\sqrt{\frac{2\gamma_2\kappa}{N\lambda}}$$

where γ_1 and γ_2 are given by:

$$\gamma_1 = \log(\frac{8Tr(A)}{\lambda\delta}); \qquad \gamma_2 = \log(\frac{4Tr(A)}{\lambda\delta}).$$

Hence, for $N \geq M(1,\lambda,\delta)$ we have that $(1-\eta)^{-\frac{1}{2}} \leq 2$ and one can therefore write:

$$\|\mathfrak{E}_3\| \leq 4\|B\|^2(2\gamma_1\frac{1+\kappa/\lambda}{3N}+\sqrt{\frac{2\gamma_1\kappa}{N\lambda}})$$
$$\|\mathfrak{E}_1\| = \|\mathfrak{E}_1\| \leq \frac{2\|B\|}{\sqrt{\lambda}}\|E\|$$

The error $\|E\|$ is controlled by Proposition 94 where it holds with probability greater or equal to $1-\delta$ that:

$$\|E\| \leq \frac{1}{\sqrt{N}}(\sqrt{2\sigma_1^2\log\frac{2}{\delta}}+\frac{2a\log\frac{2}{\delta}}{\sqrt{N}}) := \frac{1}{\sqrt{N}}a_{N,\delta}.$$

Finally, we have shown that provided that $N \geq M(1,\lambda,\delta)$ then with probability greater than $1-\delta$ one has:

$$\|G_{M,N}-G_M\| \leq \frac{\|B\|}{\sqrt{N\lambda}}(2a_{N,\delta}+\sqrt{2\gamma_1\kappa}+2\gamma_1\frac{\lambda+\kappa}{\sqrt{N\lambda}})+\frac{1}{N\lambda}a_{N,\delta}^2.$$

□

Proposition 90. *Let* $0 \leq \lambda \leq \|A\|$ *and define* $M(d,\lambda,\delta) := \frac{128}{9}\log\frac{4Tr(A)}{\lambda\delta}(d\kappa\lambda^{-1}+1)$. *Under Assumption 3 and Assumption (F), for any* $\delta \geq 0$ *such that* $M \geq M(d,\lambda,\delta)$ *the following holds with probability* $1-\delta$*:*

$$\|G_M - G_W\| \lesssim \lambda^{\frac{1}{2+c}}$$

Proof. We consider the error $\|G_M - G_W\|$. Recall that G_W is given by $G_W = \Phi^\top\Phi$ with Φ defined in Proposition 85. Let κ be a positive real number, we know by

Assumption 3 that there exists $F^\kappa := (f_s^\kappa)_{1\leq s\leq q}$ with $f_s^\kappa \in \mathcal{H}$ such that $\|\Phi - F^\kappa\|_{L_2(\rho_\theta)} \leq C\kappa$ and $\|f_s^\kappa\|_{\mathcal{H}} \leq C\kappa^{-c}$ for some fixed positive constant C. Therefore, we use F^κ to control the error $\|G_M - G_W\|$. Let's call $E = \Phi \circ h_\theta - LF^\kappa$ We consider the following decomposition:

$$\begin{aligned} G_M - G_W =& (L^\top \Phi \circ h_\theta)^\top \mathcal{G}_M(A_\lambda) L^\top \Phi \circ h_\theta - \Phi^\top \Phi \\ =& \underbrace{E^\top L \mathcal{G}_M(A_\lambda) L^\top E}_{\mathfrak{E}_1} - \underbrace{E^\top E}_{\mathfrak{E}_2} \\ &+ \underbrace{F_\kappa^\top \left(L^\top L \mathcal{G}_M(A_\lambda) - I\right) L^\top \Phi \circ h_\theta}_{\mathfrak{E}_3} + \underbrace{E^\top L \left(\mathcal{G}_M(A_\lambda) L^\top L - I\right) F^\kappa}_{\mathfrak{E}_4} \end{aligned}$$

First we consider the term $\mathfrak{E}_1$ one simply has:

$$\|\mathfrak{E}_1\| \leq \kappa^2 \underbrace{\|LA_\lambda^{-\frac{1}{2}}\|}_{\leq 1} \underbrace{\|A_\lambda^{\frac{1}{2}} \mathcal{G}_M(A_\lambda) A_\lambda^{\frac{1}{2}}\|}_{\leq 1} \underbrace{\|A_\lambda^{-\frac{1}{2}} L^\top\|}_{\leq 1} \leq \kappa^2$$

The second term also satisfies $\|\mathfrak{E}_1\| \leq \kappa^2$ by definition of F_κ. For the last two terms $\mathfrak{E}_3$ and $\mathfrak{E}_4$ we use Lemma 92 which allows to control the operator norm of $L(\mathcal{G}_M(A_\lambda)L^\top L - I)$. Hence, for any $\delta \geq 0$ and M such that $M \geq M(d, \lambda, \delta)$ and for $\kappa \leq 1$ it holds with probability $1 - \delta$ that:

$$\|\mathfrak{E}_3\| \lesssim \sqrt{\lambda}\kappa^{-c}; \qquad \|\mathfrak{E}_4\| \lesssim \sqrt{\lambda}\kappa^{-c}$$

We have shown so far that $\|G_M - G_W\| \lesssim (\kappa^2 + 2\kappa^{-c}\sqrt{\lambda})$. One can further optimize over κ on the interval $[0, 1]$ to get a tighter bound. The optimal value in this case is $\kappa^* = \min(1, (c\lambda^{\frac{1}{2}})^{\frac{1}{2+c}})$. By considering $\lambda > 0$ such that $(c\lambda^{\frac{1}{2}})^{\frac{1}{2+c}}) \leq 1$, it follows directly that $\|G_M - G_W\| \lesssim \lambda^{\frac{1}{2+c}}$ which shows the desired result. □

Proposition 91. *Let* $0 \leq \lambda \leq \|A\|$ *and define* $M(d, \lambda, \delta) := \frac{128}{9} \log \frac{4Tr(A)}{\lambda\delta}(d\kappa\lambda^{-1} + 1)$. *Under Assumption 4 and Assumption (F), for any* $\delta \geq 0$ *such that* $M \geq M(d, \lambda, \delta)$ *the following holds with probability* $1 - \delta$*:*

$$\|G_M - G_W\| \lesssim \lambda^{\min(1, \alpha + \frac{1}{2})}$$

Proof. Recall that G_W is given by $G_W = \Phi^\top \Phi$ with Φ defined in Proposition 85. By Assumption 4, we have that $\Phi = \nabla(A^\alpha V)$ with $V := (v_s)_{1\leq s\leq q} \in \mathcal{H}^q$. Hence, one can write

$$\begin{aligned} G_M - G_W &= (L^\top \Phi \circ h_\theta)^\top \mathcal{G}_M(A_\lambda) L^\top \Phi \circ h_\theta - \Phi^\top \Phi \\ &= V^\top (A^\alpha (A\mathcal{G}_M(A_\lambda)A - A)A^\alpha V \end{aligned}$$

we can therefore directly apply Lemma 92 and get $\|G_M - G_W\| \lesssim \lambda^{\min(1,\alpha+\frac{1}{2})}$ with probability $1-\delta$ for any $\delta \geq 0$ such that $M \geq M(d,\lambda,\delta)$. □

Lemma 92. *Let* $0 \leq \lambda \leq \|A\|$, $\alpha \geq 0$ *and define* $M(d,\lambda,\delta) := \frac{128}{9}\log\frac{4Tr(A)}{\lambda\delta}(d\kappa\lambda^{-1}+1)$. *Under Assumption* ***(F)****, for any* $\delta \geq 0$ *such that* $M \geq M(d,\lambda,\delta)$ *the following holds with probability* $1-\delta$*:*

$$\|L(\mathcal{G}_M(A_\lambda)L^\top L - I)A^\alpha\| \lesssim \lambda^{min(1,\alpha+\frac{1}{2})}$$

Proof. We have the following identities:

$$\begin{aligned} L(\mathcal{G}_M(A_\lambda)L^\top L - I)A^\alpha =& L(\mathcal{G}_M(A_\lambda)A_\lambda - I - \lambda\mathcal{G}_M(A_\lambda))A^\alpha \\ =& \underbrace{LA_\lambda^{-\frac{1}{2}}A_\lambda^{\frac{1}{2}}(\mathcal{G}_M(A_\lambda)A_\lambda P_M - I)A^\alpha}_{\mathfrak{E}_1} \\ &- \underbrace{\lambda LA_\lambda^{-\frac{1}{2}}A_\lambda^{\frac{1}{2}}\mathcal{G}_M(A_\lambda)A_\lambda^{\frac{1}{2}}A_\lambda^{-\frac{1}{2}}A^\alpha}_{\mathfrak{E}_3} \\ &+ \underbrace{LA_\lambda^{-\frac{1}{2}}A_\lambda^{\frac{1}{2}}\mathcal{G}_M(A_\lambda)A_\lambda^{\frac{1}{2}}A_\lambda^{\frac{1}{2}}(I-P_M)A^\alpha}_{\mathfrak{E}_2}. \end{aligned}$$

