Training generative neural networks via Maximum Mean Discrepancy optimization

Gintare Karolina Dziugaite
University of Cambridge

Daniel M. Roy
University of Toronto

Zoubin Ghahramani
University of Cambridge

Abstract

We consider training a deep neural network to generate samples from an unknown distribution given i.i.d. data. We frame learning as an optimization minimizing a two-sample test statistic—informally speaking, a good generator network produces samples that cause a two-sample test to fail to reject the null hypothesis. As our two-sample test statistic, we use an unbiased estimate of the maximum mean discrepancy, which is the centerpiece of the nonparametric kernel two-sample test proposed by Gretton et al. [2]. We compare to the *adversarial nets* framework introduced by Goodfellow et al. [1], in which learning is a two-player game between a generator network and an adversarial discriminator network, both trained to outwit the other. From this perspective, the MMD statistic plays the role of the discriminator. In addition to empirical comparisons, we prove bounds on the generalization error incurred by optimizing the empirical MMD.

1 INTRODUCTION

In this paper, we consider the problem of learning generative models from i.i.d. data with unknown distribution $\mathcal{P}$. We formulate the learning problem as one of finding a function G, called the *generator*, such that, given an input Z drawn from some fixed *noise* distribution $\mathcal{N}$, the distribution of the output $G(Z)$ is close to the data's distribution $\mathcal{P}$. Note that, given G and $\mathcal{N}$, we can easily generate new samples despite not having an explicit representation for the underlying density.

We are particularly interested in the case where the generator is a deep neural network whose parameters we must learn. Rather than being used to classify or predict, these networks transport input randomness to output randomness, thus inducing a distribution. The first direct instantiation of this idea is due to MacKay [7], although MacKay draws connections even further back to the work of Saund [11] and others on autoencoders, suggesting that generators can be understood as decoders. MacKay's proposal, called *density networks*, uses multi-layer perceptrons (MLP) as generators and learns the parameters by approximating Bayesian inference.

Since MacKay's proposal, there has been a great deal of progress on learning generative models, especially over high-dimensional spaces like images. Some of the most successful approaches have been based on restricted Boltzmann machines [10] and deep Boltzmann networks [3]. A recent example is the Neural Autoregressive Density Estimator due to Uria, Murray, and Larochelle [15]. An indepth survey, however, is beyond the scope of this article.

This work builds on a proposal due to Goodfellow et al. [1]. Their *adversarial nets* framework takes an indirect approach to learning deep generative neural networks: a discriminator network is trained to recognize the difference between training data and generated samples, while the generator is trained to confuse the discriminator. The resulting two-player game is cast as a minimax optimization of a differentiable objective and solved greedily by iteratively performing gradient descent steps to improve the generator and then the discriminator.

Given the greedy nature of the algorithm, Goodfellow et al. [1] give a careful prescription for balancing the training of the generator and the discriminator. In particular, two gradient steps on the discriminator's parameters are taken for every iteration of the generator's parameters. It is not clear at this point how sensitive this balance is as the data set and network vary. In this paper, we describe an approximation to adversarial learning that replaces the adversary with a closed-form nonparametric two-sample test statistic based on the Maximum Mean Discrepancy (MMD), which we adopted from the kernel two sample test [2]. We call our proposal *MMD nets*.[1] We give bounds on the estimation

[1] In independent work reported in a recent preprint, Li, Swer-

error incurred by optimizing an empirical estimator rather than the true population MMD and give some illustrations on synthetic and real data.

2 LEARNING TO SAMPLE AS OPTIMIZATION

It is well known that, for any distribution $\mathcal{P}$ and any continuous distribution $\mathcal{N}$ on sufficiently regular spaces $\mathbb{X}$ and $\mathbb{W}$, respectively, there is a function $G : \mathbb{W} \to \mathbb{X}$, such that $G(W) \sim \mathcal{P}$ when $W \sim \mathcal{N}$. (See, e.g., [4, Lem. 3.22].) In other words, we can transform an input from a fixed input distribution $\mathcal{N}$ through a deterministic function, producing an output whose distribution is $\mathcal{P}$. For a given family $\{G_\theta\}$ of functions $\mathbb{W} \to \mathbb{X}$, called *generators*, we can cast the problem of learning a generative model as an optimization

$$\arg\min_\theta \delta(\mathcal{P}, G_\theta(\mathcal{N})), \tag{1}$$

where δ is some measure of discrepancy and $G_\theta(\mathcal{N})$ is the distribution of $G_\theta(W)$ when $W \sim \mathcal{N}$. In practice, we only have i.i.d. samples $X_1, X_2, \dots$ from $\mathcal{P}$, and so we optimize an empirical estimate of $\delta(\mathcal{P}, G_\theta(\mathcal{N}))$.

2.1 ADVERSARIAL NETS

Adversarial nets [1] can be cast within this framework: Let $\{D_\phi\}$ be a family of functions $\mathbb{X} \to [0, 1]$, called *discriminators*. We recover the adversarial nets objective with the discrepancy

$$\delta_{\mathrm{AN}}(\mathcal{P}, G_\theta(\mathcal{N})) = \max_\phi E\big[\log D_\phi(X) + \log(1 - D_\phi(Y))\big],$$

where $X \sim \mathcal{P}$ and $Y \sim G_\theta(\mathcal{N})$. In this case, Eq. (1) becomes

$$\min_\theta \max_\phi V(G_\theta, D_\phi)$$

where

$$V(G_\theta, D_\phi) = E\big[\log D_\phi(X) + \log(1 - D_\phi(G_\theta(W)))\big]$$

for $X \sim \mathcal{P}$ and $W \sim \mathcal{N}$. The output of the discriminator D_ϕ can be interpreted as the probability it assigns to its input being drawn from $\mathcal{P}$, and so $V(G_\theta, D_\phi)$ is the expected log loss incurred when classifying the origin of a point equally likely to have been drawn from $\mathcal{P}$ or $G_\theta(\mathcal{N})$. Therefore, optimizing ϕ maximizes the probability of distinguishing samples from $\mathcal{P}$ and $G_\theta(\mathcal{N})$. Assuming that the optimal discriminator exists for every θ, the optimal generator G is that whose output distribution is closest to $\mathcal{P}$, as measured by the Jensen–Shannon divergence, which is minimized when $G_\theta(\mathcal{N}) = \mathcal{P}$.

In [1], the generators G_θ and discriminators D_ϕ are chosen to be multilayer perceptrons (MLP). In order to find a minimax solution, they propose taking alternating gradient steps along D_ϕ and G_θ. Note that the composition $D_\phi(G_\theta(\cdot))$ that appears in the value function is yet another (larger) MLP. This fact permits the use of the back-propagation algorithm to take gradient steps.

2.2 MMD AS AN ADVERSARY

In their paper introducing adversarial nets, Goodfellow et al. [1] remark that a balance must be struck between optimizing the generator and optimizing the discriminator. In particular, the authors suggest k maximization steps for every one minimization step to ensure that D_ϕ is well synchronized with G_θ during training. A large value for k, however, can lead to overfitting. In their experiments, for every step taken along the gradient with respect to G_θ, they take two gradient steps with respect to D_ϕ to bring D_ϕ closer to the desired optimum (Goodfellow, pers. comm.).

