A MEAN-FIELD THEORY FOR KERNEL ALIGNMENT WITH RANDOM FEATURES IN GENERATIVE ADVERSARIAL NETWORKS

MASOUD BADIEI KHUZANI†, LIYUE SHEN‡, SHAHIN SHAHRAMPOUR§, LEI XING‡

Abstract. We propose a novel supervised learning method to optimize the kernel in maximum mean discrepancy generative adversarial networks (MMD GANs). Specifically, we characterize a distributionally robust optimization problem to compute a good distribution for the random feature model of Rahimi and Recht to approximate a good kernel function. Due to the fact that the distributional optimization is infinite dimensional, we consider a Monte-Carlo sample average approximation (SAA) to obtain a more tractable finite dimensional optimization problem. We subsequently leverage a particle stochastic gradient descent (SGD) method to solve finite dimensional optimization problems. Based on a mean-field analysis, we then prove that the empirical distribution of the interactive particles system at each iteration of the SGD follows the path of the gradient descent flow on the Wasserstein manifold. We also establish the non-asymptotic consistency of the finite sample estimator. Our empirical evaluation on synthetic data-set as well as MNIST and CIFAR-10 benchmark data-sets indicates that our proposed MMD GAN model with kernel learning indeed attains higher inception scores well as Fréchet inception distances and generates better images compared to the generative moment matching network (GMMN) and MMD GAN with untrained kernels.

1. Introduction

A fundamental and long-standing problem in unsupervised learning systems is to capture the underlying distribution of data. While deep generative models such as Boltzmann machines [41] and auto-encoding variational Bayes [20] accomplish this task to some extent, they are inadequate for many intractable probabilistic computations that arise in maximum likelihood estimation. Moreover, in many machine learning tasks such as caption generation [53], the main objective is to obtain new samples rather than to accurately estimate the underlying data distribution. Generative adversarial networks (GANs) [12] provides a framework to draw samples from a high dimensional data distribution without estimating the distribution.

The structure of a GAN consists of a deep feedforward network to generate new samples from a base distribution (e.g. Gaussian distribution), and a discriminator network to accept or reject the generated samples. However, training GAN requires finding a Nash equilibrium of a non-convex minimax game with continuous, high-dimensional parameters. Consequently, it is highly unstable and prone to miss modes [42, 5]. To obtain more stable generative models, the generative moment matching networks (GMMNs) [26] are proposed, wherein instead of training a discriminator network, a non-parametric statistical hypothesis

†Department of Management Science, Stanford University, CA 94043, USA
‡Department of Electrical Engineering, Stanford University, CA 94043, USA
§Department of Radiation Oncology, Stanford University, CA 94043, USA
§Department of Industrial Engineering, Texas A&M University, TX 77843, USA
(♦Co-first Authors with Equal Contributions)
test is performed to accept or reject the generated samples via the computation of the kernel maximum mean discrepancy [13]. While leveraging a statistical test simplifies the loss function for training GMMNs, in practice, the diversity of generated samples by GMMNs is highly sensitive to the choice of the kernel. Thus, to improve the sampling performance, the kernel function also needs to be jointly optimized with the generator. Along this direction, the MMD GAN model [24] is proposed in which an embedding function is optimized in conjunction with a fixed user-defined kernel (e.g. RBF Gaussian kernel). However, there are no theoretical guarantees that the user-defined kernel is the optimal kernel for embedded features.

1.1. Main contributions. To address the kernel model selection problem in MMD GAN [24], in this paper we put forth a novel framework to learn a good kernel function from training data. Our kernel learning approach is based on a distributional optimization problem to learn a good distribution for the random feature model of Rahimi and Recht [36, 37]. The main contributions of this paper is summarized as follows:

- Based on the notion of the kernel-target alignment, we characterize a distributional optimization problem to learn a good distribution for Rahimi and Recht’s random feature model of a kernel function [36, 37].
- To obtain a tractable optimization problem, we leverage the sample average approximation (SAA) method to transform the infinite dimensional distributional optimization problem to a finite dimensional optimization problem.
- We establish the consistency of the finite sample average approximation. In particular, we show that as the number of samples in SAA tends to infinity, the optimal value of the finite sample estimates tend to their population values.
- We propose a particle stochastic gradient descent method to solve the finite dimensional optimization problem associated with the Monte-Carlo sampling method. Using a mean-field analysis, we then show that the interactive particle system with SGD dynamics follows the gradient decent path on the Wasserstein manifold to minimize the distributional optimization problem. In this sense, we establish the consistency of the proposed particle SGD method for solving the distributional optimization problem.

