Proceedings of Machine Learning Research vol 145:1013–1043, 2021 2nd Annual Conference on Mathematical and Scientific Machine Learning

# Generalization and Memorization: The Bias Potential Model

**Hongkang Yang** HONGKANG@PRINCETON.EDU
*Program in Applied and Computational Mathematics, Princeton University*

**Weinan E** WEINAN@MATH.PRINCETON.EDU
*Department of Mathematics and PACM, Princeton University*

**Editors:** Joan Bruna, Jan S Hesthaven, Lenka Zdeborova

## Abstract

Models for learning probability distributions such as generative models and density estimators behave quite differently from models for learning functions. One example is found in the memorization phenomenon, namely the ultimate convergence to the empirical distribution, that occurs in generative adversarial networks (GANs). For this reason, the issue of generalization is more subtle than that for supervised learning. For the bias potential model, we show that dimension-independent generalization accuracy is achievable if early stopping is adopted, despite that in the long term, the model either memorizes the samples or diverges.

**Keywords:** probability distribution, machine learning, generalization error, curse of dimensionality, early stopping.

## 1. Introduction

Distribution learning models such as GANs have achieved immense popularity from their empirical success in learning complex high-dimensional probability distributions, and they have found diverse applications such as generating images (Brock et al., 2018) and paintings (Elgammal et al., 2017), writing articles (Brown et al., 2020), composing music (Oord et al., 2016), editing photos (Bau et al., 2019), designing new drugs (Prykhodko et al., 2019) and new materials (Mao et al., 2020), generating spin configurations (Zhang et al., 2018) and modeling quantum gases (Casert et al., 2020), to name a few.

As a mathematical problem, distribution learning is much less understood. Arguably, the most fundamental question is the generalization ability of these models. One puzzling issue is the following.

0. Generalization vs. memorization:

   Let $Q_*$ be the target distribution, and $Q_*^{(n)}$ be the empirical distribution associated with $n$ sample points. Let $Q(f)$ be the probability distribution generated by some machine learning model parametrized by the function $f$ in some hypothesis space $\mathcal{F}$. It has been argued, for example in the case of GAN, that as training proceeds, one has (Goodfellow et al., 2014)

$$\lim_{t\to\infty} Q(f(t)) = Q_*^{(n)} \tag{1}$$

   where $f(t)$ is the parameter we obtain at training step $t$. We refer to (1) as the "memorization phenomenon". When it happens, the model learned does not give us anything other than the samples we already have. This is in sharp contrast to supervised learning, where models are typically trained till interpolation and can generalize well to unseen data both in practice (Zhang et al., 2016) and in theory (E et al., 2019d).

Despite this, these distribution-learning models perform surprisingly well in practice, being able to come close to the unseen target $Q_*$ and allowing us to generate new samples. This counterintuitive result calls for a closer examination of their training dynamics, beyond the statement (1).

There are many other mysteries for distribution learning, and we list a few below.

1. Curse of dimensionality:

   The superb performance of these models (e.g. on generating high-resolution, lifelike and diverse images (Brock et al., 2018; Donahue and Simonyan, 2019; Vahdat and Kautz, 2020)) indicates that they can approximate the target $Q_*$ with satisfactorily small error. Yet, in theory, this should not be possible, because to estimate a general distribution in $\mathbb{R}^d$ with error $\leq \epsilon$, we need $n = \epsilon^{-\Omega(d)}$ amount of samples (discussed below), which becomes astronomical for real-world tasks. For instance, the BigGAN (Brock et al., 2018) was trained on the ILSVRC dataset (Russakovsky et al., 2015) with $\leq 10^7$ images of resolution $512 \times 512$, but the theoretical sample size should be like $\gg 10^{512\times 512}$.

   Of course, for restricted distribution families like the Gaussians, the sample complexity is only $n = \mathsf{poly}(d)$. Yet, one is really interested in complex distributions such as the distribution of facial images that *a priori* do not belong to any known family, so these tasks require the models to possess not only a dimension-independent sample complexity but also the universal approximation property.

2. The fragility of the training process:

   It is well-known that distribution-learning models like GANs and VAE (variational autoencoder) are difficult to train. They are especially vulnerable to issues like mode collapse (Che et al., 2016; Salimans et al., 2016), instability and oscillation (Radford et al., 2015), and vanishing gradient (Arjovsky and Bottou, 2017). The current treatment is to find by trial-and-error a delicate combination of the right architectures and hyper-parameters (Radford et al., 2015). The need to understand these issues calls for a mathematical treatment.

This paper offers a partial answer to these questions. We focus on the bias potential model, an expressive distribution-learning model that is relatively transparent, and uncover the mechanisms for its generalization ability.

Specifically, we establish a dimension-independent *a priori* generalization error estimate with early-stopping. With appropriate function spaces $f \in \mathcal{F}$, the training process consists of two regimes:

- First, by implicit regularization, the training trajectory $Q(f(t))$ comes very close to the unseen target $Q_*$, and this is when early-stopping should be performed.
- Afterwards, $Q(f(t))$ either converges to the sample distribution $Q_*^{(n)}$ or it diverges.

This paper is structured as follows. In Section 2, we introduce the bias potential model and pose it as a continuous calculus of variations problem. Section 3 analyzes the training behavior of this model and presents this paper's main results on generalization error and memorization. Section 4 presents some numerical examples. Section A contains all the proofs. Section 5 concludes this paper with remarks on future directions.

Notation: denote vectors by bold letters $\mathbf{x}$. Let $C(K)$ be the space of continuous functions over some subset $K \subseteq \mathbb{R}^d$ equipped with supremum norm. Let $\mathcal{P}(K), \mathcal{P}_{ac}(K), \mathcal{P}_2(K)$ be the space of probability measures over $K$, the subset of absolutely continuous measures, and the subset of measures with finite second moments. Denote the support of a distribution $Q \in \mathcal{P}(K)$ by sprt$Q$. Let $W_2$ be the Wasserstein metric over $\mathcal{P}_2(K)$.

### 1.1. Related works

- Generalization ability: Among distribution-learning models, GANs have attracted the most attention and their generalization ability has been discussed in (Arora et al., 2017; Zhang et al., 2017; Bai et al., 2019; Gulrajani et al., 2020) from the perspective of the neural network-based distances. For trained models, dimension-independent generalization error estimates have been obtained only for certain restricted models, such as GANs whose generators are linear maps or one-layer networks (Wu et al., 2019; Lei et al., 2020; Feizi et al., 2020).

- Curse of dimensionality (CoD): If the sampling error is measured by the Wasserstein metric $W_2$, then for any absolutely continuous $Q_*$ and any $\delta > 0$, it always holds that (Weed and Bach, 2017)
$$W_2(Q_*^{(n)}, Q_*) \gtrsim n^{-\frac{1}{d-\delta}}$$
  To achieve an error of $\epsilon$, the required sample size is $n = \epsilon^{-\Omega(d)}$.

  If sampling error is measured by KL divergence, then $KL(Q_* \| Q_*^{(n)}) = \infty$ since $Q_*^{(n)}$ is singular. If kernel smoothing is applied to $Q_*^{(n)}$, it is known that the error scales like $O(n^{-\frac{4}{d+4}})$ (Wand and Jones, 1994) (technically the norm used in (Wand and Jones, 1994) is the $L^2$ difference between densities, but one should expect that CoD would likewise be present for KL divergence.)

- Continuous perspective: (E et al., 2019c, 2020) provide a framework to study supervised learning as continuous calculus of variations problems, with emphasis on the role of the function representation, e.g. continuous neural networks (E et al., 2019b). In particular, the function representation largely determines the trainability (Chizat and Bach, 2018; Rotskoff and Vanden-Eijden, 2019) and generalization ability (E et al., 2018, 2019a,d) of a supervised learning model. This framework can be applied to studying distribution learning in general, and we use it to analyze the bias potential model.

- Exponential family:

  The density function of the bias-potential model is an instance of the exponential families. These distributions have long been applied to density estimation (Barron and Sheu, 1991; Canu and Smola, 2006) with theoretical guarantees (Yuan et al., 2012; Sriperumbudur et al., 2017). Yet, existing theories focus only on black-box estimators, instead of the training process. It has also been popular to adopt a mixture of exponential distributions (Kiefer and Wolfowitz, 1956; Jewell, 1982; Redner and Walker, 1984), but it will not be covered in this paper.

## 2. Bias Potential Model

This section introduces the bias potential model, a simple distribution-learning model proposed by (Valsson and Parrinello, 2014; Bonati et al., 2019) and also known as "variationally enhanced sampling".

To pose a supervised-learning model as a calculus of variations problem, one needs to consider four factors: function representation, training objective, training rule, and discretization (E et al., 2019c). For distribution learning, there is the additional factor of distribution representation, namely how probability distributions are represented through functions. These are general issues for any distribution-learning model. For future reference, we go through these components in some detail.

### 2.1. Distribution representation

The bias potential model adopts the following representation:

$$Q = \frac{1}{Z} e^{-V} P, \quad Z = \mathbb{E}_P[e^{-V}] \tag{2}$$

where $V$ is some potential function and $P$ is some base distribution. This representation commonly appears in statistical mechanics as the Boltzmann distribution. It is suitable for density estimation, and can also be applied to generative modeling via sampling techniques like rejection sampling, MCMC, Langevin diffusion (Roberts and Tweedie, 1996), hit-and-run (Lovász and Vempala, 2007), etc.

Typically the partition function $Z$ can be ignored, since it is not involved in the training objectives or most of the sampling algorithms.

### 2.2. Training objective

Since the representation (2) is defined by a density function, it is natural to define a density-based training objective. Given a target distribution $Q_*$, one convenient choice is the backward KL divergence

$$KL(Q_* \| Q) = \mathbb{E}_{Q_*}[\log Q_* - \log P] + \mathbb{E}_{Q_*}[V] + \log \mathbb{E}_P[e^{-V}]$$

An alternative way introduced in (Valsson and Parrinello, 2014) is to define the "biased distribution"

$$P_* = \frac{e^V Q_*}{\mathbb{E}_{Q_*}[e^V]}$$

so that $Q = Q_*$ iff $P = P_*$. Then, we can define an objective by the forward KL

$$KL(P \| P_*) = \mathbb{E}_P[\log P - \log Q_*] - \mathbb{E}_P[V] + \log \mathbb{E}_{Q_*}[e^V]$$

Removing constant terms, we obtain the following objectives

$$\begin{aligned} L^-(V) &:= \mathbb{E}_{Q_*}[V] + \log \mathbb{E}_P[e^{-V}] \\ L^+(V) &:= -\mathbb{E}_P[V] + \log \mathbb{E}_{Q_*}[e^V] \end{aligned} \tag{3}$$

Both objectives are convex in $V$ (Lemma 14). Suppose $Q_*$ can be written as (2) with potential $V_*$, then $Q = Q_*$ iff $V = V_* + c$ for some constant $c$ iff $L^+(V)$ or $L^-(V) = 0$, so we have a unique global minimizer up to constants. Otherwise, if $Q_*$ does not have the form (2), then the minimizer does not exist.

In practice, when $Q_*$ is available only through its sample $Q_*^{(n)}$, we simply substitute all the expectation terms $\mathbb{E}_{Q_*}$ in (3) by $\mathbb{E}_{Q_*^{(n)}}$.

### 2.3. Function representation

A good function representation (or function space $\mathcal{F}$) should have two conflicting properties:

1. $\mathcal{F}$ is expressive so that distributions generated by $f \in \mathcal{F}$ satisfy universal approximation property.
2. $\mathcal{F}$ has small complexity so that the generalization gap is small.

One approach is to adopt an integral transform-based representation (E et al., 2019c),

$$V(\mathbf{x}) = \mathbb{E}_{\rho(\theta)}\big[\phi(\mathbf{x}, \theta)\big]$$

for some feature function $\phi(\cdot;\theta)$ and parameter distribution $\rho$. Then, $V$ can be approximated with Monte-Carlo rate by

$$V_m(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^{m} \phi(\mathbf{x}; \theta_i) \tag{4}$$

where $\{\theta_j\}$ are i.i.d. samples from $\rho(\theta)$.