For the first $\mathfrak{E}_1$ we use [Sutherland et al., 2018, Lemma 1 (iii)] which implies that $\mathcal{G}_M(A_\lambda)A_\lambda P_M = P_M$. Thus $\mathfrak{E}_1 = LA_\lambda^{-\frac{1}{2}}A_\lambda^{\frac{1}{2}}(P_M - I)A^\alpha$. Moreover, by Lemma 93 we have that $\|A_\lambda^{\frac{1}{2}}(I-P_M)\| \leq 2\sqrt{\lambda}$ with probability $1-\delta$ for $M > M(d,\lambda,\delta)$. Therefore, recalling that $(I-P_M)^2 = I - P_M$ since P_M is a projection, one can

further write:

$$\|\mathfrak{E}_1\| \leq \underbrace{\|LA_\lambda^{-\frac{1}{2}}\|}_{\leq 1} \underbrace{\|A_\lambda^{\frac{1}{2}}(P_M - I)\|^2}_{\leq \lambda} \|A_\lambda^{-\frac{1}{2}} A^\alpha\|$$
$$\|\mathfrak{E}_2\| \leq \underbrace{\|LA_\lambda^{-\frac{1}{2}}\|}_{\leq 1} \underbrace{\|A_\lambda^{\frac{1}{2}} \mathcal{G}_M(A_\lambda) A_\lambda^{\frac{1}{2}}\|}_{\leq 1} \underbrace{\|A_\lambda^{\frac{1}{2}}(P_M - I)\|^2}_{\leq 4\lambda} \|A_\lambda^{-\frac{1}{2}} A^\alpha\|$$
$$\|\mathfrak{E}_3\| \leq \lambda \underbrace{\|LA_\lambda^{-\frac{1}{2}}\|}_{\leq 1} \underbrace{\|A_\lambda^{\frac{1}{2}} \mathcal{G}_M(A_\lambda) A_\lambda^{\frac{1}{2}}\|}_{\leq 1} \|A_\lambda^{-\frac{1}{2}} A^\alpha\|$$

It remains to note that $\|A_\lambda^{-\frac{1}{2}} A^\alpha\| \leq \lambda^{\alpha - \frac{1}{2}}$ when $0 \leq \alpha \leq \frac{1}{2}$ and that $\|A_\lambda^{-\frac{1}{2}} A^\alpha\| \leq \|A\|^{\alpha - \frac{1}{2}}$ for $\alpha > \frac{1}{2}$ which allows to conclude. □

B .4 Auxiliary Results

Lemma 93. *Let $0 \leq \lambda \leq \|A\|$. Under Assumption (F), for any $\delta \geq 0$ such that $M \geq M(d, \lambda, \delta) := \frac{128}{9} \log \frac{4Tr(A)}{\lambda \delta}(\kappa \lambda^{-1} + 1)$ the following holds with probability $1 - \delta$:*

$$\|A_\lambda^{\frac{1}{2}}(I - P_M)\| \leq 2\sqrt{\lambda}$$

Proof. The proof is an adaptation of the results in Rudi et al. [2015], Sutherland et al. [2018]. Here we recall Q_M defined in (7.20). Its transpose $Q_M^\top$ sends vectors in $\mathbb{R}^M$ to elements in the span of the Nyström basis points, hence P_M and $Q_M^\top$ have the same range, i.e.: $range(P_M) = range(Q_M^\top)$. We are in position to apply [Rudi et al., 2015, Proposition 3.] which allows to find an upper-bound on $A_\lambda^{\frac{1}{2}}(P_M - I)$ in terms of Q_M:

$$\|A_\lambda^{\frac{1}{2}}(P_M - I)\| \leq \sqrt{\lambda} \|A_\lambda^{\frac{1}{2}}(Q_M^\top Q_M + \lambda I)^{-\frac{1}{2}}\|.$$

For simplicity we write $\hat{A}_M := Q_M^\top Q_M$ and $E_2 := A_\lambda^{-\frac{1}{2}}(A - \hat{A}_M)A_\lambda^{-\frac{1}{2}}$. We also denote by $\beta = \lambda_{max}(E_2)$ the highest eigenvalue of E_2. We can therefore control $\|A_\lambda^{\frac{1}{2}}(\hat{A}_M + \lambda I)^{-\frac{1}{2}}\|$ in terms of β using [Rudi et al., 2015, Proposition 7] provided that $\beta < 1$:

$$\|A_\lambda^{\frac{1}{2}}(P_M - I)\| \leq \sqrt{\lambda} \frac{1}{\sqrt{1 - \beta}}.$$

Now we need to make sure that $\beta < 1$ for M large enough. To this end, we will apply [Rudi et al., 2015, Proposition 8.] to $\hat{A}_M$. Denote by $v_m = \sqrt{d}\partial_{i_m} k(Y_m, .)$. Hence, by definition of $\hat{A}_M$ it follows that $\hat{A}_M = \frac{1}{M}\sum_{m=1}^{M} v_m \otimes v_m$. Moreover, $(v_m)_{1\leq m\leq M}$ are independent and identically distributed and satisfy:

$$\mathbb{E}[v_m \otimes v_m] = \int \sum_{i=1}^{q} \partial_i k(y, .) \otimes \partial_i k(y, .)\, \mathrm{d}p_\theta(y) = A.$$

We also have by Assumption **(F)** that $\langle v_m, A_\lambda^{-1} v_m\rangle \leq \frac{d\kappa}{\lambda}$ almost surely and for all $\lambda > 0$. We can therefore apply [Rudi et al., 2015, Proposition 8.] which implies that for any $1 \geq \delta \geq 0$ and with probability $1-\delta$ it holds that:

$$\beta \leq \frac{2\gamma}{3M} + \sqrt{\frac{2\gamma d\kappa}{M\lambda}}$$

with $\gamma = \log \frac{4Tr(A)}{\lambda\delta}$ provided that $\lambda \leq \|A\|$. Thus by choosing $M \geq \frac{128\gamma}{9}(d\kappa\lambda^{-1}+1)$ we have that $\beta \leq \frac{3}{4}$ with probability $1-\delta$ which allows to conclude. □

Proposition 94. *There exist $a > 0$ and $\sigma_1 > 0$ such that for any $0 \leq \delta \leq 1$, it holds with probability greater of equal than $1-\delta$ that:*

$$\|\widehat{L^\top B} - L^\top B\| \leq \frac{2a\log\frac{2}{\delta}}{N} + \sqrt{\frac{2\sigma_1^2 \log\frac{2}{\delta}}{N}}$$

Proof. denote by $v_n = \sum_{i=1}^{d} \partial_i k(X_n, .) B_i(Z_n)$, we have that $\mathbb{E}[v_n] = L^\top B$. We will apply Bernstein's inequality for sum of random vectors. For this we first need to find $a > 0$ and $\sigma_1 > 0$ such that $\mathbb{E}[\|z_n - L^\top B\|_{\mathcal{H}}^p] \leq \frac{1}{2}p!\sigma_1^2 a^{p-2}$. To simplify

notations, we write x and x' instead of $h_\theta(z)$ and $h_\theta(z')$. We have that:

$$\begin{aligned}\mathbb{E}[\|z_n - L^\top B\|_{\mathcal{H}}^p] &= \int \left\| \sum_{i=1}^d \partial_i k(x,.) B_i(z) - \int \sum_{i=1}^d \partial_i k(x',.) B_i(z')\, \mathrm{d}\nu(z') \right\|^p \mathrm{d}\nu(z) \\ &\leq 2^{p-1} \underbrace{\int \left\| \sum_{i=1}^d \int \left(\partial_i k(x,.) - \partial_i k(x',.)\right) B_i(z)\, \mathrm{d}\nu(z') \right\|^p \mathrm{d}\nu(z)}_{\mathfrak{E}_1} \\ &\quad + 2^{p-1} \underbrace{\int \left\| \int \sum_{i=1}^d \partial_i k(x,.) \left(B_i(z) - B_i(z')\right) \mathrm{d}\nu(z') \right\|^p \mathrm{d}\nu(z)}_{\mathfrak{E}_2}\end{aligned}$$