1.2. Paper outline. The rest of this paper is organized as follows:

- Review of the MMD GAN. In Section 2, we review the MMD GAN model described in [24], and characterize the kernel optimization problem underlying this generative model.
- Kernel learning Approach and the proposed MMD GAN architecture. In Section 3, we first describe a connection between the MMD loss in MMD GAN model, and the notion of the kernel alignment for kernel learning. Using random features, we then formulate an optimization problem in terms of the distribution of the random features. We then solve this optimization problem using the particle SGD algorithm.
- Theoretical results. In Section 4, we state our main theoretical results. Due to variety of works on kernel optimization, we postpone the detailed discussion of the related literature after presenting the main results.
- Empirical evaluations. In Section 5, we provide empirical evaluation of our proposed GAN method on synthetic data-set, as well as on MNIST [23] and CIFAR-10 [21] benchmark datasets. Compared to GMNN and MMD GAN, our proposed network attain higher inception score on these data-sets, without requiring to tune the bandwidth of Gaussian RBF kernel.
Assume we are given data \({\{v_i\}}_{i=1}^{n}\) that are sampled from an unknown distribution \(P_V\) with the support \(\mathcal{V}\). In many unsupervised tasks, we wish to attain new samples from the distribution \(P_V\) without directly estimating it. Generative Adversarial Network (GAN) [12] provides such a framework. In vanilla GAN, a deep network \(G(\cdot; W)\) parameterized by \(W \in W\) is trained as a generator to transform the samples \(Z \sim P_Z, Z \in Z\) from a user-defined distribution \(P_Z\) (e.g. Gaussian distribution) into a new sample \(G(Z; W) \sim P_W\), such that the distributions \(P_W\) and \(P_V\) are close under some specified metric. In addition, a discriminator network \(D(\cdot; \delta)\) parameterized by \(\delta \in \Delta\) is also trained to reject or accept the generated samples as a realization of the data distribution. The training of the generator and discriminator networks is then accomplished via solving a minimax optimization problem as below

\[
\min_{W \in W} \max_{\delta \in \Delta} \mathbb{E}_{P_V}[D(X; \delta)] + \mathbb{E}_{P_Z}[\log(1 - D(G(Z; W); \delta))].
\]

In the high dimensional settings, the generator trained via the minimax program of Eq. (2.1) can potentially collapse to a single mode of distribution where it always emits the same point \([5]\). To overcome this shortcoming, other adversarial generative models are developed in the literature, which propose to modify or replace the discriminator network by a statistical two-sample test based on the notion of the maximum mean discrepancy (MMD). Below, we formalize the notion of MMD:

**Definition 2.1. (Maximum Mean Discrepancy)** Let \((\mathcal{X}, d)\) be a metric space, \(\mathcal{F}\) be a class of functions \(f : \mathcal{X} \rightarrow \mathbb{R}\), and \(P, Q \in \mathcal{B}(\mathcal{X})\) be two probability measures from the set of all Borel probability measures \(\mathcal{B}(\mathcal{X})\) on \(\mathcal{X}\). The maximum mean discrepancy (MMD) between the distributions \(P\) and \(Q\) with respect to the function class \(\mathcal{F}\) is defined below

\[
\text{MMD}_{\mathcal{F}}[P, Q] = \sup_{f \in \mathcal{F}} \int_X f(x)(P - Q)(dx).
\]

Adapting different function classes \(\mathcal{F}\) in Eq. (2.2) of Definition (2.1) yield different adversarial models such as Wasserstein GANs (WGAN) [1], \(f\)-GANs [31], GMMN [24], and MMD GAN [26]. In the latter two cases, the function class \(\mathcal{F}\) corresponds to a reproducing kernel Hilbert space (RKHS) of functions with a kernel \(K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}\), denoted by \((\mathcal{H}_K, K)\). For RKHS function class, the squared MMD in Eq. (2.2) between the distributions \(P = P_V\) and \(Q = P_W\) has the following expression\(^1\)

\[
\text{MMD}_K[P_V, P_W] = \sup_{f \in \mathcal{H}_K} \left\| \int_{\mathcal{X}} K(\cdot, x)(dP_V(x) - dP_W(x)) \right\|_{\mathcal{H}_K}^2
\]

\[
= \mathbb{E}_{P_V^{\otimes 2}}[K(V; V')] + \mathbb{E}_{P_W^{\otimes 2}}[K(W, W')] - 2\mathbb{E}_{P_{V,W}}[K(V; W)],
\]

where \(\mathcal{X} = \mathcal{V} \cup \mathcal{W}\).