Let us consider function representations built from neural networks:

- 2-layer neural networks: Define the continuous 2-layer network by

$$V(\mathbf{x}) = \mathbb{E}_{\rho(a,\mathbf{w},b)}\big[a\,\sigma(\mathbf{w}\cdot\mathbf{x} + b)\big] \tag{5}$$

  with an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ and weights $\mathbf{w} \in \mathbb{R}^d$ and $a, b \in \mathbb{R}$. The natural functional norm is the Barron norm (E et al., 2019b):

$$\|V\|_{\mathcal{B}} := \inf_{\rho} \|\rho\|_P, \quad \|\rho\|_P^2 := \mathbb{E}_{\rho(a,\mathbf{w},b)}\big[a^2(\|\mathbf{w}\|^2 + b^2)\big] \tag{6}$$

  where $\rho$ ranges over all parameter distributions that satisfy (5) and $\|\rho\|_P$ is known as the path norm.

- Random feature model: Rewrite (5) as

$$V(\mathbf{x}) = \mathbb{E}_{\rho_0(\mathbf{w},b)}\big[a(\mathbf{w}, b)\,\sigma(\mathbf{w}\cdot\mathbf{x} + b)\big] \tag{7}$$

  with fixed parameter distribution $\rho_0(\mathbf{w}, b)$ and

$$a(\mathbf{w}, b) := \frac{d\int a\, d\rho(a, \mathbf{w}, b)}{d\int \rho(a, \mathbf{w}, b) da}$$

  The natural functional norm is the RHKS (reproducing kernel Hilbert space) norm (E et al., 2019b; Rahimi and Recht, 2008):

$$\|V\|_{\mathcal{H}}^2 := \mathbb{E}_{\rho_0}\big[a(\mathbf{w}, b)^2\big] = \|a\|_{L^2(\rho_0)}^2 \tag{8}$$

  It corresponds to the RKHS $\mathcal{H}$ induced by the kernel

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{\rho_0(\mathbf{w},b)}[\sigma(\mathbf{w}\cdot\mathbf{x} + b)\sigma(\mathbf{w}\cdot\mathbf{x}' + b)]$$

It is straightforward to establish the universal approximation theorem for these two representations and we provide such results below: Denote by $\mathcal{P}_{ac}(K) \cap C(K)$ the distributions over $K$ with continuous density functions, and by $\|\cdot\|_{TV}$ the total variation distance, which is equivalent to the $L_1$ norm when restricted to $\mathcal{P}_{ac}(\mathbb{R}^d)$.

**Proposition 1 (Universal approximation)** *Let $K \subseteq \mathbb{R}^d$ be any compact set with positive Lebesgue measure, let $P$ be the uniform distribution over $K$, and let $\mathcal{V}$ be any class of functions that is dense in $C(K)$. Then, the class of probability distributions (2) generated by $V \in \mathcal{V}$ and $P$ are*

- *dense in $\mathcal{P}(K)$ under the Wasserstein metric $W_p$ ($1 \leq p < \infty$),*
- *dense in $\mathcal{P}_{ac}(K)$ under the total variation norm $\|\cdot\|_{TV}$,*
- *dense in $\mathcal{P}_{ac}(K) \cap C(K)$ under KL divergence.*

*Given assumption A.1, this result applies if $\mathcal{V}$ is the Barron space $\{\|V\|_{\mathcal{B}} < \infty\}$ or RKHS space $\{\|V\|_{\mathcal{H}} < \infty\}$.*

The Monte-Carlo approximation (4) suggests that these continuous models can be approximated efficiently by finite neural networks. Specifically, we can establish the following *a priori* error estimates:

**Proposition 2 (Efficient approximation)** *Suppose that the base distribution $P$ is compactly-supported in a ball $B_R(0)$, and the activation function $\sigma$ is Lipschitz with $\sigma(0) = 0$. Given $\|V\|_{\mathcal{B}} < \infty$, for every $m \in \mathbb{N}$, there exists a finite 2-layer network $V_m$ with $m$ neurons that satisfies:*

$$KL(Q\|Q_m) \leq \frac{\|V\|_{\mathcal{B}}}{\sqrt{m}} \cdot 2\sqrt{3}\|\sigma\|_{Lip}\sqrt{R^2+1}$$
$$\|V_m\|_{\mathcal{B}} \leq \sqrt{2}\|V\|_{\mathcal{B}}$$

*where $Q_m$ is the distribution generated by $V_m$. Similarly, assume that the fixed parameter distribution $\rho_0$ in (7) is compactly-supported in a ball $B_r(0)$, then given $\|V\|_{\mathcal{H}} < \infty$, for every $m$, there exists $V_m$ such that*

$$KL(Q\|Q_m) \leq \frac{\|V\|_{\mathcal{H}}}{\sqrt{m}} \cdot 2\sqrt{3}\|\sigma\|_{Lip}\sqrt{R^2+1}\, r$$
$$\|V_m\|_{\mathcal{H}} \leq \sqrt{2}\|V\|_{\mathcal{H}}$$

### 2.4. Training rule

We consider the simplest training rule, the gradient flow.

For continuous function representations, there are generally two kinds of flows:

- Non-conservative gradient flow

  For the random feature model (7), we can train the function $a(\mathbf{w}, b)$ using its variational gradient
  $$\partial_t a(\mathbf{w}, b) = -\frac{\delta L}{\delta a}(\mathbf{w}, b)$$

Specifically, for the training objectives $L^{\pm}(V)$ in (3), the corresponding flows are defined by

$$\frac{d}{dt}a = -\frac{\delta L^+}{\delta a} = -\mathbb{E}_{P_*-P}[\sigma(\mathbf{w}\cdot\mathbf{x}+b)]$$
$$\frac{d}{dt}a = -\frac{\delta L^-}{\delta a} = -\mathbb{E}_{Q_*-Q}[\sigma(\mathbf{w}\cdot\mathbf{x}+b)]$$

- Conservative gradient flow:

  For the 2-layer neural network (5), we train the parameter distribution $\rho(a,\mathbf{w},b)$. Its gradient flow is constrained by the conservation of local mass and obeys the continuity equation (in the weak sense):

$$\partial_t\rho - \nabla\cdot\big(\rho\nabla\frac{\delta L}{\delta\rho}\big) = 0 \tag{9}$$

  With the objectives (3), the gradient fields are given by

$$\begin{aligned}\nabla\frac{\delta L^+}{\delta\rho} &= \nabla_{(a,\mathbf{w},b)}\mathbb{E}_{P_*-P}\big[a\,\sigma(\mathbf{w}\cdot\mathbf{x}+b)\big] = \mathbb{E}_{P_*-P}\begin{bmatrix}\sigma(\mathbf{w}\cdot\mathbf{x}+b)\\ a\,\sigma'(\mathbf{w}\cdot\mathbf{x}+b)\,\mathbf{x}\\ a\,\sigma'(\mathbf{w}\cdot\mathbf{x}+b)\end{bmatrix}\\ \nabla\frac{\delta L^-}{\delta\rho} &= \nabla_{(a,\mathbf{w},b)}\mathbb{E}_{Q_*-Q}\big[a\,\sigma(\mathbf{w}\cdot\mathbf{x}+b)\big] = \mathbb{E}_{Q_*-Q}\begin{bmatrix}\sigma(\mathbf{w}\cdot\mathbf{x}+b)\\ a\,\sigma'(\mathbf{w}\cdot\mathbf{x}+b)\,\mathbf{x}\\ a\,\sigma'(\mathbf{w}\cdot\mathbf{x}+b)\end{bmatrix}\end{aligned} \tag{10}$$

### 2.5. Discretization

So far we have only discussed the continuous formulation of distribution learning models. In practice, we implement these continuous models using discretized versions, with the hope that the discretized models inherit these properties up to a controllable discretization error.

Let us focus on the discretization in the parameter space, and in particular, the most popular "particle discretization", since this is the analog of Monte-Carlo for dynamic problems. Consider the parameter distribution $\rho(a,\mathbf{w},b)$ of the 2-layer net (5) and its approximation by the empirical distribution

$$\rho^{(m)} = \frac{1}{m}\sum_{j=1}^{m}\delta_{(a_j,\mathbf{w}_j,b_j)}$$

where the particles $\{(a_j,\mathbf{w}_j,b_j)\}$ are i.i.d. samples of $\rho$. The potential function represented by this empirical distribution is given by:

$$V_m(\mathbf{x}) = \mathbb{E}_{\rho^{(m)}}\big[a\,\sigma(\mathbf{w}\cdot\mathbf{x}+b)\big] = \frac{1}{m}\sum_j a_j\,\sigma(\mathbf{w}_j\cdot\mathbf{x}+b_j)$$

Suppose we train $\rho^{(m)}$ by conservative gradient flow (9,10) with the objective $L^-$. The continuity equation (9) implies that, for any smooth test function $f(a,\mathbf{w},b)$, we have

$$\frac{d}{dt}\int f\,d\rho^{(m)} = -\int\nabla f\cdot\nabla\frac{\delta L}{\delta\rho}d\rho^{(m)} = -\frac{1}{m}\sum_{j=1}^{m}\nabla f(a_j,\mathbf{w}_j,b_j)^T\cdot\nabla\frac{\delta L}{\delta\rho}(a_j,\mathbf{w}_j,b_j)$$

Meanwhile, we also have

$$\frac{d}{dt}\int f\, d\rho^{(m)} = \frac{1}{m}\sum_{j=1}^{m} \nabla f(a_j, \mathbf{w}_j, b_j)^T \cdot \frac{d}{dt}\begin{bmatrix} a_j \\ \mathbf{w}_j \\ b_j \end{bmatrix}$$

Thus we have recovered the gradient flow for finite *scaled* 2-layer networks:

$$\frac{d}{dt}\begin{bmatrix} \mathbf{a}_j \\ \mathbf{w}_j \\ b_j \end{bmatrix} = -\nabla \frac{\delta L^-(V_m)}{\delta \rho}(a_j, \mathbf{w}_j, b_j) = -m \cdot \frac{\partial L^-(V_m)}{\partial (a_j, \mathbf{w}_j, b_j)} = -\mathbb{E}_{Q_* - Q}\begin{bmatrix} \sigma(\mathbf{w}_j \cdot \mathbf{x} + b_j) \\ a_j\, \sigma'(\mathbf{w}_j \cdot \mathbf{x} + b_j)\, \mathbf{x} \\ a_j\, \sigma'(\mathbf{w}_j \cdot \mathbf{x} + b_j) \end{bmatrix}$$

This example shows that the particle discretization of continuous 2-layer networks (5) leads to the same result as the mean-field modeling of 2-layer nets (Mei et al., 2018; Rotskoff and Vanden-Eijden, 2019).

## 3. Training Dynamics

This section studies the training behavior of the bias potential model and presents the main result of this paper, on the relation between generalization and memorization: When trained on a finite sample set,

- With early stopping, the model reaches dimension-independent generalization error rate.
- As $t \to \infty$, the model necessarily memorizes the samples unless it diverges.

### 3.1. Trainability

We begin with the training dynamics on the population loss. First, we consider the random feature model (7) and establish global convergence:

**Proposition 3 (Trainability)** *Suppose that the target distribution $Q_*$ is generated by a potential $V_*$ ($\|V_*\|_{\mathcal{H}} < \infty$). Suppose that our distribution $Q_t$ is generated by potential $V_t$ with parameter function $a_t$ trained by gradient flow on either of the objectives (3). Then,*

$$L^{\pm}(V_t) - L^{\pm}(V_*) \leq \frac{||V_* - V_0||^2_{\mathcal{H}}}{2t}$$

Next, for 2-layer neural networks, we show that whenever the conservative gradient flow converges, it must converge to the global minimizer. In particular, it will not be trapped at bad local minima and thus avoids mode collapse. This result is analogous to the global optimality guarantees for supervised learning and regression problems (Chizat and Bach, 2018; Rotskoff and Vanden-Eijden, 2019).