We used the convexity of the norm and the triangular inequality to get the last line. We introduce the notation $\gamma_i(x) := \partial_i k(x,.) - \int \partial_i k(h_\theta(z'),.)\, \mathrm{d}\nu(z')$ and by $\Gamma(x)$ we denote the matrix whose components are given by $\Gamma(x)_{ij} := \langle \gamma_i(x), \gamma_j(x) \rangle_{\mathcal{H}}$. The first term $\mathfrak{E}_1$ can be upper-bounded as follows:

$$\begin{aligned}\mathfrak{E}_1 &= \int \left| Tr(B(z)B(z)^\top \Gamma(x)) \right|^{\frac{p}{2}} \\ &\leq \int \left| \|B(z)\|^2 \, Tr(\Gamma(x)^2)^{\frac{1}{2}} \right|^{\frac{p}{2}}.\end{aligned}$$

Moreover, we have that $Tr(\Gamma(x)^2)^{\frac{1}{2}} = (\sum_{1 \leq i,j \leq d} \langle \gamma_i(x), \gamma_j(x) \rangle_{\mathcal{H}}^2)^{\frac{1}{2}} \leq \sum_{i=1}^d \|\gamma_i(x)\|^2$. We further have that $\|\gamma_i(x)\| \leq \partial_i \partial_{i+d} k(x,x)^{\frac{1}{2}} + \int \partial_i \partial_{i+d} k(h_\theta(z), h_\theta(z))^{\frac{1}{2}}\, \mathrm{d}\nu(z)$ and by Assumption **(F)** it follows that $\|\gamma_i(x)\| \leq 2\sqrt{\kappa}$. Hence, one can directly write that: $\mathfrak{E}_1 \leq (2\sqrt{\kappa d})^p \int \|B(z)\|^p \, \mathrm{d}\nu(z)$. Recalling Assumptions **(B)** and **(C)** we get:

$$\mathfrak{E}_1 \leq 2^{p-1} (2\sqrt{\kappa d})^p C(\theta)^p (1 + \frac{1}{2} p! \zeta^{p-2} \sigma^2)$$

Similarly, we will find an upper-bound on $\mathfrak{E}_2$. To this end, we introduce the matrix $Q(x', x")$ whose components are given by $Q(x', x")_{i,j} = \partial_i \partial_{i+d} k(x', x")$. One,

therefore has:

$$\begin{aligned}\mathfrak{E}_2 &= \int \left| \int \int Tr((B(z) - B(z'))\,(B(z) - B(z"))^\top\, Q(x', x")\,\mathrm{d}\nu(z')\,\mathrm{d}\nu(z") \right|^{\frac{p}{2}} \mathrm{d}\nu(z) \\ &\leq \int \left| \int \int \|B(z) - B(z')\|\,\|B(z) - B(z")\|\, Tr(Q(x', x")^2)^{\frac{1}{2}}\,\mathrm{d}\nu(z')\,\mathrm{d}\nu(z") \right|^{\frac{p}{2}} \mathrm{d}\nu(z)\end{aligned}$$

Once again, we have that

$$Tr(Q(x', x")^2)^{\frac{1}{2}} \leq (\sum_{i=1}^{d} \partial_i \partial_{i+d} k(x', x'))^{\frac{1}{2}} (\sum_{i=1}^{d} \partial_i \partial_{i+d} k(x", x"))^{\frac{1}{2}} \leq d\kappa$$

thanks to Assumption **(F)**. Therefore, it follows that:

$$\begin{aligned}\mathfrak{E}_2 &\leq (\sqrt{d\kappa})^p \int | \int \|B(z) - B(z')\|\,\mathrm{d}\nu(z)|^p\,\mathrm{d}\nu(z) \\ &\leq 3^{p-1}(\sqrt{d\kappa})^p C(\theta)^p (2^p + \int \|z\|^p\,\mathrm{d}\nu(z) + \left(\int \|z\|\,\mathrm{d}\nu(z)\right)^p) \\ &\leq 3^{p-1}(\sqrt{d\kappa})^p C(\theta)^p (2^p + \frac{1}{2} p! \zeta^{p-2}\sigma^2 + \left(\int \|z\|\,\mathrm{d}\nu(z)\right)^p).\end{aligned}$$

The second line is a consequence of Assumption **(C)** while the last line is due to Assumption **(B)**. These calculations, show that it is possible to find constants a and σ_1 such that $\mathbb{E}[\|z_n - L^\top B\|_{\mathcal{H}}^p] \leq \frac{1}{2} p! \sigma_1^2 a^{p-2}$. Hence one concludes using Bernstein's inequality for a sum of random vectors [see for instance Rudi et al., 2015, Proposition 11]. □

C Connection to the Negative Sobolev distance

To obtain the Wasserstein natural gradient, one can exploit a Taylor expansion of W which is given in terms of the *Negative Sobolev distance* $\|\rho_{\theta+u} - \rho_\theta\|_{H^{-1}(\rho_\theta)}$ as done in Mroueh et al. [2019]:

$$W_2^2(\rho_\theta, \rho_{\theta+u}) = \|\rho_{\theta+u} - \rho_\theta\|_{H^{-1}(\rho_\theta)}^2 + o(\|u\|^2).$$

Further performing a Taylor expansion of $\|\rho_{\theta+u} - \rho_\theta\|_{H^{-1}(\rho_\theta)}$ in u leads to a quadratic term $u^\top G_W(\theta_t) u$ where we call $G_W(\theta_t)$ the *Wasserstein information matrix*. This

two steps approach is convenient conceptually and allows to use the dual formulation of the *Negative Sobolev distance* to get an estimate of the quadratic term $u^\top G_W(\theta_t)u$ using kernel methods as proposed in Mroueh et al. [2019] for learning non-parametric models. However, with such approach, $\|\rho_{\theta+u} - \rho_\theta\|_{H^{-1}(\rho_\theta)}$ needs to be well defined for u small enough. This requirement does not exploit the parametric nature of the problem and can be restrictive if $\rho_{\theta+u}$ and ρ_θ do not share the same support as we discuss now.

As shown in [Villani, 2003, Theorem 7.26] and discussed in Theorem 71 and Proposition 72, the Wasserstein distance between two probability distributions ρ and ρ' admits a first order expansion in terms of the Negative Sobolev Distance:

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon} W_2(\rho, \rho + \epsilon(\rho' - \rho)) = \|\rho - \rho'\|_{H^{-1}(\rho)}$$

when ρ' admits a bounded density w.r.t. ρ. When such assumption fails to hold, there are cases when this first order expansion is no longer available. For instance, in the simple case when the parametric family consists of dirac distributions δ_θ located at a value θ, the Wasserstein distance admits a closed form expression of the form:

$$W_2(\delta_\theta, \delta_\theta + \epsilon(\delta_{\theta'} - \delta_\theta)) = \sqrt{\epsilon}\|\theta - \theta'\|$$

Hence, $\frac{1}{\epsilon} W_2(\delta_\theta, \delta_\theta + \epsilon(\delta_{\theta'} - \delta_\theta))$ diverges to infinity. One can consider a different perturbation of the model $\delta_{\theta+\epsilon u}$ for some vector u which the one we are interested in here. In this case, the Wasserstein distance admits a well-defined asymptotic behavior:

$$\lim_{\epsilon\to} \frac{1}{\epsilon} W_2(\delta_\theta, \delta_{\theta+\epsilon u}) = \|u\|.$$

On the other hand the Negative Sobolev Distance is infinite for any value of ϵ. To see this, we consider its dual formulation as in Mroueh et al. [2019]:

$$\frac{1}{2}\|\rho - \rho'\|_{H^{-1}(\rho)} = \sup_{\substack{f\in C_c^\infty(\Omega) \\ \int f(x)\,\mathrm{d}\rho(x)=0}} \int f(x)\,\mathrm{d}\left(\rho - \rho'\right)(x) - \frac{1}{2}\int \|\nabla_x f(x)\|^2\,\mathrm{d}\rho(x)$$

Evaluating this expression for δ_θ and $\delta_{\theta+\epsilon u}$ for any value $\epsilon > 0$ and for any u that is non zero, on has:

$$\frac{1}{2\epsilon}\|\delta_\theta - \delta_{\theta+\epsilon u}\|_{H^{-1}(\delta_\theta)} = \sup_{\substack{f\in C_c^\infty(\Omega)\\ f(\theta)=0}} \frac{1}{\epsilon}(f(\theta) - f(\theta+\epsilon u)) - \frac{1}{2}\|\nabla_x f(\theta)\|^2 \quad (7.25)$$

One can always find a function f such that $\nabla f(\theta) = 0$, $f(\theta) = 0$ and $-f(\theta + \epsilon u)$ can be arbitrarily large, thus the Negative Sobolev distance is infinite. This is not the case of the metric $u^\top G_W(\theta)u$ which can be computed in closed form:

$$\frac{1}{2}u^\top G_W(\theta)u = \sup_{\substack{f\in C_c^\infty(\Omega)\\ f(\theta)=0}} \nabla f(\theta)^\top u - \frac{1}{2}\|\nabla_x f(\theta)\|^2 \quad (7.26)$$

In this case, choosing $f(\theta) = 0$ and $\nabla f(\theta) = u$ achieves the supremum which is simply given by $\frac{1}{2}\|u\|^2$. Equation (7.26) can be seen as a limit case of (7.25) when $\epsilon \to 0$:

$$\frac{1}{2}u^\top G_W(\theta)u := \sup_{\substack{f\in C_c^\infty(\Omega)\\ f(\theta)=0}} \lim_{\epsilon\to 0} \frac{1}{\epsilon}(f(\theta) - f(\theta+\epsilon u)) - \frac{1}{2}\|\nabla_x f(\theta)\|^2$$

However, the order between the supremum and the limit cannot be exchanged in this case, which makes the two objects behave very differently in the case of singular probability distributions.