It is unclear how sensitive this balance is. Regardless, while adversarial networks deliver impressive sampling performance, the optimization takes approximately 7.5 hours to train on the MNIST dataset running on a GeForce GTX TITAN GPU from nVidia with 6GB RAM. Can we potentially speed up the process with a more tractable choice of adversary?

Our proposal is to replace the adversary with the kernel two-sample test introduced by Gretton et al. [2]. In particular, we replace the family of discriminators with a family $\mathcal{H}$ of test functions $\mathbb{X} \to \mathbb{R}$, closed under negation, and use the maximum mean discrepancy between $\mathcal{P}$ and $G_\theta(\mathcal{N})$ over $\mathcal{H}$, given by

$$\delta_{\mathrm{MMD}_\mathcal{H}}(\mathcal{P}, G_\theta(\mathcal{N})) = \sup_{f \in \mathcal{H}} E[f(X)] - E[f(Y)], \tag{2}$$

where $X \sim \mathcal{P}$ and $Y \sim G_\theta(\mathcal{N})$. See Fig. 1 for a comparison of the architectures of adversarial and MMD nets.

While Eq. (2) involves a maximization over a family of functions, Gretton et al. [2] show that it can be solved in closed form when $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

More carefully, let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) of real-valued functions on Ω and let $\langle \cdot, \cdot \rangle_\mathcal{H}$ denote its inner product. By the reproducing property it follows that there exists a *reproducing kernel* $k \in \mathcal{H}$ such that every $f \in \mathcal{H}$ can be expressed as

$$f(x) = \langle f, k(\cdot, x) \rangle_\mathcal{H} = \sum \alpha_i k(x, x_i) \tag{3}$$

The functions *induced by a kernel* k are those functions in the closure of the span of the set $\{k(\cdot, x) : x \in \Omega\}$, which is necessarily an RKHS. Note, that for every positive definite kernel there is a unique RKHS $\mathcal{H}$ such that every function in $\mathcal{H}$ satisfies Eq. (3).

sky, and Zemel [6] also propose to use MMD as a training objective for generative neural networks. We leave a comparison to future work.

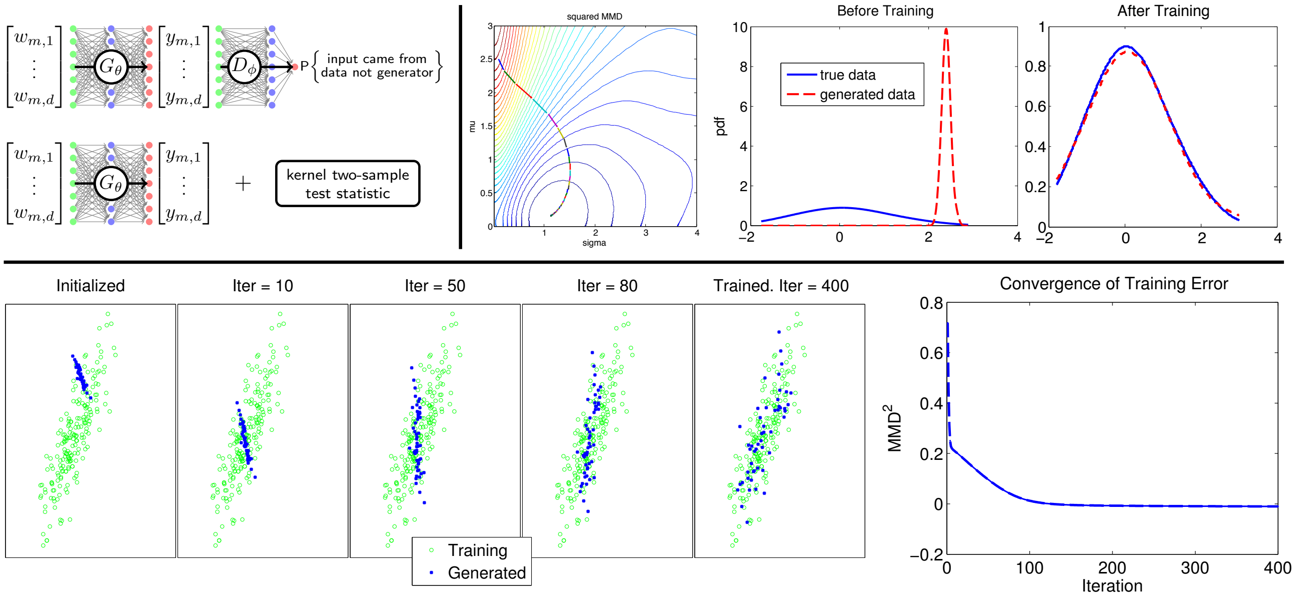

Figure 1: (top left) Comparison of adversarial nets and MMD nets. **(top right)** Here we present a simple one-dimensional illustration of optimizing a generator via MMD. Both the training data and noise data are Gaussian distributed and we consider the class of generators given by $G_{(\mu,\sigma)}(w) = \mu + \sigma w$. The plot on the left shows the isocontours of the MMD-based cost function and the path taken by gradient descent. On right, we show the distribution of the generator before and after a number of training iterations, as compared with the data generating distribution. Here we did not resample the generated points and so we do not expect to be able to drive the MMD to zero and match the distribution exactly. **(bottom)** The same procedure is repeated here for a two-dimensional dataset. On the left, we see the gradual alignment of the Gaussian-distributed input data to the Gaussian-distributed output data as the parameters of the generator G_θ are optimized. The learning curve on the right shows the decrease in MMD obtained via gradient descent.

Assume that $\mathbb{X}$ is a nonempty compact metric space and $\mathcal{F}$ a class of functions $f : \mathbb{X} \to \mathbb{R}$. Let p and q be Borel probability measures on $\mathbb{X}$, and let X and Y be random variables with distribution p and q, respectively. The *maximum mean discrepancy* (MMD) between p and q is

$$\mathrm{MMD}(\mathcal{F}, p, q) = \sup_{f \in \mathcal{F}} E[f(X)] - E[f(Y)]$$

If $\mathcal{F}$ is chosen to be an RKHS $\mathcal{H}$, then

$$\mathrm{MMD}^2(\mathcal{F}, p, q) = \|\mu_p - \mu_q\|^2_{\mathcal{H}}$$

where $\mu_p \in \mathcal{H}$ is the *mean embedding* of p, given by

$$\mu_p = \int_{\mathbb{X}} k(x, \cdot)\, p(\mathrm{d}x) \in \mathcal{H}$$

and satisfying, for all $f \in \mathcal{H}$,

$$E[f(X)] = \langle f, \mu_p \rangle_{\mathcal{H}}.$$

The properties of $\mathrm{MMD}(\mathcal{H}, \cdot, \cdot)$ depend on the underlying RKHS $\mathcal{H}$. For our purposes, it suffices to say that if we take $\mathbb{X}$ to be $\mathbb{R}^D$ and consider the RKHS $\mathcal{H}$ induced by Gaussian or Laplace kernels, then MMD is a metric, and so the minimum of our learning objective is achieved uniquely by $\mathcal{P}$, as desired. (For more details, see Sriperumbudur et al. [12].)