**Proposition 4** *Assume that the distribution $Q_t$ is generated by potential $V_t$, a 2-layer network with parameter distribution $\rho_t$ trained by gradient flow on either of the objectives (3). Assume that the assumption A.2 holds. If the flow $\rho_t$ converges in $W_1$ metric (or any $W_p$, $1 \leq p \leq \infty$) to some $\rho_\infty$ as $t \to \infty$, then $\rho_\infty$ is a global minimizer of $L^{\pm}$: Let $V_\infty$ be the corresponding 2-layer network, then*

$$Q_* = Q_\infty = \frac{e^{-V_\infty} P}{\mathbb{E}_P[e^{-V_\infty}]}$$

### 3.2. Generalization ability

Now we consider the most important issue for the model, the generalization error, and prove that a dimension-independent *a priori* error rate is achievable within a convenient early-stopping time interval.

We study the training dynamics on the empirical loss. For convenience, we make the following assumptions:

- Let the base distribution $P$ in (2) be supported on $[-1,1]^d$ (the $l^\infty$ ball). Without loss of generality, we use the $l^\infty$ norm on $[-1,1]^d$.
- Let the objective $L$ be $L^-(V)$ from (3) (The analysis of $L^+$ would be more involved). Recall that if the target $Q_*$ is generated by a potential $V_*$, then

$$L(V) - L(V_*) = KL(Q_* \| Q)$$

  Denote by $L^{(n)}$ the objective that corresponds to $Q_*^{(n)}$:

$$L^{(n)}(V) = \mathbb{E}_{Q_*^{(n)}}[V] + \log \mathbb{E}_P[e^{-V}] = L(V) + \mathbb{E}_{Q_*^{(n)} - Q_*}[V]$$

- Model $V$ by the random feature model (7) with RKHS norm $\|V\|_{\mathcal{H}} = \|a\|_{L^2(\rho_0)}$ from (8). Assume that the activation function $\sigma$ is ReLU, and that the fixed parameter distribution $\rho_0$ is supported inside the $l^1$ ball, that is, $\|\mathbf{w}\|_1 + |b| \leq 1$ for $\rho_0$ almost all $(\mathbf{w}, b)$. Denote $(\mathbf{x}, 1)$ by $\tilde{\mathbf{x}}$ and $(\mathbf{w}, b)$ by $\mathbf{w}$, so the activation can be written as $\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}})$.

  **Remark 5 (Universal approximation)** *If we further assume that $\rho_0$ covers all directions (e.g. $\rho_0$ is uniform over the $l^1$ sphere $\{\|\mathbf{w}\|_1 + |b| = 1\}$) and $P$ is uniform over some $K \subseteq [-1,1]^d$, then Proposition 1 implies that this model enjoys universal approximation over distributions on $K$.*

- Training rule: We train $a$ by gradient flow (Section 2.4). Let $a_t, V_t, Q_t$ and $a_t^{(n)}, V_t^{(n)}, Q_t^{(n)}$ be the training trajectories under $L$ and $L^{(n)}$. Assume the same initialization $a_0 = a_0^{(n)}$.

**Theorem 6 (Generalization ability)** *Suppose $Q_*$ is generated by a potential function $V_*$ ($\|V_*\|_{\mathcal{H}} < \infty$). For any $\delta \in (0,1)$, with probability $1-\delta$ over the sampling of $Q_*^{(n)}$, the testing error of $Q_t^{(n)}$ is bounded by*

$$KL\big(Q_* \| Q_t^{(n)}\big) \leq \frac{\|V_* - V_0\|_{\mathcal{H}}^2}{2t} + 2\Big(4\frac{\sqrt{2\log 2d}}{\sqrt{n}} + \frac{\sqrt{2\log(2/\delta)}}{\sqrt{n}}\Big)t$$

**Corollary 7** *Given the condition of Theorem 6, if we choose an early-stopping time $T$ such that*

$$T = \Theta\Big(\|V_* - V_0\|_{\mathcal{H}}\big(\frac{n}{\log d}\big)^{1/4}\Big)$$

*then the testing error obeys*

$$KL\big(Q_* \| Q_T^{(n)}\big) \lesssim \|V_* - V_0\|_{\mathcal{H}}\Big(\frac{\log d}{n}\Big)^{1/4}$$

This rate is significant in that it is dimension-independent up to a negligible $(\log d)^{1/4}$ term. Although the upper bound $n^{-1/4}$ is slower than the desirable Monte-Carlo rate of $n^{-1/2}$, it is much better than the rate $n^{-1/d}$ and we believe there is room for improvement. In addition, the early-stopping time interval is reachable within a time that is dimension-independent and the width of this interval is at least on the order of $n^{1/4}$.

This result is enabled by the function representation of the model, specifically:

1. Learnability: If the target $Q_*$ lives in the right space for our function representation, then the optimization rate (for the population loss $L(V_t) - L(V_*)$) is fast and dimension-independent. In this case, the right space consists of distributions generated by random feature models, and the $O(1/t)$ rate is provided by Proposition 3.

2. Insensitivity to high dimensional structures: The function representations have small Rademacher complexity, so they are insensitive to the empirical error $Q_* - Q_*^{(n)}$ and the resulting deviation of the training trajectory $Q_t^{(n)} - Q_t$ scales as $O(n^{-1/2})$ instead of $O(n^{-1/d})$. This result is provided by Lemmas 8 and 9 below.

**Lemma 8** *For any distribution $Q_*$ supported on $[-1,1]^d$ and any $\delta \in (0,1)$, with probability $1-\delta$ over the i.i.d. sampling of the empirical distribution $Q_*^{(n)}$, we have*

$$\sup_{\|\mathbf{w}\|_1 \leq 1} \mathbb{E}_{Q_* - Q_*^{(n)}}[\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}})] \leq 4\sqrt{\frac{2\log 2d}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}$$

**Lemma 9** *Let $L$ be a convex Fréchet-differentiable function over a Hilbert space $H$ with Lipschitz constant $l$. Let $h$ be a Fréchet-differentiable function with Lipschitz constant $\epsilon$. Define two gradient flow trajectories $x_t, y_t$:*

$$x_0 = y_0, \ \frac{d}{dt}x_t = -\nabla L(x_t), \ \frac{d}{dt}y_t = -\nabla \widetilde{L}(y_t)$$

*where $\widetilde{L} = L + h$ represents a perturbed function. Then,*

$$L(y_t) - L(x_t) \leq l\epsilon t$$

*for all time $t \geq 0$.*

Numerical examples for the training process and generalization error are provided in Section 4.

### 3.3. Memorization

Despite that the model enjoys good generalization accuracy with early stopping, we show that in the long term the solution $Q_t^{(n)}$ necessarily deteriorates.

**Proposition 10 (Memorization)** *Under the condition of Theorem 6 and Remark 5,*

1. *If the trajectory $Q_t^{(n)}$ has only one weak limit, then $Q_t^{(n)}$ converges weakly to the empirical distribution $Q_*^{(n)}$.*

2. *The true target distribution $Q_*$ can never be a limit point of $Q_t^{(n)}$. The generalization error and the potential function's norm both diverge*

$$\lim_{t\to\infty} KL(Q_*\|Q_t^{(n)}) = \lim_{t\to\infty} \|V_t^{(n)}\|_{\mathcal{H}} = \infty$$

Hence, the model either memorizes the samples or diverges (coming to more than one limit, which are all degenerate), even though it may not manifest within realistic training time.

The proof is based on the following observation.

**Lemma 11** *Let $K \subseteq \mathbb{R}^d$ be a compact set with positive Lebesgue measure, let the base distribution $P$ be uniform over $K$, and let $k$ be a continuous and integrally strictly positive definite kernel on $K$. Given any target distribution $Q' \in \mathcal{P}(K)$ and any initialization $V_0 \in C(K)$, train the potential $V_t$ by*

$$\frac{d}{dt}V_t(\mathbf{x}) = \mathbb{E}_{(Q_t - Q')(\mathbf{x}')}[k(\mathbf{x},\mathbf{x}')]$$

*If $Q_t$ has only one weak limit, then $Q_t$ converges weakly to $Q'$. Else, none of the limit points cover the support of $Q'$.*

A numerical demonstration of memorization is provided in Section 4.

### 3.4. Regularization

Instead of early stopping, one can also consider explicit regularization: With the empirical loss $L^{(n)}$, define the problem

$$\min_{\|V\|\le R} L^{(n)}(V)$$

for some appropriate functional norm $\|\cdot\|$ and adjustable bound $R$. For the special case of random feature models (8), this problem becomes

$$\min_{\|a\|_{L^2(\rho_0)}\le R} L^{(n)}(a) \tag{11}$$

where $L^{(n)}(a)$ denotes $L^{(n)}(V)$ with potential $V$ generated by $a$.

By convexity, $L^{(n)}(a_t^{(n)})$ can always converge to the minimum value as $t\to\infty$ if $a_t^{(n)}$ is trained by gradient flow constrained to the ball $\{\|a\|_{L^2(\rho_0)} \le R\}$. Denote the minimizer of (11) by $a_R^{(n)}$ (which exists by Lemma 19) and denote the corresponding distribution by $Q_R^{(n)}$.

**Proposition 12** *Given the condition of Theorem 6, choose any $R \ge \|V_*\|_{\mathcal{H}}$. With probability $1-\delta$ over the sampling of $Q_*^{(n)}$, the minimizer $a_R^{(n)}$ satisfies*

$$KL(Q_*\|Q_R^{(n)}) \lesssim \frac{\sqrt{\log d} + \sqrt{\log 1/\delta}}{\sqrt{n}} R$$

This result can be straightforwardly extended to the case when $V, V_*$ are implemented as 2-layer networks or deep residual networks, equipped with the norms defined in (E et al., 2019b). The proof only involves the Rademacher complexity, and it is known that these functions' complexity scales as $O(\frac{R}{\sqrt{n}})$ (E et al., 2019b).

## 4. Numerical Experiments

Corollary 7 and Proposition 10 tell us that the training process roughly consists of two phases: the first phase in which a dimension-independent generalization error rate is reached, and a second phase in which the model deteriorates into memorization or divergence. We now examine how these happen in practice.

### 4.1. Empirical sample rate

The key aspect of the generalization estimate of Corollary 7 is that its sample complexity $O(n^{-\alpha})$ ($\alpha \geq 1/4$) is dimension-independent.

To verify dimension-independence, we estimate the exponent $\alpha$ for varying dimension $d$. We adopt the set-up of Theorem 6 and train our model $Q_t^{(n)}$ by SGD on a finite sample set $Q_*^{(n)}$. Specifically, $P$ is uniform over $[-1,1]^d$, the target and trained distributions $Q_*, Q_t^{(n)}$ are generated by the potentials $V_*, V_t^{(n)}$, these potentials are random feature functions (7) with $\rho_0$ being uniform over the $l^1$ sphere $\{\|\mathbf{w}\|_1 + |b| = 1\}$, with parameter functions $a_*, a_t^{(n)}$ and ReLU activation. The samples $Q_*^{(n)}$ are obtained by Projected Langevin Monte Carlo (Bubeck et al., 2018). We approximate $\rho_0$ using $m = 500$ samples (particle discretization) and set $a_* \equiv 50$. We initialize training with $a_0^{(n)} \equiv 0$ and train $a_t^{(n)}$ by scaled gradient descent with learning rate $0.5m$.

The generalization error is measured by $KL(Q_*\|Q_t^{(n)})$. Denote the optimal stopping time by

$$T_o = \text{argmin}_{t>0} KL(Q_*\|Q_t^{(n)})$$

and the corresponding optimal error by $L_o$. The most difficult part of this experiment turned out to be the computation of the KL divergence: Monte-Carlo approximation has led to excessive variance. Therefore we computed by numerical integration on a uniform grid on $[-1,1]^d$. This limits the experiments to low dimensions.

For each $d \leq 5$, we estimate $\alpha$ by linear regression between $\log n$ and $\log L_o$. The sample size $n$ ranges in $\{25, 50, 100, 200\}$, each setting is repeated 20 times with a new sample set $Q_*^{(n)}$. Also, we solve for the dependence of $T_o$ on $n$ by linear regression between $\log n$ and $\log L_o$. Here are the results:

| Dimension $d$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Exponent $-\alpha$ of $L_o$ | $-0.74$ | $-0.71$ | $-0.81$ | $-0.74$ | $-0.87$ |
| Exponent of $T_o$ | 0.30 | 0.29 | 0.27 | 0.26 | 0.31 |

Table 1: Upper: empirically, the exponent $\alpha$ of the sample complexity is dimension-independent. Lower: the optimal stopping time grows with $n$.