D Expression of WNG for the Multivariate Gaussian

Multivariate Gaussian. Consider a multivariate gaussian with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^d \times \mathbb{R}^d$ parametrized using its lower triangular components $s = T(\Sigma)$. We denote by $\Sigma = T^{-1}(s)$ the inverse operation that maps any vector $s \in \mathbb{R}^{\frac{d(d+1)}{2}}$ to its corresponding symmetric matrix in $\mathbb{R}^d \times \mathbb{R}^d$. The concatenation of the mean μ and s will be denoted as $\theta : \theta = (\mu, s)$. Given two parameter vectors $u = (m, T(S))$ and $v = (m', T(S'))$ where m and m' are vectors in $\mathbb{R}^d$ and S and

S' are symmetric matrices in $\mathbb{R}^d \times \mathbb{R}^d$ the metric evaluated at u and v is given by:

$$u^\top G(\theta) v = m^\top m' + Tr(A\Sigma A')$$

where A and A' are symmetric matrices that are solutions to the Lyapunov equation:

$$S = A\Sigma + \Sigma A, \qquad S' = A'\Sigma + \Sigma A'.$$

A and A' can be computed in closed form using standard routines making the evaluation of the metric easy to perform. Given a loss function $\mathcal{L}(\theta)$ and gradient direction $\nabla_\theta \mathcal{L}(\theta) = \nabla_\mu \mathcal{L}(\theta), \nabla_s \mathcal{L}(\theta)$, the corresponding natural gradient $\nabla^W_\theta \mathcal{L}(\theta)$ can also be computed in closed form:

$$\nabla^W_\theta \mathcal{L}(\theta) = (\nabla_\mu \mathcal{L}(\theta), T(\Sigma(A + diag(A)) + (A + diag(A))\Sigma)),$$

where $A = T^{-1}(\nabla_s \mathcal{L}(\theta))$. To use the estimator proposed in Proposition 77 we take advantage of the parametrization of the Gaussian distribution as a push-forward of a standard normal vector:

$$X \sim \mathcal{N}(\mu, \Sigma) \iff X = \Sigma^{\frac{1}{2}} Z + \mu, Z \sim \mathcal{N}(0, I_d)$$

Part IV

Conclusion

In this chapter, we summarize our contributions and propose directions for future research.

Conditional score estimation

In Chapter 3, we have proposed a framework for conditional density estimation using a non-parametric class of models. These models take the form of an exponential family with natural parameters belonging to a vector-valued RKHS. We then extended the score matching procedure from Sriperumbudur et al. [2017] to estimate these natural parameters by solving a finite-dimensional convex problem. Finally, we provided convergence rates of the resulting estimators in the well-specified setting where the data distribution belongs to the class defined by the model. This work resulted in a general and flexible framework for estimating joint densities that factorize according to a graph, and has a clear computational advantage when the graph is space. When the graph is dense, this framework can still exploit a smooth dependence on the conditional variables, often resulting in an improved empirical performance over direct estimation of the joint density. In the following, we discuss two extensions of this framework.

Future works

Scalable conditional score estimation. To approximate a conditional density $p(y|x)$, the proposed estimator requires solving a linear system of size $N \times d_y$ with d_y being the dimension of y. Hence, the memory and time complexities are $N^2 \times d_y^2$ and $N^3 \times d_y^3$, which is still prohibitive for large sample size N and even moderate dimensions d_y. This cost could be reduced without compromising the convergence rate using Nyström method as done in Chapter 7 and Sutherland et al. [2018]. This method would reduce the complexity in memory and in time to $O(NMd_y^2)$ and $O(NM^2d_y^3)$. The dependence in the dimension d_y results from evaluating all partial derivatives of the kernel. Using an approach similar to Chapter 7, this dependence in the dimension d_y can become linear by randomly subsampling the partial derivatives with respect to the M Nyström points. We expect these improvements to come without compromising the convergence rate, provided the number of basis points M

scales with the sample size as a suitable optimal rate, usually between $N^{\frac{1}{2}}$ and $N^{\frac{1}{3}}$ in the well-specified setting.

Learning graphical structure under sparsity constraints. The framework of Chapter 3 assumes the graphical structure to be known. Moreover, unless the graph is directed and acyclic, the procedure doesn't guarantee that estimating each conditional density independently results in a consistent probability distribution. A possible extension is to learn a graphical structure by imposing sparsity constraints similarly to Sun et al. [2015]. In this case the logarithm of the density can be decomposed as a sum of a constant term C and unary and pairwise terms:

$$\log p(x) = C + \sum_{1 \leq i \leq d} T_i(x_i) + \sum_{\substack{1 \leq i,j \leq d \\ i \neq j}} T_{ij}(x_i, x_j).$$

The unary function T_i would belong to an RKHS $\mathcal{H}_i$, while the pairwise function $x_j \mapsto T_{i,j}(., x_j)$ would belong to a particular vector-valued RKHS taking values in $\mathcal{H}_i$. Hence, similarly to Chapter 3, it is possible to fit each conditional distribution using the conditional score. However, instead of the usual ridge regression, using a joint sparsity-inducing penalty as in Rakotomamonjy et al. [2011] would encourage only a small number of non-trivial functions $T_{ij}(x_i, x_j)$ while setting the others to 0. Independently learning each conditional would still result in a computational gain compared to Sun et al. [2015]. However, further investigation is needed to provide guarantees for the estimation and recovery of the sparsity structure.

Structuring and Regularizing implicit models

We considered the problem of estimating distributions under a low-intrinsic dimension assumption and using Implicit Generative models (IGMs). To address the stability problem in IGMs, we introduced a self-regularizing loss for learning these models using the Maximum Mean Discrepancy in Chapter 4. We then provided practical conditions on the kernel's parametrization that ensure continuity of the loss in the models' parameters. We empirically showed that this loss increases training stability and yields state-of-the-art results in image generation tasks. This

method resulted in a robust and generic approach to learning IGMs. In Chapter 5, we introduced a new model that augments IGMs with an explicit model using an importance sampling strategy. This hybrid model enables the use of expressive latent noise distributions at the same training cost compared to IGMs. We have shown how to train the explicit component using a generalized maximum likelihood relative to the support of the IGM, thus taking advantage of a stronger topology of convergence during training. We then proposed a simple MCMC scheme in latent space to sample from such models. This resulted in a flexible class of models generalizing both IGMs and Energy-Based Models, that is suited for modeling data with low intrinsic dimension.

Future work

Regularity of the loss. We have shown in Chapter 4 the self-regularizing loss to be continuous in the weak topology of convergence of measures when viewed as a divergence between probability distributions. However, when using gradient methods during optimization, this condition is insufficient to guarantee convergence towards a local optimum. In Chapter 5, we established the loss to be L-Lipschitz and weakly convex under practical conditions on the critic functions and IGM models. The resulting regularity of the loss is enough to guarantee convergence of gradient methods to local optima. However, this weak regularity of the loss can result in slower convergence compared to losses that are L-smooth.

The recent work of Chu et al. [2020] proposed a general framework for regularizing the discriminator used for adversarial training so that the loss used to learn the IGM is L-smooth. This framework guarantees local convergence of gradient descent methods provided the optimal discriminator is computed exactly. Most of the proposed regularizations can be practically achieved using existing methods such as spectral normalization Miyato et al. [2018]. However, one of these conditions in [Chu et al., 2020, Section 6] requires the set of critic functions to belong to a smooth RKHS space such as the one defined by a gaussian kernel and to control their RKHS norm. Unfortunately, this condition is rarely satisfied in practice when the critic function is parameterized by a Deep Neural network, as often done in

practice. Identifying more practical conditions to ensure the L-smoothness of the loss would result in faster convergence guarantees. The work of Bietti et al. [2018], Bietti and Mairal [2017] provides a first insight on this question by viewing deep neural networks as functions in an abstract RKHS space whose smoothness can be controlled by their RKHS norm. Further investigation is required to understand if such RKHS penalty could yield L-smooth losses for IGMs and devise efficient methods for estimating it.