In practice, we often do not have access to p or q. Instead, we are given independent i.i.d. data $X, X', X_1, \dots, X_N$ and $Y, Y', Y_1, \dots, Y_M$ fom p and q, respectively, and would like to estimate the MMD. Gretton et al. [2] showed that

$$\mathrm{MMD}^2[\mathcal{H}, p, q] = E[k(X, X') - 2k(X, Y) + k(Y, Y')]$$

and then proposed an unbiased estimator

$$\begin{aligned}\mathrm{MMD}^2_u[\mathcal{H}, X, Y] = & \frac{1}{N(N-1)} \sum_{n \neq n'} k(x_n, x_{n'}) \\ & + \frac{1}{M(M-1)} \sum_{m \neq m'} k(y_m, y_{m'}) \\ & - \frac{2}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} k(x_n, y_m).\end{aligned} \tag{4}$$

3 MMD NETS

With an unbiased estimator of the MMD objective in hand, we can now define our proposal, *MMD nets*: Fix a neural network G_θ, where θ represents the parameters of the network. Let $W = (w_1, \dots, w_M)$ denote noise inputs drawn from $\mathcal{N}$, let $Y_\theta = (y_1, \dots, y_m)$ with $y_j = G_\theta(w_j)$ denote

Algorithm 1 Stochastic gradient descent for MMD nets.

```
Initialize $M$, $\theta$, $\alpha$, $k$
Randomly divide training set $X$ into $N_{\text{mini}}$ mini batches
for $i \leftarrow 1$, number-of-iterations do
    Regenerate noise inputs $\{w_i\}_{i=1,...,M}$ every $r$ iterations
    for $n_{\text{mini}} \leftarrow 1, N_{\text{mini}}$ do
        for $m \leftarrow 1, M$ do
            $y_m \leftarrow G_\theta(w_m)$
        end for
        compute the n'th minibatch's gradient $\nabla C^{(n)}$
        update learning rate $\alpha$ (e.g., RMSPROP)
        $\theta \leftarrow \theta - \alpha \nabla C_n$
    end for
end for
```

the noise inputs transformed by the network G_θ, and let $X = (x_1, .., x_N)$ denote the training data in $\mathbb{R}^D$. Given a positive definite kernel k on $\mathbb{R}^D$, we minimize $C(Y_\theta, X)$ as a function of θ, where

$$C(Y_\theta, X) = \frac{1}{M(M-1)} \sum_{m \neq m'} k(y_m, y_{m'}) - \frac{2}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} k(y_m, x_n).$$

Note that $C(Y_\theta, X)$ is composed of only those parts of the unbiased estimator (Eq. (4)) that depend on θ.

In practice, the minimization is solved by gradient descent, possibly on subsets of the data. More carefully, the chain rule gives us

$$\nabla C(Y_\theta, X) = \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{\partial C_n(Y_\theta, X_n)}{\partial y_m} \frac{\partial G_\theta(w_m)}{\partial \theta},$$

where

$$C_n(Y_\theta, X_n) = \frac{1}{M(M-1)} \sum_{m \neq m'} k(y_m, y_{m'}) - \frac{2}{M} \sum_{m=1}^{M} k(y_m, x_n).$$

Each derivative $\frac{\partial C_n(Y_\theta, X_n)}{\partial y_m}$ is easily computed for standard kernels like the RBF kernel. Our gradient $\nabla C(Y_\theta, X_n)$ depends on the partial derivatives of the generator with respect to its parameters, which we can compute using back propagation.

4 MMD GENERALIZATION BOUNDS

MMD nets operate by minimizing an empirical estimate of the MMD. This estimate is subject to Monte Carlo error and so the network weights (parameters) $\hat{\theta}$ that are found to minimize the empirical MMD may do a poor job at minimizing the exact population MMD. We show that, for sufficiently large data sets, this estimation error is bounded, despite the space of parameters θ being continuous and high dimensional.

Let Θ denote the space of possible parameters for the generator G_θ, let $\mathcal{N}$ be the distribution on $\mathcal{W}$ for the noisy inputs, and let $p_\theta = G_\theta(\mathcal{N})$ be the distribution of $G_\theta(W)$ when $W \sim \mathcal{N}$ for $\theta \in \Theta$. Let $\hat{\theta}$ be the value optimizing the unbiased empirical MMD estimate, i.e.,

$$\mathrm{MMD}_u^2(\mathcal{H}, X, Y_{\hat{\theta}}) = \inf_\theta \mathrm{MMD}_u^2(\mathcal{H}, X, Y_\theta), \qquad (5)$$

and let θ^* be the value optimizing the population MMD, i.e.,

$$\mathrm{MMD}^2(\mathcal{H}, p_{\text{data}}, p_{\theta^*}) = \inf_\theta \mathrm{MMD}^2(\mathcal{H}, p_{\text{data}}, p_\theta).$$

We are interested in bounding the difference

$$\mathrm{MMD}^2(\mathcal{H}, p_{\text{data}}, p_{\hat{\theta}}) - \mathrm{MMD}^2(\mathcal{H}, p_{\text{data}}, p_{\theta^*}).$$

To that end, for a measured space $\mathcal{X}$, write $L_\infty(\mathcal{X})$ for the space of essentially bounded functions on $\mathcal{X}$ and write $B(L_\infty(\mathcal{X}))$ for the unit ball under the sup norm, i.e.,

$$B(L_\infty(\mathcal{X})) = \{f \colon \mathcal{X} \to \mathbb{R} : (\forall x \in \mathcal{X}) f(x) \in [-1, 1]\}.$$

The bounds we obtain will depend on a notion of complexity captured by the fat-shattering dimension:

Definition 1 (Fat-shattering [8]). Let $\mathcal{X}_N = \{x_1, \ldots, x_N\} \subset \mathcal{X}$ and $\mathcal{F} \subset B(L_\infty(\mathcal{X}))$. For every $\varepsilon > 0$, $\mathcal{X}_N$ is said to be *ε-shattered by $\mathcal{F}$* if there is some function $h : \mathcal{X} \to \mathbb{R}$, such that for every $I \subset \{1, \ldots, N\}$ there is some $f_I \in \mathcal{F}$ for which

$$f_I(x_n) \geq h(x_n) + \varepsilon \text{ if } n \in I,$$
$$f_I(x_n) \leq h(x_n) - \varepsilon \text{ if } n \notin I.$$

For every ε, the *fat-shattering dimension of $\mathcal{F}$*, written $\mathrm{fat}_\varepsilon(\mathcal{F})$, is defined as

$$\mathrm{fat}_\varepsilon(\mathcal{F}) = \sup \{|\mathcal{X}_N| : \mathcal{X}_N \subset \mathcal{X}, \mathcal{X}_N \text{ is } \varepsilon\text{-shattered by } \mathcal{F}\}.$$