Our experiments suggest that the generalization error of the early-stopping solution scales as $n^{-0.8}$ and is dimension-independent, and the optimal early-stopping time is around $n^{0.3}$. This error is much better than the upper bound $O(n^{-1/4})$ given by Corollary 7, indicating that our analysis has much room for improvement.

Shown in Figure 1 is the generalization error $KL(Q_*\|Q_t^{(n)})$ during training, for dimension $d = 5$.

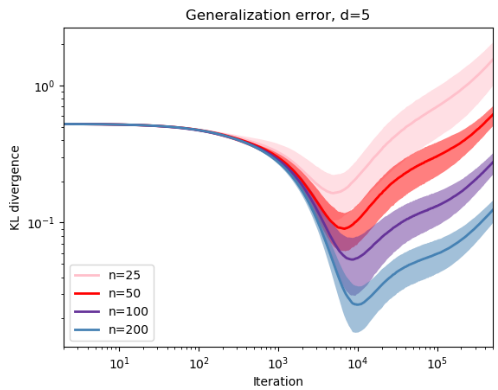


Figure 1: Generalization error curves with log axes. The solid curves are averages over 20 trials, and the shaded regions are $\pm 1$ standard deviations. The results for other $d$ are similar.

All error curves go through a rapid descent, followed by a slower but gradual ascent due to memorization. In fact, the convergence rate prior to the optimal stopping time appears to be exponential. Note that if exponential convergence indeed holds, then the generalization error estimate of Corollary 7 can be improved to $O(n^{-1/2}\log n)$.

### 4.2. Deterioration and memorization

Proposition 10 indicates that as $t \to \infty$ the model either memorizes the sample points or diverges. Our result shows that in practice we obtain memorization.

We adopt the same set-up as in Section 4.1. Since memorization occurs very slowly with SGD, we accelerate training using Adam. Figure 2 shows the result for $d = 1, n = 25$.

We see that there is a time interval during which the trained model closely fits the target distribution, but it eventually concentrates around the samples, and this memorization process does not seem to halt within realistic training time.

Figure 3 suggests that memorization is correlated with the growth of the function norm of the potential.

## 5. Discussion

Let us summarize some of the insights obtained in this paper:

- For distribution-learning models, good generalization can be characterized by dimension-independent *a priori* error estimates for early-stopping solutions. As demonstrated by the proof of Theorem 6, such estimates are enabled by two conditions:

    1. Fast global convergence is guaranteed for learning distributions that can be represented by the model, with an explicit and dimension-independent rate. For our example, this results from the convexity of the model.

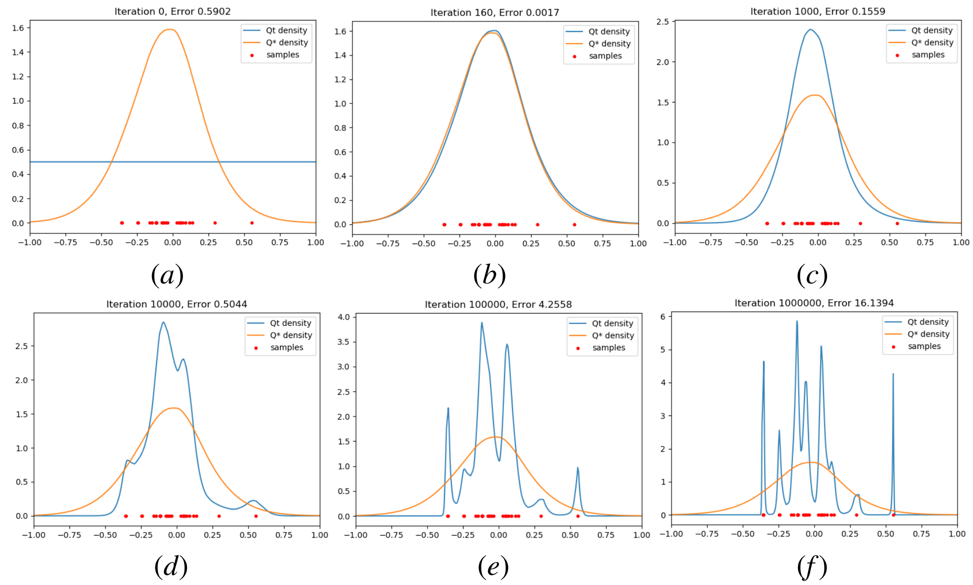


Figure 2: From top left to bottom right: Initialization, optimal stopping time at iteration 160, long time solutions at iterations $10^3, 10^4, 10^5$ and $10^6$. The orange curve is the density of the target distribution $Q_*$, and the blue curves are $Q_t^{(n)}$. The red dots are the samples $Q_*^{(n)}$.

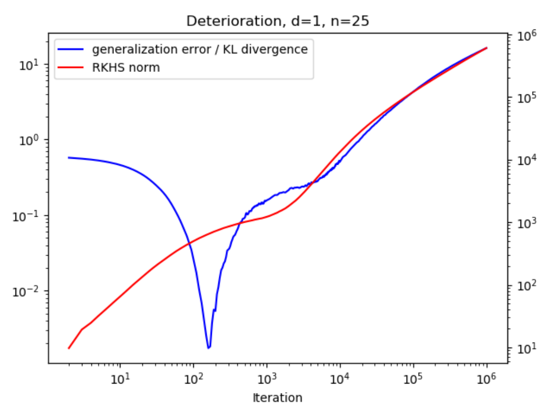


Figure 3: Left: generalization error with Adam optimizer. Right: RKHS norm $\|V_t^{(n)}\|_{\mathcal{H}}$.

2. The model is insensitive to the sampling error $Q_* - Q_*^{(n)}$, so memorization happens very slowly and early-stopping solutions generalize well. For our example, this is enabled by the small Rademacher complexity of the random feature model.

- Memorization seems inevitable for all sufficiently expressive models (Proposition 10), and the generalization error $\widetilde{L}$ will eventually deteriorate to either $n^{-O(1/d)}$ or $\infty$. Thus, instead of the long time limit $t \to \infty$, one needs to consider early-stopping.

  The basic approach, as suggested by Theorem 6, is to choose an appropriate function representation such that, with absolute constants $\alpha_1, \alpha_2 > 0$, there exists an early-stopping interval $[T_{\min}, T_{\max}]$ with $T_{\min} \ll n^{\alpha_1} \ll T_{\max}$ and

$$\sup_{t \in [T_{\min}, T_{\max}]} \widetilde{L}\big(Q_*, Q_t^{(n)}\big) = O(n^{-\alpha_2}) \tag{12}$$

  Then, with a reasonably large sample set (polynomial in precision $\epsilon^{-1}$), the early-stopping interval will become sufficiently wide and hard to miss, and the corresponding generalization error will be satisfactorily small.

- A distribution-learning model can be posed as a calculus of variations problem. Given a training objective $L(Q)$ and distribution representation $Q(f)$, this problem is entirely determined by the function representation or function space $\{\|f\| < \infty\}$. Given a training rule, the choice of the function representation then determines the trainability (Proposition 4) and generalization ability (Theorem 6) of the model.

Future work can be developed from the above insights:

- Generalization error estimates for GANs

  The Rademacher complexity argument should be applicable to GANs to bound the deviation $\|G_t - G_t^{(n)}\|_{L^2(P)}$, where $G_t, G_t^{(n)}$ are the generators trained on $Q_*$ and $Q_*^{(n)}$ respectively. Nevertheless, the difficulty is in the convergence analysis. Unlike bias potential models, the training objective of GAN is non-convex in the generator $G$, and the solutions to $G\#P = Q_*$ are in general not unique.

- Mode collapse

  If we consider mode collapse as a form of bad local minima, then it can benefit from a study of the critical points of GAN, once we pose GAN as a calculus of variations problem. Unlike the bias potential model whose parameter function $V$ ranges in the Hilbert space $\mathcal{H}$, GANs are formulated on the Wasserstein manifold whose tangent space $L^2(Q; \mathbb{R}^d)$ depends significantly on the current position $Q$. In particular, the behavior of gradient flow differs whether $Q$ is absolutely continuous or not, and we expect that successful GAN models can maintain the absolutely continuity of the trajectory $Q_t$.

- New designs

  The design of distribution-learning model can benefit from a mathematical understanding. For instance, consider the early-stopping interval (12), can there be better training rules than gradient flow that reduces $T_{\min}$ or postpones $T_{\max}$ so that early-stopping becomes easier to perform?

## Acknowledgement

The work reported here is supported in part by a gift from Iflytek to Princeton University.

## References

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks, 2017.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). *arXiv preprint arXiv:1703.00573*, 2017.

Yu Bai, Tengyu Ma, and Andrej Risteski. Approximability of discriminators implies diversity in GANs, 2019.

Andrew R Barron and Chyong-Hwa Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, pages 1347–1369, 1991.

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics*, 38(4):1–11, Jul 2019. ISSN 1557-7368. doi: 10.1145/3306346.3323023. URL http://dx.doi.org/10.1145/3306346.3323023.

Luigi Bonati, Yue-Yu Zhang, and Michele Parrinello. Neural networks-based variationally enhanced sampling. In *Proceedings of the National Academy of Sciences*, volume 116, pages 17641–17647, 2019.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 59(4):757–783, 2018.

Stéphane Canu and Alex Smola. Kernel methods and the exponential family. *Neurocomputing*, 69 (7-9):714–720, 2006.

Corneel Casert, Kyle Mills, Tom Vieijra, Jan Ryckebusch, and Isaac Tamblyn. Optical lattice experiments at unobserved conditions and scales through generative adversarial deep learning, 2020.

Tong Che, Yanran Li, Athul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*, 2016.

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

Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in neural information processing systems*, pages 10542–10552, 2019.

Weinan E, Chao Ma, and Lei Wu. A priori estimates for two-layer neural networks. *arXiv preprint arXiv:1810.06397*, 2018.

Weinan E, Chao Ma, and Qingcan Wang. A priori estimates of the population risk for residual networks. *arXiv preprint arXiv:1903.02154*, 1(7), 2019a.

Weinan E, Chao Ma, and Lei Wu. Barron spaces and the compositional function spaces for neural network models. *arXiv preprint arXiv:1906.08039*, 2019b.

Weinan E, Chao Ma, and Lei Wu. Machine learning from a continuous viewpoint. *arXiv preprint arXiv:1912.12777*, 2019c.

Weinan E, Chao Ma, and Lei Wu. On the generalization properties of minimum-norm solutions for over-parameterized neural network models. *arXiv preprint arXiv:1912.06987*, 2019d.

Weinan E, Chao Ma, Stephan Wojtowytsch, and Lei Wu. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't, 2020.

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms, 2017.

S. Feizi, F. Farnia, T. Ginart, and D. Tse. Understanding GANs in the LQG setting: Formulation, generalization and stability. *IEEE Journal on Selected Areas in Information Theory*, 1(1):304–311, 2020. doi: 10.1109/JSAIT.2020.2991375.

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*, pages 2672–2680, 2014.

Ishaan Gulrajani, Colin Raffel, and Luke Metz. Towards GAN benchmarks which require generalization. *arXiv preprint arXiv:2001.03653*, 2020.

Nicholas P. Jewell. Mixtures of exponential distributions. *Ann. Statist.*, 10(2):479–484, 06 1982. doi: 10.1214/aos/1176345789.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, 27(4):887–906, 12 1956. doi: 10.1214/aoms/1177728066.

Qi Lei, Jason D. Lee, Alexandros G. Dimakis, and Constantinos Daskalakis. Sgd learns one-layer networks in wgans, 2020.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Yunwei Mao, Qi He, and Xuanhe Zhao. Designing complex architectured materials with generative adversarial networks. *Science Advances*, 6(17), 2020.

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.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8(1): 143–195, 1999.

Oleksii Prykhodko, Simon Viet Johansson, Panagiotis-Christos Kotsias, Josep Arús-Pous, Esben Jannik Bjerrum, Ola Engkvist, and Hongming Chen. A de novo molecular generation method using latent vector based generative adversarial network. *Journal of Cheminformatics*, 11(74): 1–11, Dec 2019.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 555–561. IEEE, 2008.