Instead of training the generator via solving the minimax optimization in Eq. (2.1), the MMD GAN model of [26] proposes to optimize the discrepancy between two distributions via optimization of an embedding function \(\iota : \mathbb{R}^d \mapsto \mathbb{R}^p, p \leq d\), i.e.,

\[
\min_{W \in W} \max_{\iota \in \Theta} \text{MMD}_{K_{\Theta}}[P_V, P_W],
\]

\(^1\)We note that the notation \(\text{MMD}_{K}[P_V, P_W]\) is conventionally used to denote the square root of the expression on the right hand side of Eq. (2.3). However, to simplify the arguments in subsequent results and their proofs, we adapt the notation of Eq. (2.3).
where $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a user-defined fixed kernel. In [26], the proposal for the kernel $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a mixture of the Gaussians,

$$
(2.5) \quad k \circ \iota(x, y) = k(\iota(x), \iota(y)) = \sum_{i=1}^m \exp \left( \frac{\|\iota(x) - \iota(y)\|^2}{\sigma_i^2} \right),
$$

where the bandwidth parameters $\sigma_1, \ldots, \sigma_m > 0$ are manually selected. However, there are two challenges associated with the optimization of the embedding map with a pre-fixed Gaussian mixture kernel:

(i) The Gaussian mixture kernel is not expected to be optimal. Even if the Gaussian mixture kernel yields a good performance on a given data-set, there are no theoretical guarantees to suggest that it performs well across all data-sets that are encountered in practice.

(ii) When the Gaussian mixture kernel model is admissible, the choice of the Gaussian bandwidths $\sigma_1, \ldots, \sigma_m$ poses a statistical model selection problem. Although such model selection issues can be addressed using the cross-validation or jackknife, in practice such methods slow down the training process due to repeatedly re-fitting the model.

To address the aforementioned issues, in this paper we propose to optimize the MMD loss with respect to the underlying kernel

$$
(2.6) \quad \min_{W \in W} \max_{K \in K} \text{MMD}_K[P_V, P_W],
$$

over a suitable kernel class $K$. Notice that the optimization of the embedding function in Eq. (2.4) is a special case of the general kernel learning problem formulated in Eq. (2.6) when the kernel class is selected as $K^\text{def} = \{K : K(x, y) = k(\iota(x), \iota(y)), \iota \in \mathcal{Q}\}$.

3. Proposed approach: kernel learning with random features for MMD GANs

In this section, we first expound our kernel learning approach. Then, we describe a novel MMD GAN model based on the proposed kernel learning approach.

3.1. Robust distributional optimization for kernel learning. To address the kernel model selection issue in MMD GAN [24], we consider a kernel optimization scheme with random features [36, 37]. Let $\varphi : \mathbb{R}^D \times \mathbb{R}^D \to [-1, 1]$ denotes the explicit feature maps and $\mu \in \mathcal{M}(\mathbb{R}^D)$ denotes a probability measure from the space of probability measures $\mathcal{M}(\mathbb{R}^D)$ on $\mathbb{R}^D$. The kernel function is characterized via the explicit feature maps using the following integral equation

$$
(3.1) \quad K_\mu(x, y) = \mathbb{E}_\mu[\varphi(x; \xi)\varphi(y; \xi)] = \int_\Xi \varphi(x; \xi)\varphi(y; \xi)\mu(d\xi).
$$

Let $\text{MMD}_\mu[P_V, P_W] \text{def} = \text{MMD}_K[P_V, P_W]$. Then, the inner kernel optimization problem in Eq. (2.6) can alternatively be formulated in terms of the distribution of random features as follows, i.e,

$$
(3.2) \quad \min_{W \in W} \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W].
$$
Here, $\mathcal{P}$ is the set of probability distributions corresponding to a kernel class $\mathcal{K}$. In the sequel, we consider $\mathcal{P}$ to be the distribution ball of radius $R$ as below

$$
\mathcal{P} \overset{\text{def}}{=} \mathcal{B}_R(\mu_0) = \{ \mu \in \mathcal{M}(\mathbb{R}^D) : d(\mu, \mu_0) \leq R \},
$$

where $\mu_0$ is a user-defined base distribution, and $d(\cdot, \cdot) : \mathcal{M}(\mathbb{R}^D) \times \mathcal{M}(\mathbb{R}^D) \to \mathbb{R}$ is a distance on the measure space $\mathcal{M}(\mathbb{R}^D)$.