Consider the class

$$\mathcal{G}_{k+}^{\mathbb{X}} = \{g = k(x, G_\theta(\cdot)) : x \in \mathbb{X}, \theta \in \Theta\}$$

of functions from $\mathcal{W}$ to $\mathbb{R}$ that are compositions of some generator and the kernel with some fixed input, and the (sub)class

$$\mathcal{G}_{k+} = \{g = k(G_\theta(w), G_\theta(\cdot)) : w \in \mathcal{W}, \theta \in \Theta\}.$$

We then have the following bound on the estimation error:

Theorem 1 (estimation error)**.** *Assume the kernel is bounded by one and that there exists $\gamma_1, \gamma_2 > 1$ and $p_1, p_2 \in \mathbb{N}$ such that, for all $\varepsilon > 0$, it holds that $fat_\varepsilon(\mathcal{G}_{k+}) \leq \gamma_1 \varepsilon^{-p_1}$ and $fat_\varepsilon(\mathcal{G}^X_{k+}) \leq \gamma_2 \varepsilon^{-p_2}$. Then with probability at least $1 - \delta$,*

$$\mathrm{MMD}^2(\mathcal{H}, p_{\mathrm{data}}, p_{\hat{\theta}}) < \mathrm{MMD}^2(\mathcal{H}, p_{\mathrm{data}}, p_{\theta^*}) + \varepsilon,$$

with

$$\varepsilon = r(p_1, \gamma_1, M) + r(p_2, \gamma_2, M-1) + 12M^{-\frac{1}{2}}\sqrt{\log\frac{2}{\delta}},$$

where the rate $r(p, \gamma, M)$ is

$$r(p, \gamma, M) = C_p\sqrt{\gamma}\begin{cases} M^{-\frac{1}{2}} & \textit{if } p < 2, \\ M^{-\frac{1}{2}}\log^{\frac{3}{2}}(M) & \textit{if } p = 2, \\ M^{-\frac{1}{p}} & \textit{if } p > 2, \end{cases}$$

for constants C_{p_1} and C_{p_2} depending on p_1 and p_2 alone.

The proof appears in the appendix. We can obtain simpler, but slightly more restrictive, hypotheses if we bound the fat-shattering dimension of the class of generators $\{G_\theta : \theta \in \Theta\}$ alone: Take the observation space $\mathbb{X}$ to be a bounded subset of a finite-dimensional Euclidean space and the kernel to be Lipschitz continuous and translation invariant. For the RBF kernel, the Lipschitz constant is proportional to the inverse of the length-scale: the resulting bound loosens as the length scale shrinks.

5 EMPIRICAL EVALUATION

In this section, we demonstrate the approach on an illustrative synthetic example as well as the standard MNIST digits and Toronto Face Dataset (TFD) benchmarks. We show that MMD-based optimization of the generator rapidly delivers a generator that produces recognizable samples, but these samples are inferior to those produced by adversarial networks, both visually and as measured by an estimate of the mean log density on a held-out test set.

5.1 GAUSSIAN DATA, KERNEL, AND GENERATOR

Under an RBF kernel and Gaussian generator with parameters $\theta = \{\mu, \sigma\}$, it is straightforward to find the gradient of $C(Y_\theta, X)$ by applying the chain rule. Using fixed random standard normal numbers $\{w_1, ..., w_M\}$, we have $y_m = \mu + \sigma w_m$ for $m \in \{1, .., M\}$. The result of these illustrative synthetic experiments can be found in Fig. 1. The dataset consisted of $N = 200$ samples from a standard normal and $M = 50$ noise input samples were generated from a standard normal with a fixed random seed. The algorithm was initialized at values $\{\mu, \sigma\} = \{2.5, 0.1\}$. We fixed the learning rate to 0.5 and ran gradient descent steps for $K = 250$ iterations.

5.2 MNIST DIGITS

We evaluated MMD nets on MNIST digits [5]. The generator was chosen to be a fully connected, 3 hidden layer neural network with sigmoidal activation functions. Following Gretton et al. [2], we used a radial basis function (RBF) kernel, but also evaluated the rational quadratic (RQ) kernel [9] and Laplacian kernel, but found that the RBF performed best in the parameter ranges we evaluated. We used Bayesian optimization (WHETLab) to set the bandwidth of the RBF and the number of neurons in each layer on initial test runs of 50,000 iterations. However, one can get a similar-quality generator simply using the median heuristic [2] to set the kernel bandwidth. The learning rate was adjusting during optimization by RMSPROP [14].

Fig. 2 presents the digits learned after 1,000,000 iterations. (Doubling the number of iterations produced similar images.) We performed minibatch stochastic gradient descent, resampling the generated digits every 300 iterations, with minibatches of 500 training and generated points. It is clear that the digits produced have many artifacts not appearing in the MNIST data set. Indeed, the mean log density of held-out test data was estimated to be only **113 ± 2**, as compared with the reported **225 ± 2** achieved by adversarial nets. On the other hand, most of the gain is achieved by MMD nets in the first 100-200k iterations, and so perhaps MMD nets could be used to initialize a network further optimized by other means.

5.3 TORONTO FACE DATASET

We also evaluated MMD nets on the Toronto face dataset (TFD) [13]. We used a 3-hidden-layer sigmoidal MLP with similar architecture (1000, 600, and 1000 units) and RBF kernel for the cost function with the same hyper parameter. We used 500 training and generated points per batch. The generated points were resampled every 500 iterations. The network was optimized for 500,000 iterations. Samples from the resulting network are plotted in Fig. 3. Again, the samples produced by MMD nets are clearly distinguishable from the training samples and this is reflected in a much lower mean log density than adversarial nets.

6 CONCLUSION

MMD offers a closed-form surrogate for the discriminator in the adversarial nets framework. After using Bayesian optimization for the parameters, we found that the network produced samples that were visually similar, but far from indistinguishable from those used to train the network. On one hand, adversarial nets handedly outperformed MMD nets in terms of mean log density. On the other, MMD nets achieve most of their gain quickly and so it seems promising to combine MMD nets with another technique, perhaps using MMD nets to initialize a more costly procedure.

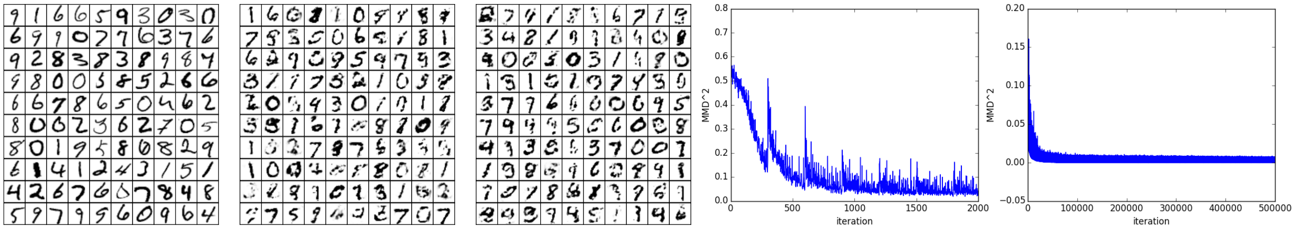

Figure 2: (top-left) MNIST digits from the training set. **(top-right)** Newly generated digits produced after 1,000,000 iterations (approximately 5 hours). Despite the remaining artifacts, the resulting kernel-density estimate of the test data is state of the art. **(top-center)** Newly generated digits after 300 further iterations optimizing the associated empirical MMD. **(bottom-left)** MMD learning curves for first 2000 iterations. **(bottom-right)** MMD learning curves from 2000 to 500,000 iterations. Note the difference in y-axis scale. No appreciable change is seen in later iterations.