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Grant M Rotskoff and Eric Vanden-Eijden. Trainability and accuracy of neural networks: An interacting particle system approach. *stat*, 1050:30, 2019.

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, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Carl-Johann Simon-Gabriel, Alessandro Barp, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies, 2020.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18, 2017.

Yitong Sun, Anna Gilbert, and Ambuj Tewari. On the approximation properties of random relu features. *arXiv preprint arXiv:1810.04374*, 2018.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*, 2020.

Omar Valsson and Michele Parrinello. Variational approach to enhanced sampling and free energy calculations. *Physical review letters*, 113, 2014.

Matt P Wand and M Chris Jones. *Kernel smoothing*. Crc Press, 1994.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. arxiv e-prints, art. *arXiv preprint arXiv:1707.00087*, 2017.

Shanshan Wu, Alexandros G Dimakis, and Sujay Sanghavi. Learning distributions generated by one-layer relu networks. In *Advances in Neural Information Processing Systems 32*, pages 8107–8117. Curran Associates, Inc., 2019.

Lin Yuan, Sergey Kirshner, and Robert Givan. Estimating densities with non-parametric exponential families. *arXiv preprint arXiv:1206.5036*, 2012.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Linfeng Zhang, Weinan E, and Lei Wang. Monge-Ampère flow for generative modeling. *arXiv preprint arXiv:1809.10188*, 2018.

Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in GANs. *arXiv preprint arXiv:1711.02771*, 2017.

## Appendix A. Proofs

### A.1. Proof of the Universal Approximation Property

**Assumption A.1** *For 2-layer neural networks (5), assume that the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is continuous and is not a polynomial.*

*For the random feature model (7), assume that the activation function is continuous, non-polynomial and grows at most linearly at infinity, $\sigma(x) = O(|x|)$. In addition, we assume that the fixed parameter distribution $\rho_0(\mathbf{w}, b)$ has full support over $\mathbb{R}^{d+1}$. (See Theorem 1 of (Sun et al., 2018) for more general conditions.) Alternatively, one can assume that $\sigma$ is ReLU and $\rho_0$ covers all directions, that is, for all $(\mathbf{w}, b) \neq \mathbf{0}$, we have $\lambda(\mathbf{w}, b) \in sprt\rho_0$ for some $\lambda > 0$.*

*By Theorem 3.1 of (Pinkus, 1999), Theorem 1 and Proposition 1 of (Sun et al., 2018), the Barron space $\mathcal{B}$ and RKHS space $\mathcal{H}$ defined by (6) and (8) are dense in the space of continuous functions over any compact subset of $\mathbb{R}^d$.*

**Proof** [Proof of Proposition 1] Denote the set of distributions generated by $\mathcal{V}$ by

$$\mathcal{Q} = \{Q \in \mathcal{P}(K) \mid Q \text{ is given by (2) with } V \in \mathcal{V}\}$$

First, for any $Q_* \in \mathcal{P}_{ac}(K) \cap C(K)$, assume that its density function is strictly positive $Q_*(\mathbf{x}) \geq \epsilon > 0$ over $K$. Define $V_* = \log Q_* \in C(K)$. Let $\{V_m\} \subseteq \mathcal{V}$ be a sequence that approximates $V_*$ in the supremum norm, and let $Q_m$ be the distributions (2) generated by $V_m$. Lemma 13 implies that

$$\lim_{m\to\infty} KL(Q_*\|Q_m) \leq \lim_{m\to\infty} 2\|V_* - V_m\|_{C(K)} = 0$$

For the general case $Q_* \in \mathcal{P}_{ac}(K) \cap C(K)$, define for any $\epsilon \in (0,1)$,

$$Q_*^{\epsilon} = (1-\epsilon)Q_* + \epsilon P$$

For each $m \in \mathbb{N}$, let $Q_m$ be a distribution generated by some $V_m \in \mathcal{V}$ such that $\|\log Q_*^{1/m} - V_m\|_{C(K)} < 1/m$. Then,

$$\begin{aligned}\lim_{m\to\infty} KL(Q_*\|Q_m) &= \lim_{m\to\infty} KL(Q_*\|Q_*^{1/m}) + \mathbb{E}_{Q_*} \log \frac{Q_*^{1/m}(\mathbf{x})}{Q_m(\mathbf{x})} \\ &\leq \lim_{m\to\infty} \frac{1}{m} KL(Q_*\|P) + \|\log Q_*^{1/m} - V_m\|_{C(K)} = 0\end{aligned}$$

where the inequality follows from the convexity of KL. Hence, the set $\mathcal{Q}$ is dense in $\mathcal{P}_{ac}(K) \cap C(K)$ under KL divergence.

Next, consider the total variation norm. Since $\mathcal{P}_{ac}(K) \cap C(K)$ is dense in $\mathcal{P}_{ac}(K)$ under $\|\cdot\|_{TV}$, and since Pinsker's inequality bounds $\|\cdot\|_{TV}$ from above by KL divergence, we conclude that $\mathcal{Q}$ is also dense in $(\mathcal{P}_{ac}(K), \|\cdot\|_{TV})$.

Now consider the $W_1$ metric. $\|\cdot\|_{TV}$ can be seen as an optimal transport cost with cost function $c(\mathbf{x}, \mathbf{x}') = \mathbf{1}_{\mathbf{x} \neq \mathbf{x}'}$, so for any $Q_1, Q_2 \in \mathcal{P}(K)$,

$$W_1(Q_1, Q_2) \leq \text{diam}(K)\ \|Q_1 - Q_2\|_{TV}$$

Since $\mathcal{P}_{ac}(K)$ is dense in $\mathcal{P}(K)$ under the $W_1$ metric, we conclude that $\mathcal{Q}$ is dense in $(\mathcal{P}(K), W_1)$.

Finally, note that for any $p \in [1, \infty)$,

$$W_p \lesssim \text{diam}(K)^{1-1/p}\ W_1^{1/p}$$

So $\mathcal{Q}$ is dense in $(\mathcal{P}(K), W_p)$. ■

### A.2. Estimating the Approximation Error

**Lemma 13** *For any base distribution $P$ and any potential functions $V_1, V_2$,*

$$\left|\log \mathbb{E}_P[e^{-V_1}] - \log \mathbb{E}_P[e^{-V_2}]\right| \leq \|V_1 - V_2\|_{L^\infty(P)}$$

**Proof** Denote $V_{\max}, V_{\min} = \max(V_1, V_2), \min(V_1, V_2)$. Then,

$$
\begin{aligned}
&\big|\log \mathbb{E}_P[e^{-V_1}] - \log \mathbb{E}_P[e^{-V_2}]\big| \\
&\leq \log \mathbb{E}_P[e^{-V_{\min}}] - \log \mathbb{E}_P[e^{-V_{\max}}] \\
&\leq \log \big(\|e^{-V_{\max}}\|_{L^1(P)}\|e^{V_{\max}-V_{\min}}\|_{L^\infty(P)}\big) - \log \|e^{-V_{\max}}\|_{L^1(P)} \\
&= \log \|e^{V_{\max}-V_{\min}}\|_{L^\infty(P)} \\
&= \|V_1 - V_2\|_{L^\infty(P)}
\end{aligned}
$$

■

**Proof** [Proof of Proposition 2] The proof follows the standard argument of Monte-Carlo estimation (Theorem 4 of (E et al., 2019b)). First, consider the case $\|V\|_{\mathcal{B}} < \infty$. For any $\epsilon \in (0, 0.01)$, let $\rho$ be a parameter distribution of $V$ with path norm $\|\rho\|_P < (1+\epsilon)\|V\|_{\mathcal{B}}$. Define the finite neural network

$$
V_m(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^m a_j \sigma(\mathbf{w}_j \cdot \mathbf{x} + b_j) =: \frac{1}{m}\sum_{j=1}^m \phi(\mathbf{x}; \theta_j)
$$

where $\theta_j = (a_j, \mathbf{w}_j, b_j)$ are i.i.d. samples from $\rho$. Denote $\Theta = (\theta_j)_{j=1}^m$.

Let $Q_m$ be the distribution generated by $V_m$. The approximation error is given by

$$
KL(Q\|Q_m) = \mathbb{E}_Q\big[V_m - V\big] + (\log \mathbb{E}_P[e^{-V_m} - \log \mathbb{E}_P[e^{-V}]])
$$

By Lemma 13,

$$
\begin{aligned}
KL(Q\|Q_m) &\leq \|V_m - V\|_{L^\infty(Q)} + \|V_m - V\|_{L^\infty(P)} \\
&\leq 2\|V_m - V\|_{L^\infty(P)}
\end{aligned}
$$

Given that sprt$P \subseteq B_R(0)$, we can bound

$$
\begin{aligned}
\mathbb{E}_\Theta\big[\|V - V_m\|^2_{L^\infty(P)}\big] &\leq \mathbb{E}_\Theta\Big[\sup_{\|\mathbf{x}\|\leq R}\Big(\frac{1}{m}\sum_{j=1}^m \phi(\mathbf{x};\theta_j) - \mathbb{E}_{\theta\sim\rho}[\phi(\mathbf{x};\theta)]\Big)^2\Big] \\
&\leq \mathbb{E}_\Theta\Big[\frac{1}{m^2}\sup_{\|\mathbf{x}\|\leq R}\sum_{j=1}^m \big(\phi(\mathbf{x};\theta_j) - \mathbb{E}_{\theta'}[\phi(\mathbf{x};\theta')]\big)^2\Big] \\
&= \mathbb{E}_{\theta\sim\rho}\Big[\frac{1}{m}\sup_{\|\mathbf{x}\|\leq R}\big(\phi(\mathbf{x};\theta) - \mathbb{E}_{\theta'}[\phi(\mathbf{x};\theta')]\big)^2\Big] \\
&\leq \mathbb{E}_{\theta\sim\rho}\Big[\frac{1}{m}\sup_{\|\mathbf{x}\|\leq R}\phi(\mathbf{x};\theta)^2\Big] \\
&\leq \mathbb{E}_{\theta\sim\rho}\Big[\frac{1}{m}\sup_{\|\mathbf{x}\|\leq R} a^2\|\sigma\|^2_{Lip}(\|\mathbf{w}\|^2 + b^2)(\|\mathbf{x}\|^2 + 1)\Big] \\
&\leq \frac{1}{m}\|\rho\|^2_P(R^2+1)\|\sigma\|^2_{Lip} \\
&\leq \frac{1}{m}(1+\epsilon)^2\|V\|^2_{\mathcal{B}}(R^2+1)\|\sigma\|^2_{Lip}
\end{aligned}
$$

Meanwhile, denote the empirical measure on $\Theta = (\theta_j)$ by $\rho^{(m)} = \frac{1}{m}\sum_{j=1}^m \delta_{\theta_j}$. Then, its expected path norm is bounded by

$$\mathbb{E}_\Theta\big[\|\rho^{(m)}\|_P^2\big] = \frac{1}{m}\sum_{j=1}^m \mathbb{E}_{\theta_j}\big[a_j^2(\|\mathbf{w}_j\|^2 + b_j^2)\big] = \|\rho\|_P^2 \leq (1+\epsilon)^2\|V\|_{\mathcal{B}}^2$$

Define the events

$$\begin{aligned} E_1 &:= \Big\{\Theta \mid \|V - V_m\|_{L^\infty(P)}^2 \leq 3 \cdot \frac{1}{m}\|V\|_{\mathcal{B}}^2 (R^2+1)\|\sigma\|_{Lip}^2\Big\} \\ E_2 &:= \Big\{\Theta \mid \|\rho^{(m)}\|_P^2 \leq 2\|V\|_{\mathcal{B}}^2\Big\} \end{aligned}$$

By Markov's inequality,

$$\begin{aligned} \mathbb{P}(E_1) &= 1 - \mathbb{P}(E_1^C) \geq 1 - \frac{\mathbb{E}\big[\|V - V_m\|_{L^\infty(P)}^2\big]}{\frac{3}{m}\|V\|_{\mathcal{B}}^2(R^2+1)\|\sigma\|_{Lip}^2} \geq 1 - \frac{(1+\epsilon)^2}{3} \\ \mathbb{P}(E_2) &= 1 - \mathbb{P}(E_2^C) \geq 1 - \frac{\mathbb{E}\big[\|\rho^{(m)}\|_P^2\big]}{2\|V\|_{\mathcal{B}}^2} \geq 1 - \frac{(1+\epsilon)^2}{2} \end{aligned}$$