The kernel MMD loss function in Eq. (3.2) is defined with respect to the unknown distributions of the data-set $P_V$ and the model $P_W$. Therefore, we construct an unbiased estimator for the MMD loss function in Eq. (3.2) based on the training samples. To describe the estimator, sample the labels from a uniform distribution $y_1, \cdots, y_n \sim \text{i.i.d.} \text{Uniform}\{-1, +1\}$, where we assume that the number of positive and negative labels are balanced. In particular, consider the set of positive labels $\mathcal{I} = \{ i \in \{1, \cdots, n\} : y_i = +1 \}$, and negative labels $\mathcal{J} = \{1, 2, \cdots, n\} / \mathcal{I}$, where their cardinality is $|\mathcal{I}| = |\mathcal{J}| = \frac{n}{2}$. We consider the following assignment of labels:

- **Positive class labels**: If $y_i = +1$, sample the corresponding feature map from data-distribution $x_i = v_i \sim P_V$.
- **Negative class labels**: If $y_i = -1$, sample from the corresponding feature map from the generated distribution $x_i = \mathcal{G}(Z_i, W) \sim P_W$, $Z_i \sim P_Z$.

By this construction, the joint distribution of features and labels $P_{Y, X}$ has the marginals $P_{X|Y = +1} = P_V$, and $P_{X|Y = -1} = P_W$. Moreover, the following statistic, known as the kernel alignment in the literature (see, e.g., [43, 6]), is an unbiased estimator of the MMD loss in Eq. (3.2),

$$
\min_{W \in \mathcal{W}} \sup_{\mu \in \mathcal{P}} \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} y_i y_j K_\mu(x_i, x_j).
$$

See Appendix A.1 for the related proof. The kernel alignment in Eq. (3.4) can also be viewed through the lens of the risk minimization

$$
\min_{W \in \mathcal{W}} \inf_{\mu \in \mathcal{P}} \frac{1}{n(n-1)\alpha} \sum_{1 \leq i < j \leq n} (\alpha y_i y_j - K_\mu(x_i, x_j))^2
$$

(3.5b)

Here, $\alpha > 0$ is a scaling factor that determines the separation between feature vectors, and $K_\mu \overset{\text{def}}{=} \alpha yy^T$ is the ideal kernel that provides the maximal separation between the feature vectors over the training data-set, i.e., $K_\mu(x_i, x_j) = \alpha$ when features have identical labels $y_i = y_j$, and $K_\mu(x_i, x_j) = -\alpha$ otherwise. Upon expansion of the risk function in Eq. (3.5), it can be easily shown that it reduces to the kernel alignment in Eq. (3.4) when $\alpha \to +\infty$. Intuitively, the risk minimization in Eq. (3.5) gives a feature space in which pairwise distances are similar to those in the output space $\mathcal{Y} = \{-1, +1\}$.