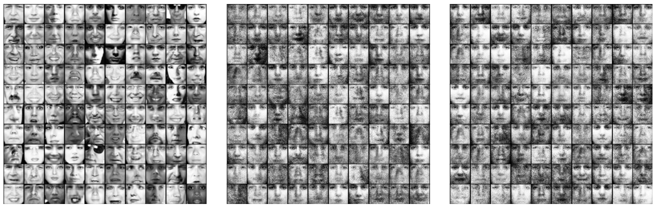

Figure 3: (left) TFD. **(right)** Faces generated by network trained for 500,000 iterations. **(center)** After an additional 500 iterations.

A PROOFS

We begin with some preliminaries and known results:

Definition 2 ([8])**.** A random variable σ is said to be a *Rademacher random variable* if it takes values in $\{-1, 1\}$, each with probability $1/2$.

Definition 3 ([8])**.** Let μ be a probability measure on $\mathcal{X}$, and let $\mathcal{F}$ be a class of uniformly bounded functions on $\mathcal{X}$. Then the *Rademacher complexity* of $\mathcal{F}$ (with respect to μ) is

$$R_N(\mathcal{F}) = E_\mu E_{\sigma_1,\dots,\sigma_N}\left[\frac{1}{\sqrt{N}}\sup_{f\in\mathcal{F}}\left|\sum_{n=1}^{N}\sigma_n f(X_n)\right|\right],$$

where $\sigma = (\sigma_1, \sigma_2, \dots)$ is a sequence of independent Rademacher random variables, and $X_1, X_2, \dots$ are independent, μ-distributed random variables, independent also from σ.

Theorem 2 (McDiarmids Inequality [8])**.** *Let $f : \mathcal{X}_1 \times \dots \times \mathcal{X}_N \to \mathbb{R}$ and assume there exists $c_1, \dots, c_N \geq 0$ such that, for all $k \in \{1, \dots, N\}$, we have*

$$\sup_{x_1,\dots,x_k,x_k',\dots,x_N} |f(x_1, \dots, x_k, \dots, x_N) - f(x_1, \dots, x_k', \dots, x_N)| \leq c_k.$$

Then, for all $\varepsilon > 0$ and independent random variables $\xi_1, \dots, \xi_n$ in $\mathcal{X}$,

$$\Pr\{f(\xi_1, \dots, \xi_N) - E(f(\xi_1, \dots, \xi_N)) \geq \varepsilon)\} < \exp\left(\frac{-2\varepsilon^2}{\sum_{n=1}^{N} c_n^2}\right).$$

Theorem 3 ([8, Thm. 2.35])**.** *Let $\mathcal{F} \subset B(L_\infty(\mathcal{X}))$. Assume there exists $\gamma > 1$, such that for all $\varepsilon > 0$, $fat_\varepsilon(\mathcal{F}) \leq \gamma\varepsilon^{-p}$ for some $p \in \mathbb{N}$. Then there exists constants C_p depending on p only, such that $R_N(\mathcal{F}) \leq C_p\Psi(p, N, \gamma)$ where*

$$\Psi(p, N, \gamma) = \gamma^{\frac{1}{2}}\begin{cases} 1 & \textit{if } 0 < p < 2 \\ \log^{\frac{3}{2}} N & \textit{if } p = 2 \\ N^{\frac{1}{2}-\frac{1}{p}} & \textit{if } p > 2. \end{cases}$$

Theorem 4 ([2])**.** *Assume $0 \leq k(x_i, x_j) \leq K$, $M = N$. Then*

$$\Pr\left[|\mathrm{MMD}_u^2(\mathcal{H}, X, Y_\theta) - \mathrm{MMD}^2(\mathcal{H}, p_{\mathrm{data}}, p_\theta)| > \varepsilon\right] \leq \delta_\varepsilon$$

where

$$\delta_\varepsilon = 2\exp\left(-\frac{\varepsilon^2 M}{16K^2}\right).$$

The case where Θ is a finite set is elementary:

Theorem 5 (estimation error for finite parameter set)**.** *Let p_θ be the distribution of $G_\theta(W)$, with θ taking values in some finite set $\Theta = \{\theta_1, ..., \theta_T\}, T < \infty$. Then, with probability at least $1 - (T + 1)\delta_\varepsilon$, where δ_ε is defined as in Theorem 4, we have*

$$\mathrm{MMD}^2(\mathcal{H}, p_{\mathrm{data}}, p_{\hat{\theta}}) < \mathrm{MMD}^2(\mathcal{H}, p_{\mathrm{data}}, p_{\theta^*}) + 2\varepsilon.$$

Proof. Let $\mathcal{E}(\theta) = \mathrm{MMD}_u^2(\mathcal{H}, X, Y_\theta)$ and let $\mathcal{T}(\theta) = \mathrm{MMD}^2(\mathcal{H}, p_{\mathrm{data}}, p_\theta)$.

Note, that the upper bound stated in Theorem 4 holds for the parameter value θ^*, i.e.,

$$\Pr\left[|\mathcal{E}(\theta^*) - \mathcal{T}(\theta^*)| > \varepsilon\right] \leq \delta_\varepsilon. \tag{6}$$

Because $\hat{\theta}$ depends on the training data X and generator data Y, we use a uniform bound that holds over all θ. Specifically,

$$\begin{aligned}\Pr\left[|\mathcal{E}(\hat{\theta}) - \mathcal{T}(\hat{\theta})| > \varepsilon\right] &\le \Pr\left[\sup_{\theta} |\mathcal{E}(\theta) - \mathcal{T}(\theta)| > \varepsilon\right] \qquad (7)\\ &\le \sum_{t=1}^{T} \Pr\left[|\mathcal{E}(\hat{\theta}) - \mathcal{T}(\hat{\theta})| > \varepsilon\right] \le T\delta_{\varepsilon}.\end{aligned}$$

This yields that with probability at least $1 - T\delta_{\varepsilon}$,

$$\begin{aligned}2\varepsilon &\ge |\mathcal{E}(\hat{\theta}) - \mathcal{T}(\hat{\theta})| + |\mathcal{E}(\theta^*) - \mathcal{T}(\theta^*)|\\ &\ge |\mathcal{E}(\theta^*) - \mathcal{E}(\hat{\theta}) + \mathcal{T}(\hat{\theta}) - \mathcal{T}(\theta^*)|.\end{aligned} \tag{8}$$