Since $\epsilon \in (0, 0.01)$,

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - 1 \geq \frac{1 - 10\epsilon - 5\epsilon^2}{6} > 0$$

Hence, there exists $\Theta = (\theta_j)_{j=1}^m$ such that

$$\begin{aligned} KL(Q\|Q_m) &\leq 2\|V_m - V\|_{L^\infty(P)} \leq \frac{2\sqrt{3}\|V\|_{\mathcal{B}}}{\sqrt{m}}\|\sigma\|_{Lip}\sqrt{R^2+1} \\ \|V_m\|_{\mathcal{B}} &\leq \|\rho^{(m)}\|_P \leq \sqrt{2}\|V\|_{\mathcal{B}} \end{aligned}$$

The argument for the case $\|V\|_{\mathcal{H}} < \infty$ is the same. ■

### A.3. Proof of Trainability

**Lemma 14** *The objectives $L^+, L^-$ from (3) are convex in $V$.*

**Proof** It suffices to show that $\log \mathbb{E}_P[e^V]$ is convex: Given any two potential functions $V_1, V_2$ and any $t \in (0,1)$, Hölder's inequality implies that

$$\begin{aligned} \log \mathbb{E}_P\big[e^{tV_1+(1-t)V_2}\big] &= \log \mathbb{E}_P\big[(e^{V_1})^t (e^{V_2})^{(1-t)}\big] \\ &\leq \log\Big(\big\|(e^{V_1})^t\big\|_{L^{1/t}(P)}\big\|(e^{V_2})^{(1-t)}\big\|_{L^{1/(1-t)}(P)}\Big) \\ &= \log\big(\mathbb{E}_P[e^{V_1}]^t \mathbb{E}_P[e^{V_2}]^{(1-t)}\big) \\ &= t\log \mathbb{E}_P[e^{V_1}] + (1-t)\log \mathbb{E}_P[e^{V_2}] \end{aligned}$$

■

**Proof** [Proof of Proposition 3] For the target potential function $V_*$, denote its parameter function by $a_* \in L^2(\rho_0)$. Let the objective $L$ be either $L^+$ or $L^-$. The mapping

$$a \mapsto V = \mathbb{E}_{\rho_0}[a(\mathbf{w}, b)\sigma(\mathbf{w} \cdot +b)]$$

is linear while $L$ is convex in $V$ by Lemma 14, so $L$ is convex in $a \in L^2(\rho_0)$ and we simply write the objective as $L(a)$. Define the Lyapunov function

$$E(t) = t\,\big(L(a_t) - L(a_*)\big) + \frac{1}{2}||a_* - a_t||^2_{L^2(d\rho_0)}$$

Then,

$$\begin{aligned}\frac{d}{dt}E(t) &= \big(L(a_t) - L(a_*)\big) + t \cdot \frac{d}{dt}L(a_t) + \big\langle a_t - a_*,\ \frac{d}{dt}a_t\big\rangle_{L^2(\rho_0)} \\ &\leq \big(L(a_t) - L(a_*)\big) - \big\langle a_t - a_*,\ \nabla L(a_t)\big\rangle_{L^2(\rho_0)}\end{aligned}$$

By convexity, for any $a_1, a_2$,

$$L(a_1) + \langle a_2 - a_1,\ \nabla L(a_1)\rangle \leq L(a_2)$$

Hence, $\frac{d}{dt}E \leq 0$. We conclude that $E(t) \leq E(0)$ or equivalently

$$t\,\big(L(a_t) - L(a_*)\big) + \frac{1}{2}||a_* - a_t||^2_{L^2(d\rho_0)} \leq \frac{1}{2}||a^* - a_0||^2_{L^2(d\rho_0)}$$

■

**Assumption A.2** *We make the following assumptions on the activation function $\sigma(\mathbf{w} \cdot \mathbf{x} + b)$, the initialization $\rho_0$ of $\rho_t$, and the base distribution $P$:*

1. *The weights $(\mathbf{w}, b)$ are restricted to the sphere $\mathbb{S}^d \subseteq \mathbb{R}^{d+1}$.*

2. *The activation is universal in the sense that for any distributions $P, Q$,*

$$P = Q \text{ iff } \forall(\mathbf{w}, b) \in \mathbb{S}^d,\ \mathbb{E}_{P-Q}\big[\sigma(\mathbf{w} \cdot \mathbf{x} + b)\big] = 0$$

3. *$\sigma$ is continuously differentiable with a Lipschitz derivative $\sigma'$. (For instance, $\sigma$ might be sigmoid or mollified ReLU.)*

4. *The initialization $\rho_0 = \rho_0(a, \mathbf{w}, b) \in \mathcal{P}(\mathbb{R} \times \mathbb{S}^d)$ has full support over $\mathbb{S}^d$. Specifically, the support of $\rho_0$ contains a submanifold that separates the two components, $(\infty, -\overline{a}) \times \mathbb{S}^d$ and $(\overline{a}, \infty) \times \mathbb{S}^d$, for some $\overline{a}$.*

5. *$P$ is compactly-supported.*

**Proof** [Proof of Proposition 4] The proof follows the arguments of (Chizat and Bach, 2018; Rotskoff and Vanden-Eijden, 2019). For convenience, denote $(\mathbf{x}, 1)$ by $\tilde{\mathbf{x}}$ and $(\mathbf{w}, b)$ by $\mathbf{w}$, so the activation is simply $\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}})$. Denote the training objective by $L$ ($L = L^+$ or $L = L^-$). From a particle perspective, the flow (10) can be written as

$$\begin{aligned} \dot{a}_t &= -\mathbb{E}_{\Delta_t}\big[\sigma(\mathbf{w}_t \cdot \tilde{\mathbf{x}})\big] \\ \dot{\mathbf{w}}_t &= -a_t\, \mathbb{E}_{\Delta_t}\big[\sigma'(\mathbf{w}_t \cdot \tilde{\mathbf{x}})\, \tilde{\mathbf{x}}\big] \end{aligned} \tag{13}$$

where $\Delta_t = P_* - P$ if $L = L^+$ and $\Delta_t = Q_* - Q_t^-$ if $L = L^-$.

Since the velocity field (13) is locally Lipschitz over $\mathbb{R} \times \mathbb{S}^d$, the induced flow is a family of locally Lipschitz diffeomorphisms, and thus preserve the submanifold given by Assumption A.2. Denote by $\hat{\rho}_t$ and $\hat{\rho}_\infty$ the projections of $\rho_t, \rho_\infty$ onto $\mathbb{S}^d$. It follows that $\hat{\rho}_t$ has full support over $\mathbb{S}^d$ for all time $t < \infty$.

Since $\rho_\infty$ is a stationary point of $L$, the velocity field (13) vanishes at $\rho_\infty$ almost everywhere. In particular, for all $\mathbf{w}$ in the support of $\hat{\rho}_\infty$,

$$g(\mathbf{w}) := \mathbb{E}_{\Delta_\infty}\big[\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}})\big] = 0$$

We show that this condition holds for all $\mathbf{w} \in \mathbb{S}^d$. Denote $S = \mathbb{S}^d - \text{sprt}\hat{\rho}_\infty$. Assume to the contrary that $g(\mathbf{w})$ does not vanish on $S$. Let $\mathbf{w}_* \in S$ be a maximizer of $|g(\mathbf{w})|$. Without loss of generality, let $g(\mathbf{w}_*) > 0$; the same reasoning applies to $g(\mathbf{w}_*) < 0$.

Since $\rho_t \to \rho_\infty$ in $W_1$, the bias potential $V_t$ converges to $V^*$ uniformly over the compact support of $P$. Since all $\Delta_t$ are supported on sprt$P$, the velocity field (13) converges locally uniformly to

$$\begin{bmatrix} -\mathbb{E}_{\Delta_\infty}[\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}})] \\ -a\, \mathbb{E}_{\Delta_\infty}[\sigma'(\mathbf{w} \cdot \tilde{\mathbf{x}})\, \tilde{\mathbf{x}}] \end{bmatrix} = \begin{bmatrix} -g(\mathbf{w}) \\ -a g'(\mathbf{w}) \end{bmatrix}$$

For $t$ sufficiently large, we can study the flow with this approximate field. Let $(a, \mathbf{w})$ be any point with $\mathbf{w}$ sufficiently close to $\mathbf{w}_*$, consider a trajectory $(a_t, \mathbf{w}_t)$ initialized from $a_{t_0} = a, \mathbf{w}_{t_0} = \mathbf{w}$ with a large $t_0$. If $a < 0$, then $a_t$ becomes increasingly negative, while $\mathbf{w}_t$ follows a gradient ascent on $g$ and converges to $\mathbf{w}_*$ (or any maximizer nearby). Else, $a \geq 0$, but if $\mathbf{w}$ is sufficiently close to $\mathbf{w}_*$, then $\dot{\mathbf{w}}_t = O(g'(\mathbf{w}))$ is very small (since $g'(\mathbf{w}_*) = 0$ and $g'$ is Lipschitz in $\mathbf{w}$), so $\mathbf{w}_t$ will stay around $\mathbf{w}_*$ and $g(\mathbf{w}_t)$ remains positive. Then, $a_t$ eventually becomes negative, and $\mathbf{w}_t$ converges to $\mathbf{w}_*$.

Since $\hat{\rho}_t$ has positive mass in any neighborhood of $\mathbf{w}_*$ at time $t_0$, this mass will remain in $S$ as $t \to \infty$. This is a contradiction since $S$ is disjoint from sprt$\hat{\rho}^*$. It follows that $g(\mathbf{w})$ vanishes on all of $\mathbb{S}^d$. Then for any $\mathbf{w} \in \mathbb{S}^d$,

$$\mathbb{E}_{\Delta_\infty}\big[\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}})\big] = 0$$

By Assumption A.2, we conclude that $\Delta_\infty = 0$, or equivalently $Q_\infty = Q_*$ and $V_\infty = V_*$ (up to an additive constant). ■

### A.4. Proof of Generalization Ability

**Proof** [Proof of Lemma 8] Theorem 6 of (E et al., 2019b) implies that given any $n$ points with $l^\infty$ norm $\leq 1$, the Rademacher complexity of the class $\{\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}}), \|\mathbf{w}\|_1 \leq 1\}$ is bounded by

$$Rad_n \leq 2\sqrt{\frac{2 \log 2d}{n}}$$

Since $|\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})| \leq 1$ for all $\|\mathbf{w}\|_1 \leq 1$, $\|\mathbf{x}\|_\infty \leq 1$, we can apply Theorem 26.5 of (Shalev-Shwartz and Ben-David, 2014) to conclude that

$$\forall \|\mathbf{w}\| \leq 1,\ \mathbb{E}_{Q_* - Q_*^{(n)}}[\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})] \leq 2Rad_n + \sqrt{\frac{2\log(2/\delta)}{n}}$$

with probability $1-\delta$ over the sampling of $Q_*^{(n)}$. ∎

**Proof** [Proof of Lemma 9] Denote the inner product and norm of $H$ by $\langle x, y\rangle$ and $\|x\|$. Then,

$$\frac{d}{dt}\|y_t - x_t\| \leq -\big\langle \frac{y_t - x_t}{\|y_t - x_t\|},\ \nabla L(y_t) - \nabla L(x_t) + \nabla h(y_t)\big\rangle$$

Since $L$ is convex, $(y-x)\cdot(\nabla L(y) - \nabla L(x)) \geq 0$ for any $x, y \in H$. Therefore,

$$\begin{aligned}\frac{d}{dt}\|y_t - x_t\| &\leq -\langle \frac{y_t - x_t}{\|y_t - x_t\|}, \nabla h(y_t)\rangle \\ &\leq \|\nabla h(y_t)\| \leq \epsilon\end{aligned}$$

so that $\|y_t - x_t\| \leq \epsilon t$. By Lipschitz continuity, $L(y_t) - L(x_t) \leq l\epsilon t$. ∎

**Proof** [Proof of Theorem 6] For any time $T$, the testing error can be decomposed into

$$\begin{aligned}KL\big(Q_* \| Q_T^{(n)}\big) &= L(V_T^{(n)}) - L(V_*) \\ &= \big(L(V_T^{(n)}) - L(V_T)\big) + \big(L(V_T) - L(V_*)\big)\end{aligned}$$

The second term is bounded by Proposition 3, while the first term can be bounded by Lemma 9. The Hilbert space $H$ in Lemma 9 corresponds to the parameter functions $L^2(\rho_0)$ for the random feature model, the convex objective corresponds to the objective $L$ over $a \in L^2(\rho_0)$,

$$L(a) = \mathbb{E}_{Q_*}[V] + \log \mathbb{E}_P[e^{-V}], \quad V(\mathbf{x}) = \mathbb{E}_{\rho_0(\mathbf{w})}[a(\mathbf{w})\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})]$$

and the perturbation term $h$ corresponds to $L^{(n)} - L$,

$$L^{(n)}(a) - L(a) = \mathbb{E}_{Q_*^{(n)} - Q_*}[V]$$

The remaining task is to estimate the constants $l$ and $\epsilon$.