Model miss-specification. The support defined by an IGM is, in general, a *rectifiable set* Federer [2014], a notion that extends smooth manifolds to piecewise smooth sets and thus also possesses a tangent structure defined almost everywhere. However, it remains unclear whether this notion can correctly represent the support of distributions such as images for which the geometrical properties are not yet fully understood [Bartholdi et al., 2012, Introduction].

Generalization in implicit models. We focused on developing models for data with a low intrinsic dimension and methods for learning them without accounting for their generalization capabilities. The work of Uppal et al. [2019] considered the question of generalization under Integral Probability Metrics. It provided a minimax optimal statistical rate of convergence that depends on both smoothness of the target distribution and the smoothness of the critic functions. However, their analysis deals only with distributions with a well-defined density. Therefore, extending it to distributions with a low intrinsic dimension would better reflect the practical setting. In the context of regression using Deep Neural Networks, the encouraging result in Nakada and Imaizumi [2020] shows their proposed estimator to converge at rates that depend only on the Minkowski dimension of the data support. Extending such result to the case of IGMs or GEBMs is a promising avenue for future research.

Optimization of implicit generative models

We now discuss the contributions in Chapters 6 and 7 along with possible future work directions.

Wasserstein Gradient flow of weak divergence functionals

Contribution. We considered the Wasserstein gradient flow of the MMD functional as a simplified setting for analyzing the optimization properties in the context of learning IGMs. Our study showed that even this simplified setting is highly non-convex and convergence to a global solution can fail dramatically. We provided a criterion on the smoothness of the optimization trajectories that ensured global convergence of the flow and introduced an algorithm based on noise-injection to improve convergence.

Trajectory-independent criteria for convergence. Further investigation is required to understand if the trajectory dependent criterion for global convergence can be relaxed into a criterion depending only on the initial condition:

$$\text{If } \rho_0 \in \mathcal{P}^\star \Rightarrow \rho_t \to \nu^\star,$$

here, $\mathcal{P}^\star$ would represent a *basin* of attraction by analogy to finite dimensional optimization. To the best of our knowledge, identifying such a set is still an open question.

Choice of the kernel. As shown in the experiments of Section 5 .2, using an MMD with a Gaussian kernel yields optimization trajectories that are likely to fail reaching the global solution. Hence, another interesting question would be to identify *well-behaved* characteristic kernels for which the Wasserstein gradient flow has better convergence properties.

Wasserstein natural gradient

Contribution. We introduced a scalable estimator of the Wasserstein natural gradient in Chapter 7. To the best of our knowledge, this is the first time a practical and consistent estimator of the Wasserstein Natural Gradient is derived. This estimator is scalable to large models and allows to exploit the Wasserstein geometry. This work results in optimization methods that are robust to the model's parametrization and that do not require a well-defined density, thus making them particularly well suited for IGMs. We discuss below a few avenues for future work.

High dimensional distributions. We provided convergence rates in Section 3 .4 for the estimator of the Wasserstein natural gradient that is adaptive to the smoothness of the transport map. However, it remains open what regularity assumptions on the model can guarantee a smooth transport map.

The effect of the dimension on the estimator's quality also appears in the minimal number of basis points and samples that yield a consistent estimator. Such a number exhibits a polynomial dependence on the dimension regardless of the smoothness of the transport map. A possible research direction is to mitigate this dependence when an additional structure in the model is available, such as conditional independence between the dimension as done in Chapter 3.

Significance.

We showed the benefit of using the WNG to be even more crucial when the optimization problem is ill-conditioned, as illustrated in an artificial setting (Section 4). An important question is to identify more natural settings where WNG is beneficial. Recent developments already hint to use cases in the context of Reinforcement Learning in Moskovitz*, Ted et al. [2021]. However, a precise quantification of the speed of convergence of the optimization remains open.

Wasserstein Natural gradient for Reinforcement Learning. In [Moskovitz*, Ted et al., 2021], we proposed a framework for using the Wasserstein natural gradient in the context of reinforcement learning. This framework takes inspiration from Pacchiano et al. [2019] and introduces a Policy Optimization algorithm based on the Wasserstein Natural gradient to capture the 'behavioral' geometry of a policy. This approach resulted in improved learning efficiency compared to well-established algorithms such as TRPO Schulman et al. [2015] or Proximal Policy Optimization Schulman et al. [2017] both based on the Kullback-Liebler divergence between policies as a proximity measure. The proposed method relies on a pre-defined representation of the 'behavior' of a policy to compute the natural gradient. Defining general principle for learning such representations could result in improved efficiency.

Bibliography

Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev Descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2976–2985, April 2019.

Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Global convergence of neuron birth-death dynamics. In *ICML*, 2019.

David L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In *AMS CONFERENCE ON MATH CHALLENGES OF THE 21ST CENTURY*, 2000.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. 6:695–709, 2005. doi: 10.1.1.109.4126.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, August 2002. ISSN 0899-7667.

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

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, and Michael I Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *Advances in neural information processing systems*, pages 4116–4124, 2016.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794.

Ph Rigollet and Alexander B Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16(3):260–280, 2007.

John Shawe-Taylor and Nello Cristianini. *Support vector machines*, volume 2. Cambridge University Press Cambridge, 2000.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

C. Gu and C. Qiu. Smoothing spline density estimation: Theory. 21(1):217–234, 1993.

A. Barron and C-H. Sheu. Approximation of density functions by sequences of exponential families. 19(3):1347–1369, 1991.

G. Pistone and C. Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. 23(5):1543–1561, 1995.

Stephane Canu and Alex J. Smola. Kernel methods and the exponential family. 69 (7):714–720, 2006.

Kenji Fukumizu. Exponential manifold by reproducing kernel Hilbert spaces. In *Algebraic and Geometric Methods in Statistics*, pages 291–306. Cambridge University Press, 2009.

Bharath Sriperumbudur, Kenji Fukumizu, Revant Kumar, Arthur Gretton, and Aapo Hyvärinen. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 12 2017.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2001.

Michael I. Jordan. *Learning in Graphical Models*. MIT Press, 1999. ISBN 0-262-60032-3.

Charles A. Micchelli and Massimiliano A. Pontil. On learning vector-valued functions. 17(1):177–204, 01 2005. ISSN 0899-7667. doi: 10.1162/0899766052530802.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017.

Tengyuan Liang. How well can generative adversarial networks learn densities: A nonparametric view. *arXiv preprint arXiv:1712.08244*, 2017.

Rob Cornish, Anthony L. Caterini, George Deligiannidis, and Arnaud Doucet. Relaxing bijectivity constraints with continuously indexed normalising flows. 2020.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *NIPS*, 2016.

A. Müller. Integral probability metrics and their generating classes of functions. 29 (2):429–443, 1997.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2017. Curran Associates Inc.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018a.

Y. Jin, K. Zhang, M. Li, Y. Tian, H. Zhu, and Z. Fang. Towards the automatic anime characters creation with generative adversarial networks, 2017.

Vaishnavh Nagarajan and J. Zico Kolter. Gradient descent gan optimization is locally stable. 06 2017.

Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On Convergence and Stability of GANs. *arXiv:1705.07215 [cs]*, May 2017. arXiv: 1705.07215.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Cedric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Jonathan Weed, Francis Bach, et al. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A): 2620–2648, 2019.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 2012.

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 30*, pages 2203–2213. Curran Associates, Inc., 2017.

Mikołaj Bińkowski*, Danica J. Sutherland*, **Arbel, Michael**, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Olivier Bousquet, Olivier Chapelle, and Matthias Hein. Measure based regularization. In *NIPS*. 2004.

Samaneh Azadi, Catherine Olsson, Trevor Darrell, Ian Goodfellow, and Augustus Odena. Discriminator rejection sampling. In *International Conference on Learning Representations*, 2019.

Ryan Turner, Jane Hung, Eric Frank, Yunus Saatchi, and Jason Yosinski. Metropolis-Hastings generative adversarial networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6345–6353, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Kirill Neklyudov, Evgenii Egorov, and Dmitry Vetrov. The implicit metropolis-hastings algorithm. 2019.

HE Daniels. The asymptotic efficiency of a maximum likelihood estimator. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 151–163. University of California Press Berkeley, 1961.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Shun-ichi Amari. *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics. Springer-Verlag, New York, 1985. ISBN 978-0-387-96056-2.