Since θ^* was chosen to minimize $\mathcal{T}(\theta)$, we know that $\mathcal{T}(\hat{\theta}) \ge \mathcal{T}(\theta^*)$. Similarly, by Eq. (5), $\mathcal{E}(\theta^*) \ge \mathcal{E}(\hat{\theta})$. Therefore it follows that

$$\begin{aligned}2\varepsilon &\ge \mathcal{T}(\hat{\theta}) - \mathcal{T}(\theta^*)\\ &= \mathrm{MMD}^2(\mathcal{H}, p_{\mathrm{data}}, p_{\theta^*}) - \mathrm{MMD}^2(\mathcal{H}, p_{\mathrm{data}}, p_{\hat{\theta}})\end{aligned}$$

proving the theorem. □

Corollary 1. *With probability at least* $1 - \delta$,

$$\mathrm{MMD}^2(\mathcal{H}, p_{\mathrm{data}}, p_{\hat{\theta}}) < \mathrm{MMD}^2(\mathcal{H}, p_{\mathrm{data}}, p_{\theta^*}) + 2\varepsilon_{\delta},$$

where

$$\varepsilon_{\delta} = 8K\sqrt{\frac{1}{M}\log\left[2(T+1)\delta\right]}.$$

In order to prove the general result, we begin with some technical lemmas. The development here owes much to Gretton et al. [2].

Lemma 1. *Let* $\mathcal{F} = \{f : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}\}$ *and*

$$\mathcal{F}_+ = \{h = f(y, \cdot) :\ f \in \mathcal{F},\ y \in \mathcal{Y}\} \cap B(L_\infty(\mathcal{Y})).$$

Let $\{Y_n\}_{n=1}^N$ *be* μ*-distributed independent random variables in* $\mathcal{Y}$*. Assume for some* $\gamma > 1$ *and some* $p \in \mathbb{N}$*, we have* $fat_{\varepsilon}(\mathcal{F}_+) \le \gamma\varepsilon^{-p}$*, for all* $\varepsilon > 0$*. For* $y_n \in \mathcal{Y} \quad \forall n = 1, \ldots, N$*, define* $\rho(y_1, \ldots, y_N)$ *to be*

$$\sup_{f \in \mathcal{F}} \left| E\left(f(Y, Y')\right) - \frac{1}{N(N-1)} \sum_{n \ne n'} f(y_n, y_{n'}) \right|.$$

Then there exists a constant C *that depends on* p*, such that*

$$E\left(\rho(Y_1, \ldots, Y_N)\right) \le \frac{C}{\sqrt{N-1}} \Psi(\gamma, N-1, p).$$

Proof. Let us introduce $\{\zeta_n\}_{n=1}^N$, where ζ_n and $Y_{n'}$ have the same distribution and are independent for all $n, n' \in \{1, \ldots, N\}$. Then the following is true:

$$E(f(Y, Y')) = E\Big(\frac{1}{N(N-1)} \sum_{n, n' :\, n \ne n'} f(\zeta_n, \zeta_{n'})\Big)$$

Using Jensen's inequality and the independence of Y, Y' and $Y_n, Y_{n'}$, we have

$$\begin{aligned}&E\left(\rho(Y_1, \ldots, Y_N)\right)\\ &= E\Big(\sup_{f \in \mathcal{F}} \Big| E(f(Y, Y'))\\ &\qquad\qquad - \frac{1}{N(N-1)} \sum_{n \ne n'} f(Y_m, Y_{m'}) \Big|\Big)\\ &\le E\Big(\sup_{f \in \mathcal{F}} \Big| \frac{1}{N(N-1)} \sum_{n \ne n'} f(\zeta_n, \zeta'_n)\\ &\qquad\qquad - \frac{1}{N(N-1)} \sum_{n \ne n'} f(Y_n, Y_{n'}) \Big|\Big).\end{aligned} \tag{9}$$

Introducing conditional expectations allows us to rewrite the equation with the sum over n outside the expectations. I.e., Eq. (9) equals

$$\begin{aligned}&\frac{1}{N} \sum_{n} E E^{(Y_n, \zeta_n)} \Big(\sup_{f \in \mathcal{F}} \Big| \frac{1}{N-1} \sum_{n \ne n'} \Phi(\zeta_n, \zeta_{n'}, Y_n, Y_{n'}) \Big| \Big)\\ &= E E^{(Y, \zeta)} \Big(\sup_{f \in \mathcal{F}} \Big| \frac{1}{N-1} \sum_{n=1}^{N-1} \sigma_n \Phi(\zeta, \zeta_n, Y, Y_n) \Big| \Big),\end{aligned} \tag{10}$$

where $\Phi(x, x', y, y') = f(x, x') - f(y, y')$. The second equality follows by symmetry of random variables $\{\zeta_n\}_{n=1}^{N-1}$. Note that we also added Rademacher random variables $\{\sigma_n\}_{n=1}^{N-1}$ before each term in the sum since $(f(\zeta_n, \zeta_{n'}) - f(Y_n, Y_{n'}))$ has the same distribution as $-(f(\zeta_n, \zeta_{n'}) - f(Y_n, Y_{n'}))$ for all n, n' and therefore the σ's do not affect the expectation of the sum.

Note that ζ_m and Y_m are identically distributed. Thus the triangle inequality implies that Eq. (10) is less than or equal to

$$\begin{aligned}&\frac{2}{N-1} E\left(E^{(Y)} \left(\sup_{f \in \mathcal{F}} \Big| \sum_{n=1}^{N-1} \sigma_n f(Y, Y_n) \Big| \right) \right)\\ &\qquad \le \frac{2}{\sqrt{N-1}} R_{N-1}(\mathcal{F}_+),\end{aligned}$$

where $R_{N-1}(\mathcal{F}_+)$ is the Rademacher's complexity of $\mathcal{F}_+$. Then by Theorem 3, we have

$$E\left(\rho(Y_1, \ldots, Y_N)\right) \le \frac{C}{\sqrt{N-1}} \Psi(\gamma, N-1, p).$$

□

Lemma 2. *Let* $\mathcal{F} = \{f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}$ *and* $\mathcal{F}_+ = \{f : x \times \mathcal{Y} \to \mathbb{R},\ x \in \mathcal{X}\}$ *and assume* $\mathcal{F}_+ \subset B(L_\infty(\mathcal{Y}))$*. Let* $\{X_n\}_{n=1}^N$ *and* $\{Y_m\}_{n=1}^M$ *be* ν*- and* μ*-distributed independent random variables in* $\mathcal{X}$ *and* $\mathcal{Y}$*, respectively. Assume for some* $\gamma > 1$*, such that for all* $\varepsilon > 0$*,* $fat_{\varepsilon}(\mathcal{F}_+) \le \gamma\varepsilon^{-p}$*,* *for some $p \in \mathbb{N}$. For all $x_n \in \mathcal{X}$, $n \leq N$, and all $y_m \in \mathcal{Y}$, $m \leq M$, define*

$$\rho(x_1, \ldots, x_N, y_1, \ldots, y_M) = \sup_{f \in \mathcal{F}} \Big| E(f(X,Y) - \frac{1}{NM} \sum_{n,m} f(x_n, y_m) \Big|.$$