First, we have $l \leq 2$. For any $a \in L^2(\rho_0)$, let $Q$ be the modeled distribution,

$$\begin{aligned}\|\nabla L(a)\|_{L^2(\rho_0)} &= \|\mathbb{E}_{Q_* - Q}[\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})]\|_{L^2(d\rho_0(\mathbf{w}))} \\ &\leq \sup_{\|\mathbf{w}\|_1 \leq 1} |\mathbb{E}_{Q_* - Q}[\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})]| \\ &\leq \sup_{\|\mathbf{w}\|_1 \leq 1} |\mathbb{E}_{Q_*}[\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})]| + \sup_{\|\mathbf{w}\|_1 \leq 1} |\mathbb{E}_{Q}[\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})]| \\ &\leq 2\end{aligned}$$

where in the last step, since all distributions are supported on $[-1,1]^d$, $\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}}) \leq \|\mathbf{w}\|_1 \|\tilde{\mathbf{x}}\|_\infty \leq 1$.

Next, the estimate of $\epsilon$ has been provided by Lemma 8, because for any $a \in L^2(\rho_0)$,

$$\begin{aligned}
\|\nabla h(a)\|_{L^2(\rho_0)} &= \|\nabla L^{(n)}(a) - \nabla L(a)\|_{L^2} \\
&= \|\mathbb{E}_{Q_* - Q_*^{(n)}}[\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}})]\|_{L^2} \\
&\leq \|\mathbb{E}_{Q_* - Q_*^{(n)}}[\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}})]\|_{L^\infty(\rho_0)} \\
&\leq \sup_{\|\mathbf{w}\|_1 \leq 1} |\mathbb{E}_{Q_* - Q_*^{(n)}}[\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}})]|
\end{aligned}$$

■

### A.5. Proof of Memorization

To prove Proposition 10 and Lemma 11, we begin with a few useful lemmas.

Let $\mathcal{M}(K)$ be the space of finite signed measures on $K$. We say that a kernel $k$ is integrally strictly positive definite if

$$\forall m \in \mathcal{M}(K),\ \mathbb{E}_{m(\mathbf{x})}\mathbb{E}_{m(\mathbf{x}')}[k(\mathbf{x}, \mathbf{x}')] \to m = 0$$

Equip $\mathcal{M}(K)$ with the inner product

$$\forall m_1, m_2 \in \mathcal{M}(K),\ \langle m_1, m_2 \rangle_k := \mathbb{E}_{m_1(\mathbf{x})}\mathbb{E}_{m_2(\mathbf{x}')}[k(\mathbf{x}, \mathbf{x}')]$$

from which we define the MMD (maximum mean discrepancy) distance $\|\cdot\|_k$

$$\|m_1 - m_2\|_k^2 = \langle m_1 - m_2, m_1 - m_2 \rangle_k$$

Let $\mathcal{H}_k$ be the RKHS generated by $k$ with inner product $\langle , \rangle_{\mathcal{H}_k}$. Then the MMD inner product is the RKHS inner product on the mean embeddings $f_i = \mathbb{E}_{m_i(\mathbf{x})}[k(\mathbf{x}, \cdot)]$,

$$\begin{aligned}
\langle m_1, m_2 \rangle_k &= \langle f_1, f_2 \rangle_{\mathcal{H}_k} \\
\|m_1 - m_2\|_k &= \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{m_1 - m_2}[f]
\end{aligned}$$

**Lemma 15** *When restricted to the subset $\mathcal{P}(K)$, the MMD distance $\|\cdot\|_k$ induces the weak topology and thus $(\mathcal{P}(K), \|\cdot\|_k)$ is compact.*

**Proof** By Lemma 2.1 of (Simon-Gabriel et al., 2020), the MMD distance metrizes the weak topology of $\mathcal{P}(K)$, which is compact by Prokhorov's theorem. ■

As $\mathcal{P}(K)$ is a convex subset of $\mathcal{M}(K)$, we can define the tangent cone at each point $Q \in \mathcal{P}(K)$ by

$$T_Q\mathcal{P}(K) = \{\lambda\Delta \mid \lambda \geq 0,\ \Delta = \Delta^+ - \Delta^-,\ \Delta^\pm \in \mathcal{P}(K),\ \Delta^- \ll Q\}$$

and equip it with the MMD norm, $\|\Delta\|_k^2 = \mathbb{E}_{\Delta^2}[k]$.

Given the gradient flow $V_t$ defined in Lemma 11, the distribution $Q_t$ evolves by

$$\frac{d}{dt}Q_t(\mathbf{x}) = \big(v(\mathbf{x}; Q_t) - \mathbb{E}_{Q_t(\mathbf{x}')}[v(\mathbf{x}'; Q_t)]\big)Q_t(\mathbf{x})$$

$$v(\mathbf{x}; Q) := \mathbb{E}_{(Q'-Q)(\mathbf{x}')}[k(\mathbf{x}, \mathbf{x}')]$$

We can extend this flow to a dynamical system on $\mathcal{P}(K)$ in positive time $t \geq 0$, defined by

$$\begin{aligned} \frac{d}{dt}Q_t &= \overline{v}(Q_t)Q_t \\ \overline{v}(Q) &= v(\cdot\,; Q) - \mathbb{E}_{Q(\mathbf{x}')}[v(\mathbf{x}'; Q)] \end{aligned} \tag{14}$$

Each $\overline{v}(Q)Q$ is a tangent vector in $T_Q\mathcal{P}(K)$.

Note that we can rewrite $v$ and $\overline{v}$ in terms of the RKHS norm: Let $f, f'$ be the mean embeddings of $Q, Q'$,

$$\begin{aligned} v(\mathbf{x}; Q) &= \langle k(\mathbf{x}, \cdot),\ f' - f\rangle_{\mathcal{H}_k} \\ \overline{v}(\mathbf{x}; Q) &= \langle k(\mathbf{x}, \cdot) - f,\ f' - f\rangle_{\mathcal{H}_k} \end{aligned}$$

It follows that $v$ and $\overline{v}$ are uniformly continuous over the compact space $K \times (\mathcal{P}(K), \|\cdot\|_k)$.

**Lemma 16** *Given any initialization $Q_0 \in \mathcal{P}(K)$, there exists a unique solution $Q_t$, $t \geq 0$ to the dynamics (14).*

**Proof** The integral form of (14) can be written as

$$\forall t \geq 0,\ Q_t = Q_0 + \int_0^t \overline{v}(Q_s)Q_s ds \tag{15}$$

where we adopt the Bochner integral on $(\mathcal{M}(K), \langle,\rangle_k)$. In the spirit of the classical Picard-Lindelöf theorem, we consider the vector space $C([0,T], \mathcal{M}(K))$ equipped with sup-norm

$$|||\phi||| = \sup_{t\in[0,T]} \|\phi(t)\|_k$$

On the complete subspace $C([0,T], \mathcal{P}(K))$, define the operator $\phi \mapsto F(\phi)$ by

$$F(\phi)_t = \phi_0 + \int_0^t \overline{v}(\phi_s)\phi_s ds$$

Define the sequence $\phi^0 \equiv Q_0$ and $\phi^{n+1} = F(\phi^n)$.

Note that the tangent field (14) is Lipschitz

$$\forall Q_1, Q_2, \quad \|\overline{v}(Q_1)Q_1 - \overline{v}(Q_2)Q_2\|_k \leq c\|Q_1 - Q_2\|_k$$

with $c \leq 4(\|k\|^2_{C(K\times K)} + \|k\|_{C(K\times K)})$. Then, with $T \leq 1/2c$,

$$\begin{aligned} |||\phi^{n+1} - \phi^n||| &\leq \sup_{t\in[0,T]} \Big\| \int_0^t \overline{v}(\phi^n_s)\phi^n_s - \overline{v}(\phi^{n-1}_s)\phi^{n-1}_s ds \Big\|_k \\ &\leq \int_0^T \|\overline{v}(\phi^n_t)\phi^n_t - \overline{v}(\phi^{n-1}_t)\phi^{n-1}_t\|_k dt \\ &\leq cT \sup_{t\in[0,T]} \|\phi^n_t - \phi^{n-1}_t\|_k \end{aligned}$$

$$\leq \frac{1}{2}|||\phi^n - \phi^{n-1}|||$$

By the completeness of $(C([0,T],\mathcal{P}(K)), |||\cdot|||)$ and Banach fixed point theorem, the sequence $\phi^n$ converges to a unique solution $\phi$ of (15) on $[0,T]$. Then, we can extend this solution iteratively to $[T,2T],[2T,3T],\dots$ and obtain a unique solution on $[0,\infty)$. ■

Denote the set of fixed points of (14) by

$$\mathcal{P}_o = \{Q \in \mathcal{P}(K) \mid \overline{v}(Q)Q = 0\}$$

Also, define the set of distributions that have larger supports than the target distribution $Q'$

$$\mathcal{P}_* = \{Q \in \mathcal{P}(K) \mid \text{sprt}Q' \subseteq \text{sprt}Q\}$$

**Lemma 17** *We have the following inclusion*

$$\mathcal{P}_o \subseteq \{Q'\} \cup (\mathcal{P}(K) - \mathcal{P}_*)$$

*Given any initialization $Q_0 \in \mathcal{P}(K)$, let $Q_t$, $t \geq 0$ be the trajectory defined by Lemma 16 and let $\mathcal{Q}$ be the set of limit points in MMD metric*

$$\mathcal{Q} = \bigcap_{T\to\infty} \overline{\{Q_t, t \geq T\}}^{\|\cdot\|_k}$$

*then $\mathcal{Q} \subseteq \mathcal{P}_o$.*

**Proof** For any fixed point $Q \in \mathcal{P}_o$, we have $\overline{v}(\mathbf{x};Q) = 0$ for $Q$-almost every $\mathbf{x}$. By continuity, we have

$$\forall \mathbf{x} \in \text{sprt}Q, \quad v(\mathbf{x};Q) = \mathbb{E}_{Q(\mathbf{x}')}[v(\mathbf{x}';Q)] \tag{16}$$

If we further suppose that $Q \in \mathcal{P}_*$, then this equality holds for $Q'$-almost all $\mathbf{x}$, so

$$\begin{aligned} 0 &= \mathbb{E}_{(Q-Q')(\mathbf{x})}[v(\mathbf{x};Q)] \\ &= \mathbb{E}_{(Q-Q')^2(\mathbf{x},\mathbf{x}')}[k(\mathbf{x},\mathbf{x}')] \\ &= \|Q - Q'\|_k^2 \end{aligned}$$

Since $k$ is integrally strictly positive definite, we have $Q = Q'$. It follows that

$$\mathcal{P}_o \cap \mathcal{P}_* = \{Q'\}$$

or equivalently $\mathcal{P}_o \subseteq \{Q'\} \cup (\mathcal{P}(K) - \mathcal{P}_*)$.