Andrew Holbrook, Shiwei Lan, Jeffrey Streets, and Babak Shahbaba. The nonparametric Fisher geometry and the chi-square process density prior. *arXiv:1707.03117 [stat]*, July 2017. arXiv: 1707.03117.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

F. Otto and C. Villani. Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 173(2): 361–400, June 2000. ISSN 00221236. doi: 10.1006/jfan.1999.3557.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Gradient flows with metric and differentiable structures, and applications to the Wasserstein space. *Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei. Matematica e Applicazioni*, 15(3-4):327–343, 2004. ISSN 1120-6330.

Filippo Santambrogio. Gradient flows in wasserstein spaces and applications to crowd movement. *Séminaire Équations aux dérivées partielles (Polytechnique)*, pages 1–16, 2010.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. NIPS, 2018a.

José A. Carrillo, Robert J. McCann, and Cédric Villani. Contractions in the 2-Wasserstein Length Space and Thermalization of Granular Media. *Archive for Rational Mechanics and Analysis*, 179(2):217–263, February 2006. ISSN 1432-0673. doi: 10.1007/s00205-005-0386-1.

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.

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.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. G. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *NIPS*, 2009.

Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, February 1998. ISSN 0899-7667. doi: 10.1162/089976698300017746.

Mohammad Emtiyaz Khan and Wu Lin. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. *arXiv preprint arXiv:1703.04265*, 2017.

Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. Noisy natural gradient as variational inference. In *International Conference on Machine Learning*, pages 5852–5861. PMLR, 2018.

Yann Ollivier, Ludovic Arnold, Anne Auger, and Nikolaus Hansen. Information-geometric optimization algorithms: A unifying picture via invariance principles. *J. Mach. Learn. Res.*, 18:18:1–18:65, 2011.

Wuchen Li and Guido Montufar. Natural gradient via optimal transport. *arXiv:1803.07033 [cs, math]*, March 2018a. arXiv: 1803.07033.

Wuchen Li and Guido Montufar. Ricci curvature for parametric statistics via optimal transport. *arXiv:1807.07095 [cs, math, stat]*, July 2018b. arXiv: 1807.07095.

Wuchen Li. Geometry of probability simplex via optimal transport. *arXiv:1803.06360 [math]*, March 2018. arXiv: 1803.06360.

James Martens and Roger Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. *arXiv:1503.05671 [cs, stat]*, March 2015. arXiv: 1503.05671.

Roger Grosse and James Martens. A Kronecker-factored Approximate Fisher Matrix for Convolution Layers. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 573–582. JMLR.org, 2016. event-place: New York, NY, USA.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. Fast Approximate Natural Gradient Descent in a Kronecker-factored Eigenbasis. *arXiv:1806.03884 [cs, stat]*, June 2018. arXiv: 1806.03884.

Tom Heskes. On "Natural" Learning and Pruning in Multilayered Perceptrons. *Neural Computation*, 12(4):881–901, April 2000. ISSN 0899-7667. doi: 10.1162/089976600300015637.

Alberto Bernacchia, Mate Lengyel, and Guillaume Hennequin. Exact natural gradient in deep linear networks and its application to the nonlinear case. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5941–5950. Curran Associates, Inc., 2018.

Wuchen Li, Alex Tong Lin, and Guido Montufar. Affine natural proximal learning. February 2019.

Rémi Peyre. Comparison between W_2 distance and $\dot{H}^{-1}$ norm, and localisation of Wasserstein distance. *ESAIM: Control, Optimisation and Calculus of Variations*, 24(4):1489–1501, 2018.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Soc., 2003. ISBN 978-0-8218-3312-4. Google-Books-ID: R_nWqjq89oEC.

Arbel, Michael and Arthur Gretton. Kernel Conditional Exponential Family. *In AISTATS*, 2018.

Arbel*, Michael, Sutherland*, Danica J., Mikołaj Bińkowski, and Arthur Gretton. On gradient regularizers for MMD GANs. *In NeurIPS*, 2018.

Arbel, Michael, Liang Zhou, and Arthur Gretton. Generalized energy based models. In *ICLR*, 2021.

Arbel, Michael, Anna Korba, Adil Salim, and Arthur Gretton. Maximum Mean Discrepancy Gradient Flow. *In NeurIPS*, 2019a.

Arbel, Michael, Arthur Gretton, Wuhen Li, and Guido Montufar. Kernelized Wasserstein Natural Gradient. *In ICLR*, 2019b.

Moskovitz*, Ted, **Arbel*, Michael**, Ferenc Huszar, and Arthur Gretton. Efficient wasserstein natural gradients for reinforcement learning. In *ICLR*, 2021.

Danica Sutherland, Heiko Strathmann, **Arbel, Michael**, and Arthur Gretton. Efficient and principled score estimation with Nystrom kernel exponential families. In *International Conference on Artificial Intelligence and Statistics*, pages 652–660, March 2018.

Tolga Birdal, **Arbel, Michael**, Umut Simsekli, and Leonidas J Guibas. Synchronizing probability measures on rotations via optimal transport. In *In CVPR*, 2020.

Anna Korba, Adil Salim, **Arbel, Michael**, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for stein variational gradient descent. *In NeurIPS*, 2020.

Louis Thiry, **Arbel, Michael**, Eugene Belilovsky, and Edouard Oyallon. The unreasonable effectiveness of patches in deep convolutional kernels methods. In *ICLR*, 2021.

Samuel Cohen, **Arbel, Michael**, and Marc Peter Deisenroth. Estimating barycenters of measures in high dimensions. *Under review*, 2020.

Alex J Smola and Bernhard Scholkopf. *Learning with kernels*, volume 4. Citeseer, 1998.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55:58–63, 2015.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84 (3):375–393, January 2000. ISSN 0945-3245. doi: 10.1007/s002110050002.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Cedric Villani. Trend to equilibrium for dissipative equations, functional inequalities and mass transportation. *Contemporary Mathematics*, 353:95, 2004.

Radford M. Neal. Mcmc using hamiltonian dynamics. 06 2010.

C.E. Rasmussen. Gaussian Processes to Speed up Hybrid Monte Carlo for Expensive Bayesian Integrals. pages 651–659, 2003.

Heiko Strathmann, Dino Sejdinovic, Samuel Livingstone, Zoltan Szabo, and Arthur Gretton. Gradient-free hamiltonian monte carlo with efficient kernel exponential families. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 955–963. Curran Associates, Inc., 2015.

Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. Least-squares conditional density estimation. 93-D(3):583–594, 2010.

Jianqing Fan, Qiwei Yao, and Howell Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. (6704), 03 1996.

Peter Hall, Rodney C. L. Wolff, and Qiwei Yao. Methods for estimating a conditional distribution function. 94(445):154–163, 1999. ISSN 01621459.

Thomas Nagler and Claudia Czado. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. 151(C):69–89, 10 2016. ISSN 0047-259X. doi: 10.1016/j.jmva.2016.07.003.

L. Song, B. Boots, S. Siddiqi, G. Gordon, and A. Smola. Hilbert space embeddings of hilbert markov models. In *International Conference on Machine Learning (ICML)*, 2010.

S. Grunewalder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors. In *International Conference on Machine Learning (ICML)*, 2012.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

Y. Chen, M. Welling, and A. Smola. Supersamples from kernel-herding. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010.

F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In *ICML*, 2012.

M. Kanagawa, Y. Nishiyama, A. Gretton, and K. Fukumizu. Filtering with state-observation examples via kernel monte carlo filter. 28(2):382–444, 2016.

Benigno Uria, Iain Murray, and Hugo Larochelle. Rnade: The real-valued neural autoregressive density-estimator. 06 2013.

Tapani Raiko, Li Yao, Kyunghyun Cho, and Yoshua Bengio. Iterative neural autoregressive distribution estimator (nade-k). 06 2014.

Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. 17, 05 2016.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006. ISBN 0387310738.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. 7(3):331–368, 2007. ISSN 1615-3383. doi: 10.1007/s10208-006-0196-8.

R Development Core Team. *The R Manuals*, 2008.

Mixed Cumulative Distribution Networks, 08 2011.

Yichuan Tang, Ruslan Salakhutdinov, and Geoffrey Hinton. Deep mixtures of factor analysers. In *Proceedings of the 29th International Conference on Machine Learning, 2012, Edinburgh, Scotland*, 06 2012.

W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton. A test of relative similarity for model selection in generative models. In *ICLR*, 2016.

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. 10:377–408, 10 2006. doi: 10.1142/S0219530506000838.

J. R. Retherford. Review: J. diestel and j. j. uhl, jr., vector measures. 84(4):681–685, 07 1978.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. 06 2008.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *NIPS*, 2016.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. 102(477):359–378, 2007.