Then there exists C that depends on p, such that

$$E\left(\rho(X_1, \ldots, X_N, Y_1, \ldots, Y_M)\right) \leq \frac{C}{\sqrt{M}} \Psi(\gamma, M, p).$$

Proof. The proof is very similar to that of Lemma 1. □

Proof of Theorem 1. The proof follows the same steps as the proof of Theorem 5 apart from a stronger uniform bound stated in Eq. (7). I.e., we need to show:

$$\Pr\left[\sup_{\theta \in \Theta} |\mathcal{E}(\theta) - \mathcal{T}(\theta)| \geq \varepsilon\right] \leq \delta.$$

Expanding MMD as defined by Eq. (4), and substituting $Y = G_\theta(W)$, yields

$$\begin{aligned} &\sup_{\theta \in \Theta} |\mathcal{E}(\theta) - \mathcal{T}(\theta)| \\ &= \sup_{\theta \in \Theta} \Big| E(k(X, X')) \\ &\qquad - \frac{1}{N(N-1)} \sum_{n' \neq n} k(X_n, X_{n'}) \\ &\qquad + E(k(G_\theta(W), G_\theta(W'))) \\ &\qquad - \frac{1}{M(M-1)} \sum_{m \neq m'} k(G_\theta(W_m), G_\theta(W_{m'})) \\ &\qquad - 2E(k(X, G_\theta(W))) \\ &\qquad + \frac{2}{MN} \sum_{m,n} k(X_n, G_\theta(W_m)) \Big|. \end{aligned} \tag{11}$$

For all $n \in \{1, \ldots, N\}$, $k(X_n, X_{n'})$ does not depend on θ and therefore the first two terms of the equation above can be taken out of the supremum. Also, note that since $|k(\cdot, \cdot)| \leq K$, we have

$$\Big| \zeta(x_1, \ldots, x_n, \ldots, x_N) - \zeta(x_1, \ldots, x'_n, \ldots, x_N) \Big| \leq \frac{2K}{N},$$

where

$$\zeta(x_1, \ldots, x_N) = \frac{1}{N(N-1)} \sum_{n,n' : n' \neq n} k(x_n, x_{n'}),$$

and ζ is an unbiased estimate of $E(k(X, X'))$. Then from McDiarmid's inequality on ζ, we have

$$\begin{aligned} &\Pr\Big(\Big| E(k(X, X')) - \frac{1}{N(N-1)} \sum_{n' \neq n} k(X_n, X_{n'}) \Big| \geq \varepsilon \Big) \\ &\qquad \leq \exp\left(-\frac{\varepsilon^2}{2K^2} N \right). \end{aligned} \tag{12}$$

Therefore Eq. (11) is bounded by the sum of the bound on Eq. (12) and the following:

$$\begin{aligned} &\sup_{\theta \in \Theta} \Big| E(k(G_\theta(W), G_\theta(W'))) \\ &\qquad - \frac{1}{M(M-1)} \sum_{m \neq m'} k(G_\theta(W_m), G_\theta(W_{m'})) \\ &\qquad - 2E(k(X, G_\theta(W))) \\ &\qquad + \frac{2}{MN} \sum_{m,n} k(X_n, G_\theta(W_m)) \Big|. \end{aligned} \tag{13}$$

Thus the next step is to find the bound for the supremum above.

Define

$$\begin{aligned} &f(W_1, \ldots, W_M; p_{\text{noise}}) = f(\underline{W}_M) \\ &= \sup_{\theta \in \Theta} \Big| E(k(G_\theta(W), G_\theta(W'))) \\ &\qquad - \frac{1}{M(M-1)} \sum_{m \neq m'} k(G_\theta(W_m), G_\theta(W_{m'})) \Big| \end{aligned}$$

and

$$\begin{aligned} &h(X_1, \ldots, X_N, W_1, \ldots, W_M; p_{\text{data}}, p_{\text{noise}}) \\ &= h(\underline{X}_N, \underline{W}_M) \\ &= \sup_{\theta \in \Theta} \Big| \frac{1}{MN} \sum_{m,n} k(X_n, G_\theta(W_m)) - E(k(X, G_\theta(W))) \Big|. \end{aligned}$$

Then by triangle inequality, the supremum in Eq. (13) is bounded by

$$f(\underline{W}_M) + 2h(\underline{X}_N, \underline{W}_M).$$

We will first find the upper bound on $f(\underline{W}_M)$, i.e., for every $\varepsilon > 0$, we will show that there exists δ_f, such that

$$\Pr\left(f(\underline{W}_M) > \varepsilon\right) \leq \delta_f \tag{14}$$

For each $m \in \{1, \ldots, M\}$,

$$\begin{aligned} &\Big| f(W_1, \ldots, W_m, \ldots, W_M) \\ &\qquad - f(W_1, \ldots, W'_m, \ldots, W_M) \Big| \leq \frac{2K}{M} \end{aligned}$$

since the kernel is bounded by K, and therefore $k(G_\theta(W_m), G_\theta(W_{m'}))$ is bounded by K for all m. The conditions of Theorem 2 are satisfied and thus we can use McDiarmids Inequality on f:

$$\Pr\left(f(\underline{W}_M) - E(f(\underline{W}_M)) \geq \epsilon\right) \leq \exp\left(-\frac{\varepsilon^2 M}{2K^2} \right).$$

Define

$$\mathcal{G}_k = \{k(G_\theta(\cdot), G_\theta(\cdot)) : \theta \in \Theta\}$$

To show Eq. (14), we need to bound the expectation of f. We can apply Lemma 1 on the function classes $\mathcal{G}_k$ and $\mathcal{G}_{k+}$. The resulting bound is

$$E(f(\underline{W}_M)) \leq \varepsilon_{p1} = \frac{C_f}{\sqrt{M-1}} \Psi(\gamma_1, M-1, p_1), \quad (15)$$

where p_1 and γ_1 are parameters associated with fat shattering dimension of $\mathcal{G}_{k+}$ as stated in the assumptions of the theorem, and C_f is a constant depending on p_1.