Meanwhile, the MMD distance $\|Q_t - Q'\|_k^2$ is decreasing along any trajectory $Q_t$ of (14):

$$\begin{aligned} \frac{d}{dt}\frac{1}{2}\|Q_t - Q'\|_k^2 &= \mathbb{E}_{Q_t(\mathbf{x})}\mathbb{E}_{(Q_t-Q')(\mathbf{x}')}\left[k(\mathbf{x},\mathbf{x}')\,\overline{v}(\mathbf{x};Q_t)\right] \\ &= -\mathbb{E}_{Q_t(\mathbf{x})}\left[\overline{v}(\mathbf{x};Q_t)^2\right] \\ &\leq 0 \end{aligned} \tag{17}$$

Define the extended sublevel sets for every $c > 0$,

$$\mathcal{P}_c := \{Q \in \mathcal{P}(K) \mid \|Q - Q'\|_k \leq c \text{ or } Q \in \mathcal{P}_o\}$$

By Lemma 15, the space $(\mathcal{P}(K), \|\cdot\|_k)$ is compact, so the set of limit points $\mathcal{Q}$ of the trajectory $Q_t$ is nonempty. The inequality (17) is strict if $Q_t \notin \mathcal{P}_o$, so these limit points all belong to

$$\bigcap_{c \to 0^+} \mathcal{P}_c = \mathcal{P}_o$$

■

**Lemma 18** *Given any initialization $Q_0 \in \mathcal{P}_*$, if the limit point set $\mathcal{Q}$ contains only one point $Q_\infty$, then $Q_\infty \in \mathcal{P}_*$ and thus $Q_\infty = Q'$. Else, $\mathcal{Q}$ is contained in $\mathcal{P}(K) - \mathcal{P}_*$.*

**Proof** For any open subset $A$ that intersects sprt$Q_0$, we have $Q_0(A) > 0$. Also

$$\begin{aligned}\frac{d}{dt}Q_t(A) &= \mathbb{E}_{Q_t}[\mathbf{1}_A(\mathbf{x})\overline{v}(\mathbf{x}; Q_t)] \\ &\geq -Q_t(A)\|\overline{v}(Q_t)\|_{L^\infty(Q_t)} \\ &\geq -4\|k\|_{C(K \times K)}Q_t(A)\end{aligned}$$

So $Q_t(A)$ remains positive for all finite $t$. It follows that sprt$Q_0 \subseteq$ sprt$Q_t$ and $Q_t \in \mathcal{P}_*$ for all $t$.

First, consider the case $\mathcal{Q} = \{Q_\infty\}$. Assume for contradiction that $Q_\infty = \tilde{Q}$ for some $\tilde{Q} \in \mathcal{P}_o - \{Q'\} \subseteq \mathcal{P}(K) - \mathcal{P}_*$. Equation (16) implies that

$$\mathbb{E}_{\tilde{Q}}[\overline{v}(\mathbf{x}; \tilde{Q})] = 0$$

and thus

$$\begin{aligned}\mathbb{E}_{Q'}[\overline{v}(\mathbf{x}; \tilde{Q})] &= \mathbb{E}_{Q' - \tilde{Q}}[\overline{v}(\mathbf{x}; \tilde{Q})] \\ &= \mathbb{E}_{Q' - \tilde{Q}}[v(\mathbf{x}; \tilde{Q})] \\ &= \|Q' - \tilde{Q}\|_k^2 \\ &> 0\end{aligned}$$

In particular, there exists some measureable subset $S_o \subseteq$ sprt$Q'$ and some $\delta > 0$ such that

$$\forall \mathbf{x} \in S_o,\ \overline{v}(\mathbf{x}; \tilde{Q}) > 2\delta$$

By continuity, there exists some open subset $S$ ($S_o \subseteq S$) such that its closure $\overline{S}$ satisfies

$$\forall \mathbf{x} \in \overline{S},\ \overline{v}(\mathbf{x}; \tilde{Q}) \geq \delta$$

Meanwhile, since $S$ intersects sprt$Q' \subseteq$ sprt$Q_t$, we have $Q_t(S) > 0$ for all $t$. Whereas (16) implies that $\overline{S}$ is disjoint from sprt$\tilde{Q}$.

Since $\overline{v}$ is continuous over $(\mathbf{x}, Q) \in K \times (\mathcal{P}(K), \|\cdot\|_k)$ and $\overline{S}$ is compact, there exists some neighborhood $B_r(\tilde{Q}) = \{Q \in \mathcal{P}(K) \mid \|Q - \tilde{Q}\|_k < r\}$ such that

$$\forall Q \in B_r(\tilde{Q}),\ \forall \mathbf{x} \in \overline{S},\ \overline{v}(\mathbf{x}; \tilde{Q}) \geq 0$$

Since the trajectory $Q_t$ converges in the MMD distance $\|\cdot\|_k$ to $\tilde{Q}$, there exists some time $t_0$ such that for all $t \geq t_0$, $Q_t \in B_r(\tilde{Q})$. It follows that

$$\frac{d}{dt} Q_t(\overline{S}) = \mathbb{E}_{Q_t}[\mathbf{1}_{\overline{S}}(\mathbf{x}) \overline{v}(\mathbf{x}; Q_t)] \geq 0$$

so that $Q_t(\overline{S}) \geq Q_{t_0}(\overline{S})$ for all $t \geq t_0$. Yet, Lemma 15 implies that $Q_t$ converges weakly to $\tilde{Q}$, so that

$$0 = \tilde{Q}(\overline{S}) \geq \limsup_{t \to \infty} Q_t(\overline{S})$$

A contradiction. We conclude that the limit point $Q_\infty$ does not belong to $\mathcal{P}_o - \{Q'\}$. By Lemma 17, we must have $Q_\infty = Q'$.

Next, consider the case when $\mathcal{Q}$ has more than one point. Inequality (17) implies that the MMD distance $L(Q) = \|Q - Q_*^{(n)}\|_k^2$ is monotonously decreasing along the flow $Q_t$. Suppose that $Q' \in \mathcal{Q}$, then $\lim_{t\to\infty} L(Q_t) = 0$ and thus $\mathcal{Q} = \{Q'\}$, a contradiction. Hence, $\mathcal{Q} \subseteq \mathcal{P}_o - \{Q'\} \subseteq \mathcal{P}(K) - \mathcal{P}_*$.
■

**Proof** [Proof of Lemma 11] Since $V_0 \in C(K)$, the initialization $Q_0$ has full support over $K$ and thus $Q_0 \in \mathcal{P}_*$. If $Q_t$ converges weakly to some limit $Q_\infty$, Lemma 15 implies that $Q_t$ also converges in MMD metric to $Q_\infty$. Then, Lemma 18 implies that the limit $Q_\infty$ must be $Q'$.

If there are more than one limit, then Lemma 18 implies that all limit points belong to $\mathcal{P}(K) - \mathcal{P}_*$ and thus do not cover the full support of $Q'$. ■

**Proof** [Proof of Proposition 10] We simply set $Q' = Q_*^{(n)}$. Note that since $a_t^{(n)}$ is trained by

$$\frac{d}{dt} a_t^{(n)}(\mathbf{w}) = \mathbb{E}_{(Q_t^{(n)} - Q_*^{(n)})(\mathbf{x})}[\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}})]$$

the training dynamics for the potential $V_t^{(n)}$ is the same as in Lemma 11

$$\frac{d}{dt} V_t^{(n)}(\mathbf{x}) = \mathbb{E}_{(Q_t^{(n)} - Q_*^{(n)})(\mathbf{x}')}[k(\mathbf{x}, \mathbf{x}')]$$

with kernel $k$ defined by

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{\rho_0(\mathbf{w})}[\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}}) \sigma(\mathbf{w} \cdot \tilde{\mathbf{x}}')]$$

It is straightforward to check that $k$ is integrally strictly positive definite: For any $m \in \mathcal{M}(K)$, if

$$0 = \|m\|_k^2 = \mathbb{E}_{\rho_0(\mathbf{w})}\big(\mathbb{E}_{m(\mathbf{x})}[\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}}))]\big)^2$$

then for $\rho_0$-almost all $\mathbf{w}$, $\mathbb{E}_{m(\mathbf{x})}[\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}})] = 0$. It follows that for all random feature models $f$ from (7), we have $\mathbb{E}_{m(\mathbf{x})}[f(\mathbf{x})] = 0$. Assuming Remark 5, the random feature models are dense in $C(K)$ by Proposition 1, so this equality holds for all $f \in C(K)$. Hence, $m = 0$ and $k$ is integrally strictly positive definite.

Hence, Lemma 11 implies that if $Q_t^{(n)}$ has one limit point, then $Q_t^{(n)}$ converges weakly to $Q_*^{(n)}$. Else, no limit point can cover the support of $Q_*^{(n)}$ and thus do not have full support over $K$. Since the true target distribution $Q_*$ is generated by a continuous potential $V_*$, it has full support and thus does not belong to $\mathcal{Q}$ and $KL(Q_* \| Q) = \infty$ for all $Q \in \mathcal{Q}$. Similarly, we must have

$$\liminf_{t \to \infty} \|V_t^{(n)}\|_{\mathcal{H}} = \infty$$

otherwise some subsequence of $Q_t^{(n)}$ would converge to a limit with full support. ■

### A.6. Proof for the Regularized Model

**Lemma 19** *For any $R \geq 0$, there exists a minimizer of (11).*

**Proof** Since the closed ball $B_R = \{\|a\|_{L^2(\rho_0)} \leq R\}$ is weakly compact in $L^2(\rho_0)$, it suffices to show that the mapping

$$L^{(n)}(a) = \mathbb{E}_{\rho_0(\mathbf{w})}\big[a(\mathbf{w})\mathbb{E}_{Q_*^{(n)}(\mathbf{x})}[\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})]\big] + \log \mathbb{E}_{P(\mathbf{x})}\big[e^{-\mathbb{E}_{\rho_0(\mathbf{w})}[a(\mathbf{w})\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})]}\big]$$

is weakly continuous over $B_R$ (e.g. show that the term $\mathbb{E}_P\big[e^{-\mathbb{E}_{\rho_0(\mathbf{w})}[a(\mathbf{w})\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})]}\big]$ can be expressed as the uniform limit of a sequence of weakly continuous functions over $B_R$). Then, every minimizing sequence of $L^{(n)}$ in $B_R$ converges weakly to a minimizer of (11). ■

**Proof** [Proof of Proposition 12]

For any $a \in L^2(\rho_0)$,

$$\begin{aligned}|L(a) - L^{(n)}(a)| &\leq \mathbb{E}_{\rho_0(\mathbf{w})}\big[|\mathbb{E}_{Q_* - Q_*^{(n)}}[a(\mathbf{w})\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})]|\big] \\ &\leq \|a\|_{L^2(\rho_0)} \cdot \sup_{\|\mathbf{w}\|_1 \leq 1} \mathbb{E}_{Q_* - Q_*^{(n)}}[\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})]\end{aligned}$$

Thus, Lemma 8 implies that with probability $1-\delta$ over the sampling of $Q_*^{(n)}$,

$$|L(a) - L^{(n)}(a)| \leq \|a\|_{L^2(\rho_0)} \cdot \Big(4\sqrt{\frac{2\log 2d}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}\Big) \tag{18}$$

It follows that

$$\begin{aligned}L(a_R^{(n)}) &\leq L^{(n)}(a_R^{(n)}) + \frac{4\sqrt{2\log 2d} + \sqrt{2\log(2/\delta)}}{\sqrt{n}} R \\ &\leq L^{(n)}(a_*) + \frac{4\sqrt{2\log 2d} + \sqrt{2\log(2/\delta)}}{\sqrt{n}} R \\ &\leq L(a_*) + \frac{4\sqrt{2\log 2d} + \sqrt{2\log(2/\delta)}}{\sqrt{n}} (R + \|a_*\|_{L^2(\rho_*)})\end{aligned}$$

where the first and third inequalities follow from (18) and the second inequality follows from the fact that $a_* \in \{\|a\|_{L^2(\rho_0)} \leq R\}$.

Hence,

$$KL(Q_* \| Q_R^{(n)}) = L(a_R^{(n)}) - L(a_*) \leq 2R \cdot \frac{4\sqrt{2\log 2d} + \sqrt{2\log(2/\delta)}}{\sqrt{n}}$$

■