Léon Bottou, Martin Arjovsky, David Lopez-Paz, and Maxime Oquab. *Geometrical Insights for Implicit Generative Modeling*, pages 229–268. LNAI Vol. 11100. Springer, 2018.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *AISTATS*, 2018.

Yujia Li, Kevin Swersky, and Richard Zemel. Generative moment matching networks. *arXiv preprint arXiv:1502.02761*, 2015.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015.

Thomas Unterthiner, Bernhard Nessler, Calvin Seward, Günter Klambauer, Martin Heusel, Hubert Ramsauer, and Sepp Hochreiter. Coulomb GANs: Provably optimal nash equilibria via potential fields. In *ICLR*, 2018.

M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The Cramer distance as a solution to biased Wasserstein gradients, 2017.

Youssef Mroueh, Chung-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. In *ICLR*, 2018.

Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *NIPS*, 2017.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *ICML*, 2018.

Bharath Sriperumbudur. On the optimal estimation of probability mesaures in weak and strong topologies. 22(3):1839–1893, 2016.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. 12:2389–2410, 2011.

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *ICLR*, 2017.

Paul Milgrom and Ilya Segal. Envelope Theorems for Arbitrary Choice Sets. *Econometrica*, 70, 2002.

Richard M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2 edition, 2002.

H. Wendland. *Scattered Data Approximation*. Cambridge University Press, 2005.

Youssef Mroueh and Tom Sercu. Fisher GAN. In *NIPS*, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015a.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C.

Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *arXiv:1409.0575 [cs]*, September 2014. arXiv: 1409.0575.

W. Zaremba, A. Gretton, and M. B. Blaschko. B-tests: Low variance kernel two-sample tests. In *NIPS*, 2013.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception architecture for computer vision. In *CVPR*, 2016.

Shane Barratt and Rishi Sharma. A note on the inception score, 2018.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30*, pages 6626–6637. Curran Associates, Inc., 2017.

Gao Huang, Yang Yuan, Qiantong Xu, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks, 2018b.

D. Berthelot, T. Schumm, and L. Metz. BEGAN: Boundary equilibrium generative adversarial networks, 2017.

Andrew Brock, Theodore Lim, J. M. Ritchie, and Nick Weston. Neural photo editing with Introspective Adversarial Networks. In *ICLR*, 2017.

A.Dilek Güngör. Some bounds for the product of singular values. 2007.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *ICML*, 2017.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu-Jie Huang. *Predicting Structured Data*, chapter A Tutorial on Energy-Based Learning. MIT Press, 2006.

Bo Dai, Hanjun Dai, Arthur Gretton, Le Song, Dale Schuurmans, and Niao He. Kernel exponential family estimation via doubly dual embedding. In *Proceedings*

of Machine Learning Research, volume 89 of *Proceedings of Machine Learning Research*, pages 2321–2330. PMLR, 16–18 Apr 2019a.

Bo Dai, Zhen Liu, Hanjun Dai, Niao He, Arthur Gretton, Le Song, and Dale Schuurmans. Exponential Family Estimation via Adversarial Dynamics Embedding. *arXiv:1904.12083 [cs, stat]*, December 2019b. arXiv: 1904.12083.

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. 28(1):1–47, 1975. ISSN 1097-0312. doi: 10.1002/cpa.3160280102. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.3160280102.

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

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. *ICLR*, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286, 2014.

Casey Chu, Kentaro Minami, and Kenji Fukumizu. Smoothness and stability in gans. In *International Conference on Learning Representations*, 2020.

Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. On the convergence and robustness of training gans with regularized optimal transport. In *Advances in Neural Information Processing Systems 31*, pages 7091–7101. Curran Associates, Inc., 2018.

Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 2017.

Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

Akinori Tanaka. Discriminator optimal transport. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

Yan Wu, Mihaela Rosca, and Timothy Lillicrap. Deep compressed sensing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6850–6860, Long Beach, California, USA, 09–15 Jun 2019a. PMLR.

David Siegmund. Importance sampling in the monte carlo study of sequential tests. *The Annals of Statistics*, pages 673–684, 1976.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International Conference on Machine Learning*, pages 2635–2644, 2016.

Yuta Tsuboi, Hisashi Kashima, Shohei Hido, Steffen Bickel, and Masashi Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and Convergence Properties of Generative Adversarial Learning. 2017.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2):708–720, 2011.

Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 224–232. PMLR, 06–11 Aug 2017.

Arnaud Doucet, Nando de Freitas, and Neil Gordon. *Sequential Monte Carlo Methods in Practice*. Information Science and Statistics. Springer-Verlag, New York, 2001. ISBN 978-0-387-95146-1. doi: 10.1007/978-1-4757-3437-9.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3): 411–436, 2006.

Martin Haugh. Mcmc and bayesian modeling. *IEOR E4703 Monte-Carlo Simulation, Columbia University*, 2017.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. 2017.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

Matthias Sachs, Benedict Leimkuhler, and Vincent Danos. Langevin Dynamics with Variable Coefficients and Nonconservative Forces: From Stationary States to Numerical Methods. *Entropy*, 19, December 2017.

Michael Betancourt, Simon Byrne, Sam Livingstone, and Mark Girolami. The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli*, 23(4A):2257–2298, November 2017. ISSN 1350-7265. doi: 10.3150/16-BEJ810.

David Belanger and Andrew McCallum. Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992, 2016.

Jianwen Xie, Song-Chun Zhu, and Ying Nian Wu. Synthesizing dynamic patterns by spatial-temporal generative convnet. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 7093–7101, 2017.

Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Learning descriptor networks for 3d shape synthesis and analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8629–8638, 2018a.

Jianwen Xie, Song-Chun Zhu, and Ying Nian Wu. Learning energy-based spatial-temporal generative convnets for dynamic patterns. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Lifu Tu and Kevin Gimpel. Learning approximate inference networks for structured prediction. *arXiv preprint arXiv:1803.03376*, 2018.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems 32*, pages 3608–3618. Curran Associates, Inc., 2019.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. *arXiv preprint arXiv:2004.11714*, 2020.

Li Wenliang, Danica Sutherland, Heiko Strathmann, and Arthur Gretton. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning*, pages 6737–6746, May 2019.

Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. 2016.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. 2020.

Lantao Yu, Yang Song, Jiaming Song, and Stefano Ermon. Training deep energy-based models with f-divergence minimization. 2020.

Xin Ding, Z. Jane Wang, and William J. Welch. Subsampling Generative Adversarial Networks: Density Ratio Estimation in Feature Space with Softplus Loss. 2019.

John Lawson, George Tucker, Bo Dai, and Rajesh Ranganath. Energy-inspired models: Learning with sampler-induced distributions. In *Advances in Neural Information Processing Systems 32*, pages 8501–8513. Curran Associates, Inc., 2019.

Yan Wu, Jeff Donahue, David Balduzzi, Karen Simonyan, and Timothy Lillicrap. LOGAN: Latent Optimisation for Generative Adversarial Networks. *arXiv:1912.00953 [cs, stat]*, December 2019b. arXiv: 1912.00953.

Jianwen Xie, Yang Lu, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. Cooperative training of descriptor and generator networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):27–45, 2018b.

Jianwen Xie, Yang Lu, Ruiqi Gao, and Ying Nian Wu. Cooperative learning of energy-based model and latent variable model via mcmc teaching. In *AAAI*, volume 1, page 7, 2018c.

Kun Xu, Chao Du, Chongxuan Li, Jun Zhu, and Bo Zhang. Learning implicit generative models by teaching density estimators. *arXiv preprint arXiv:1807.03870*, 2018.

Tu Nguyen, Trung Le, Hung Vu, and Dinh Phung. Dual discriminator generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2670–2680, 2017.

Tong Che, Ruixiang Zhang, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yuan Cao, and Yoshua Bengio. Your GAN is secretly an energy-based model and you should use discriminator driven latent sampling. 2020.

Johann Brehmer and Kyle Cranmer. Flows for simultaneous manifold learning and density estimation. *arXiv preprint arXiv:2003.13913*, 2020.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. 2015.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. 2015b.

Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12154–12163, 2019.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

Georg Ostrovski, Will Dabney, and Rémi Munos. Autoregressive quantile networks for generative modeling. *arXiv preprint arXiv:1806.05575*, 2018.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573, 2016.

Jeff Donahue and Karen Simonyan. Large Scale Adversarial Representation Learning. *arXiv:1907.02544 [cs, stat]*, November 2019. arXiv: 1907.02544.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. *Proceedings of machine learning research*, 70:3987, 2017.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv:1802.02988 [cs, math]*, February 2018.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems 32*, pages 12680–12691. Curran Associates, Inc., 2019.

A. Klenke. *Probability Theory: A Comprehensive Course*. World Publishing Corporation, 2008. ISBN 9787510044113.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1530–1538. JMLR.org, July 2015.