Now we can write down the bound on f:

$$\Pr\left(f(\underline{W}_M) \geq \varepsilon_{p_1} + \epsilon\right) \leq \exp\left(-\frac{\varepsilon^2 M}{2K^2}\right) = \delta_f. \quad (16)$$

Similarly, $h(\underline{X}_N, \underline{W}_M)$ has bounded differences:

$$\Big| h(X_1, \ldots, X_n, \ldots, X_N, W_1, \ldots, W_M) \\ - h(X_1, \ldots, X_{n'}, \ldots, X_N, W_1, \ldots, W_M) \Big| \leq \frac{2K}{N}$$

and

$$\Big| h(X_1, \ldots, X_N, W_1, \ldots, W_m, \ldots, W_M) \\ - h(X_1, \ldots, X_N, W_1, \ldots, W_{m'}, \ldots, W_M) \Big| \leq \frac{2K}{M}.$$

McDiarmid's inequality then implies

$$\Pr\left(h(\underline{X}_N, \underline{W}_M) - E(h(\underline{X}_N, \underline{W}_M) \geq \varepsilon\right) \\ \leq \exp\left(-\frac{\varepsilon^2}{2K^2} \frac{NM}{N+M}\right). \quad (17)$$

We can bound expectation of $h(\underline{X}_N, \underline{W}_M)$ using Lemma 2 applied on $\mathcal{G}_k^{\mathbb{X}}$ and $\mathcal{G}_{k+}^{\mathbb{X}}$, where

$$\mathcal{G}_k^{\mathbb{X}} = \{k(\cdot, G_\theta(\cdot)) : \theta \in \Theta\}.$$

Then

$$E(h(\underline{X}_N, \underline{W}_M)) \leq \varepsilon_{p_2} = \frac{C_h}{\sqrt{M}} \Psi(\gamma_2, M, p_2). \quad (18)$$

for some constant C_h that depends on $p_@$. The final bound on h is then

$$\Pr\left(h(\underline{X}_N, \underline{W}_M) \geq \varepsilon_{p_2} + \varepsilon\right) \\ \leq \exp\left(-\frac{\varepsilon^2}{2K^2} \frac{NM}{N+M}\right) = \delta_h.$$

Summing up the bounds from Eq. (16) and Eq. (17), it follows that

$$\Pr\left(f(\underline{W}_M) + 2h(\underline{X}_N, \underline{W}_M) \geq \varepsilon_{p1} + 2\varepsilon_{p_2} + 3\varepsilon\right) \\ \leq \max(\delta_f, \delta_h) = \delta_h.$$

Using the bound in Eq. (12), we have obtain the uniform bound we were looking for:

$$\Pr\left[\sup_{\theta \in \Theta} |\mathcal{E}(\theta) - \mathcal{T}(\theta)| > \varepsilon_{p_1} + 2\varepsilon_{p_2} + 4\varepsilon\right] \leq \delta_h,$$

which by Eq. (7) yields

$$\Pr\left[|\mathcal{E}(\hat{\theta}) - \mathcal{T}(\hat{\theta})| > \varepsilon_{p_1} + 2\varepsilon_{p_2} + 4\varepsilon\right] \leq \delta_h.$$

Since it was assumed that $K = 1$ and $N = M$, we get $\delta_h = \exp\left(-\varepsilon^2 M/4\right)$.

To finish, we proceed as in the proof of Theorem 5. We can rearrange some of the terms to get a different form of Eq. (6):

$$\Pr\left[|\mathcal{E}(\theta^*) - \mathcal{T}(\theta^*)| > 2\varepsilon\right] \leq 2\exp\left(-\frac{\varepsilon^2 M}{4}\right) = 2\delta_h.$$

All of the above implies that for any $\varepsilon > 0$, there exists δ, such that

$$\Pr\big(\mathrm{MMD}^2(\mathcal{H}, p_{\mathrm{data}}, p_{\hat{\theta}}) - \mathrm{MMD}^2(\mathcal{H}, p_{\mathrm{data}}, p_{\theta^*}) \geq \varepsilon\big) \leq \delta,$$

where

$$\varepsilon = \varepsilon_{p_1} + 2\varepsilon_{p_2} + \frac{12}{\sqrt{M}} \sqrt{\log \frac{2}{\delta}}.$$

We can rewrite ε as:

$$\varepsilon = r(p_1, \gamma_1, M) + r(p_2, \gamma_2, M-1) + 12M^{-\frac{1}{2}} \sqrt{\log \frac{2}{\delta}},$$

The rate $r(p, \gamma, N)$ is given by Eq. (15) and Eq. (18):

$$r(p, \gamma, M) = C_p \sqrt{\gamma} \begin{cases} M^{-\frac{1}{2}} & \text{if } p < 2, \\ M^{-\frac{1}{2}} \log^{\frac{3}{2}}(M) & \text{if } p = 2, \\ M^{-\frac{1}{p}} & \text{if } p > 2, \end{cases}$$

where C_{p_1} and C_{p_2} depend on p_1 and p_2 alone. □

We close by noting that the approximation error is zero in the nonparametric limit.

Theorem 6 (Gretton et al. [2]). *Let F be the unit ball in a universal RKHS $\mathcal{H}$, defined on the compact metric space $\mathbb{X}$, with associated continuous kernel $k(\cdot, \cdot)$. Then $\mathrm{MMD}[\mathcal{H}, p, q] = 0$ if and only if $p = q$.*

Corollary 2 (approximation error). *Assume p_{data} is in the family $\{p_\theta\}$ and that $\mathcal{H}$ is an RKHS induced by a characteristic kernel. Then*

$$\inf_\theta \mathrm{MMD}(\mathcal{H}, p_{\mathrm{data}}, p_\theta) = 0$$

and the infimum is achieved at θ satisfying $p_\theta = p_{\mathrm{data}}$.

Proof. By Theorem 6, it follows that $\mathrm{MMD}^2(\mathcal{H}, \cdot, \cdot)$ is a metric. The result is then immediate. □

Acknowledgments

The authors would like to thank Bharath Sriperumbudur for technical discussions.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative Adversarial Nets”. 2014.

[2] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. “A Kernel Two-sample Test”. In: *J. Mach. Learn. Res.* 13 (Mar. 2012), pp. 723–773.

[3] G. E. Hinton and R. R. Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: *Science* 313.5786 (July 2006), pp. 504–507.

[4] O. Kallenberg. *Foundations of modern probability*. 2nd. New York: Springer, 2002, pp. xx+638.

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE*. 1998, pp. 2278–2324.

[6] Y. Li, K. Swersky, and R. Zemel. “Generative Moment Matching Networks”. http://arxiv.org/abs/1502.02761v1. 2015.

[7] D. J. MacKay. “Bayesian Neural Networks and Density Networks”. In: *Nuclear Instruments and Methods in Physics Research, A*. 1994, pp. 73–80.

[8] S. Mendelson. “A Few Notes on Statistical Learning Theory”. English. In: *Advanced Lectures on Machine Learning*. Ed. by S. Mendelson and A. Smola. Vol. 2600. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2003, pp. 1–40.

[9] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[10] R. Salakhutdinov and G. E. Hinton. “Deep Boltzmann Machines”. In: *Journal of Machine Learning Research - Proceedings Track* 5 (2009), pp. 448–455.

[11] E. Saund. “Dimensionality-Reduction Using Connectionist Networks.” In: *IEEE Trans. Pattern Anal. Mach. Intell.* 11.3 (1989), pp. 304–314.

[12] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. “Injective Hilbert Space Embeddings of Probability Measures”. In: *Conf. Comp. Learn. Theory, (COLT)*. 2008.

[13] J. M. Susskind, A. K. Anderson, and G. E. Hinton. *The Toronto face database*. Tech. rep. 2010.

[14] T. Tieleman and G. Hinton. *Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude*. COURSERA: Neural Networks for Machine Learning. 2012.

[15] B. Uria, I. Murray, and H. Larochelle. “A Deep and Tractable Density Estimator.” In: *CoRR* abs/1310.1757 (2013).