Umut Simsekli, Lingjiong Zhu, Yee Whye Teh, and Mert Gurbuzbalaban. Fractional Underdamped Langevin Dynamics: Retargeting SGD with Momentum under Heavy-Tailed Gradient Noise. *arXiv:2002.05685 [cs, stat]*, February 2020. arXiv: 2002.05685.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *NIPS*, 2017.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. 05 2016.

R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

Carl-Johann Simon-Gabriel and Bernhard Scholkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018.

I. Ekeland and R. Témam. *Convex Analysis and Variational Problems*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 01 1999.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization. *arXiv preprint arXiv:1802.09188*, 2018.

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019.

Qiang Liu. Stein variational gradient descent as gradient flow. In *Advances in neural information processing systems*, pages 3115–3123, 2017.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *arXiv preprint arXiv:1808.09372*, 2018.

Mark Kac. Foundations of kinetic theory. In *Proceedings of The third Berkeley symposium on mathematical statistics and probability*, volume 3, pages 171–197. University of California Press Berkeley and Los Angeles, California, 1956.

HP McKean Jr. A class of markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences of the United States of America*, 56(6):1907, 1966.

Lenaic Chizat and Francis Bach. A Note on Lazy Training in Supervised Differentiable Programming. *arXiv:1812.07956 [cs, math]*, December 2018b. arXiv: 1812.07956.

James Adedayo Oguntuase. On an inequality of gronwall. *Journal of Inequalities in Pure and Applied Mathematics*, 2001.

Adrien Blanchet and Jérôme Bolte. A family of functional inequalities: Lojasiewicz inequalities and displacement convex functions. *Journal of Functional Analysis*, 275(7):1650–1673, 2018.

Umut Şimşekli, Antoine Liutkus, Szymon Majewski, and Alain Durmus. Sliced-wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *ICML*, 2019.

Katy Craig and Andrea Bertozzi. A blob method for the aggregation equation. *Mathematics of computation*, 85(300):1681–1717, 2016.

José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58(2):53, 2019.

Caglar Gulcehre, Marcin Moczulski, Misha Denil, and Yoshua Bengio. Noisy activation functions. In *ICML*, 2016a.

Caglar Gulcehre, Marcin Moczulski, Francesco Visin, and Yoshua Bengio. Mollifying networks. *arXiv preprint arXiv:1608.04980*, 2016b.

Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillaume Carlier. Deep Relaxation: partial differential equations for optimizing deep neural networks. *arXiv:1704.04932 [cs, math]*, 2017.

Elad Hazan, Kfir Yehuda Levy, and Shai Shalev-Shwartz. On graduated optimization for stochastic non-convex problems. In *ICML*, 2016.

Benjamin Jourdain, Sylvie Méléard, and Wojbor Woyczynski. Nonlinear sdes driven by levy proesses and related pdes. *arXiv preprint arXiv:0707.2723*, 2007.

Kiyosi Itô. *On stochastic differential equations*, volume 4. American Mathematical Soc., 1951.

Ivan V. Shestakov and Alexander A. Shlapunov. Negative Sobolev Spaces in the Cauchy Problem for the Cauchy-Riemann Operator. January 2009.

Maximilian Behr, Peter Benner, and Jan Heiland. Solution Formulas for Differential Sylvester and Lyapunov Equations. *arXiv:1811.08327 [math]*, November 2018. arXiv: 1811.08327.

Lenaic Chizat, Bernhard Schmitzer, Gabriel Peyré, and François-Xavier Vialard. An Interpolating Distance between Optimal Transport and Fisher-Rao. *arXiv:1506.06430 [math]*, 2015.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: The hellinger–kantorovich distance and geodesic curves. *SIAM Journal on Mathematical Analysis*, 2016.

Stanislav Kondratyev, Léonard Monsaingeon, Dmitry Vorotnikov, et al. A new optimal transport distance on the space of finite radon measures. *Advances in Differential Equations*, 2016.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011. ISSN ISSN 1533-7928.

G. Hinton, N. Srivastava, and K. Swersky. Lecture 6a overview of mini–batch gradient descent., 2012.

James Martens. Deep Learning via Hessian-free Optimization. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 735–742, USA, 2010. Omnipress. ISBN 978-1-60558-907-7. event-place: Haifa, Israel.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147, February 2013.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, September 2014. arXiv: 1409.1556.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. doi: 10.1109/CVPR.2016.90.

Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient BackProp. In Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade: Second Edition*, Lecture Notes in Computer Science, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_3.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015. event-place: Lille, France.

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *arXiv:1602.07868 [cs]*, 2016.

Ronald Aylmer Fisher and Edward John Russell. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, January 1922. doi: 10.1098/rsta.1922.0009.

C. Radhakrishna Rao. Information and the Accuracy Attainable in the Estimation of Statistical Parameters. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Foundations and Basic Theory*, Springer Series in Statistics, pages 235–247. Springer New York, New York, NY, 1992. ISBN 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5_16.

Felix Otto. The Geometry of Dissipative Evolution Equations: The Porous Medium Equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, January 2001. ISSN 0360-5302. doi: 10.1081/PDE-100002243.

Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. Wasserstein Riemannian Geometry of Positive Definite Matrices. *arXiv:1801.09269 [math, stat]*, January 2018. arXiv: 1801.09269.

Klas Modin. Geometry of Matrix Decompositions Seen Through Optimal Transport and Information Geometry. *Journal of Geometric Mechanics*, 9(3):335–390, 2017. ISSN 1941-4897. doi: 10.3934/jgm.2017014. arXiv: 1601.01875.

S. Amari and A. Cichocki. Information geometry of divergence functions. *Bulletin*

of the Polish Academy of Sciences: Technical Sciences, 58(No 1):183–195, 2010. doi: 10.2478/v10175-010-0019-1.

John Lafferty and Larry Wasserman. Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1):28–63, 2008.

Yifan Chen and Wuchen Li. Natural gradient in Wasserstein statistical manifold. *arXiv:1805.08380 [cs, math]*, May 2018. arXiv: 1805.08380.

Diederik P. Kingma, Tim Salimans, and Max Welling. Variational Dropout and the Local Reparameterization Trick. *ArXiv*, abs/1506.02557, 2015. arXiv: 1506.02557.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, pages 416–426, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-44581-4.

Aaron Mishkin, Frederik Kunstner, Didrik Nielsen, Mark Schmidt, and Mohammad Emtiyaz Khan. SLANG: Fast Structured Covariance Approximations for Bayesian Deep Learning with Natural Gradient. *arXiv:1811.04504 [cs, stat]*, November 2018. arXiv: 1811.04504.

Tengyuan Liang and Alexander Rakhlin. Just Interpolate: Kernel "Ridgeless" Regression Can Generalize. *arXiv:1808.00387 [cs, math, stat]*, February 2019. arXiv: 1808.00387.

James Martens and Ilya Sutskever. Training Deep and Recurrent Networks with Hessian-Free Optimization. In Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade: Second Edition*, Lecture Notes in Computer Science, pages 479–535. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_27.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv:1707.07269 [math, stat]*, October 2018. arXiv: 1707.07269.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. 07 2015.

Siqi Sun, mladen kolar, and Jinbo Xu. Learning structured densities via infinite dimensional exponential families. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2287–2295. Curran Associates, Inc., 2015.

Alain Rakotomamonjy, Remi Flamary, Gilles Gasso, and Stephane Canu. Lp-lq penalty for Sparse Linear and Sparse Multiple Kernel Multi-Task Learning. page 14, 2011.

Alberto Bietti, Grégoire Mialon, and Julien Mairal. On Regularization and Robustness of Deep Neural Networks. *arXiv:1810.00363 [cs, stat]*, September 2018. arXiv: 1810.00363.

Alberto Bietti and Julien Mairal. Group Invariance, Stability to Deformations, and Complexity of Deep Convolutional Representations. *arXiv:1706.03078 [cs, stat]*, June 2017. arXiv: 1706.03078.

Herbert Federer. *Geometric measure theory*. Springer, 2014.

Laurent Bartholdi, Thomas Schick, Nat Smale, and Steve Smale. Hodge theory on metric spaces. *Foundations of Computational Mathematics*, 12(1):1–48, 2012.

Ananya Uppal, Shashank Singh, and Barnabás Póczos. Nonparametric density estimation & convergence rates for gans under besov ipm losses. In *Advances in Neural Information Processing Systems*, pages 9089–9100, 2019.

Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, Anna Choromanska, Krzysztof Choromanski, and Michael I Jordan. Learning to score behaviors for guided policy optimization. *arXiv preprint arXiv:1906.04349*, 2019.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *CoRR*, abs/1502.05477, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.