### 3.2. SAA for distributional optimization

The distributional optimization problem in Eq. (3.2) is infinite dimensional, and thus cannot be solved directly. To obtain a tractable optimization problem, instead of optimizing with respect to the distribution $\mu$ of random features, we optimize the i.i.d. samples (particles) $\xi_1, \cdots, \xi^N \sim \text{i.i.d.} \mu$ generated from the
distribution. The empirical distribution of these particles is accordingly defined as follows
\[
\hat{\mu}^N(\xi) \overset{\text{def}}{=} \frac{1}{N} \sum_{k=1}^{N} \delta(\xi - \xi^k),
\]
where \(\delta(\cdot)\) is the Dirac’s delta function concentrated at zero. In practice, the optimization problem in Eq. (3.5) is solved via the Monte-Carlo sample average approximation of the objective function,
\[
\min_{\mathcal{W}} \min_{\hat{\mu}^N \in \mathcal{P}_N} \text{MMD}^\alpha_{\mathcal{P}^N} \left[ P_{\mathcal{V}}, P_{\mathcal{W}} \right] = \frac{8}{n(n-1)\alpha} \sum_{1 \leq i < j \leq n} \left( \alpha y_i y_j - \mathbb{E}_{\hat{\mu}^N} [\varphi(x_i; \xi^k)\varphi(x_j; \xi^k)] \right)^2
\]
(3.7a)
\[
= \frac{8}{n(n-1)\alpha} \sum_{1 \leq i < j \leq n} \left( \alpha y_i y_j - \frac{1}{N} \sum_{k=1}^{N} \varphi(x_i; \xi^k)\varphi(x_j; \xi^k) \right)^2,
\]
where \(\mathcal{P}_N \overset{\text{def}}{=} \mathbb{B}_\mathcal{V}^N(\hat{\mu}_0^N) = \{\hat{\mu}^N \in \mathcal{M}(\mathbb{R}^D) : W_p(\hat{\mu}^N, \hat{\mu}_0^N) \leq R\},\) and \(\hat{\mu}_0^N\) is the empirical measure associated with the samples \(\xi^1, \ldots, \xi^N\). The empirical objective function in Eq. (3.7) can be optimized with respect to the samples \(\xi^1, \ldots, \xi^N\) using the particle stochastic gradient descent. For the optimization problem in Eq. (3.7), the (projected) stochastic gradient descent (SGD) takes the following recursive form,\(^2\)
\[
\xi_{m+1} = \xi_m - \eta \left( y_m \bar{y}_m - \frac{1}{\alpha N} \sum_{k=1}^{N} \varphi(x_m; \xi^k_m)\varphi(\bar{x}_m; \xi^k_m) \right) \nabla \xi \left( \varphi(x_m; \xi^k_m)\varphi(\bar{x}_m; \xi^k_m) \right),
\]
for \(k = 1, 2, \ldots, N\), where \((y_m, x_m), (\bar{y}_m, \bar{x}_m) \sim \text{i.i.d.} P_{x,y}\) and \(\eta \in \mathbb{R}^>0\) denotes the learning rate of the algorithm, and the initial particles are \(\xi^1_0, \ldots, \xi^N_0 \sim \text{i.i.d.} \mu_0\). At each iteration of the SGD dynamic in Eq. (3.8), a feasible solution for the inner optimization of the empirical risk function in Eq. (3.7) is generated via the empirical measure
\[
\hat{\mu}_m^N(\xi) = \frac{1}{N} \sum_{k=1}^{N} \delta(\xi - \xi^k_m).
\]
Indeed, we prove in Section 4 that for an appropriate choice of the learning rate \(\eta > 0\), the empirical measure in Eq. (3.9) remains inside the distribution ball \(\hat{\mu}_m^N \in \mathcal{P}_N\) for all \(m \in [0, NT] \cap \mathbb{N}\), and is thus a feasible solution for the empirical risk minimization (3.7) (see Corollary 4.2.1 in Section 4).

3.3. Proposed MMD GAN with kernel learning. In Algorithm 1, we describe the proposed method MMD GAN model with the kernel learning approach described earlier. Algorithm 1 has an inner loop for the kernel training and an outer loop for training the generator, where we employ RMSprop [46]. Our proposed MMD GAN model is distinguished from MMD GAN of [24] in that we learn a good kernel function in Eq. (3.10) of the inner loop instead of optimizing the embedding function that is implemented by an auto-encoder. However, we mention that our kernel learning approach is compatible with the auto-encoder implementation of [24] for dimensionality reduction of features (and particles). In the case of including an auto-encoder, the inner loop in Algorithm 1 must be modified to add an additional step for training the auto-encoder. However, to convey the main ideas more clearly, the training step of the auto-encoder is omitted from Algorithm 1.

\(^2\)To avoid clutter in our subsequent analysis, the normalization factor \(\frac{16}{n(n-1)}\) of the gradient is omitted by modifying the step-size \(\eta\).
Algorithm 1 MMD GAN with a supervised kernel learning Method (Monte-Carlo Approach)

**Inputs:** The learning rates $\bar{\eta}, \eta > 0$, the number of iterations of discriminator per generator update $T \in \mathbb{N}$, the batch-size $n$, the number of random features $N \in \mathbb{N}$. Regularization parameter $\alpha > 0$.

**while** $\omega$ has not converged **do**

**for** $t = 1, 2, \cdots, T$ **do**

Sample the labels $y, \tilde{y} \sim \text{i.i.d. Uniform}\{-1, 1\}$.

Sample the features $x|y = +1 \sim P_X$, and $x|y = -1 \sim P_X$. Similarly, $\tilde{x}|\tilde{y} = +1 \sim P_X$, and $\tilde{x}|\tilde{y} = -1 \sim P_X$.

For all $k = 1, 2, \cdots, N$, update the particles,

$$
(3.10) \quad \xi^k \leftarrow \xi^k - \frac{\eta}{N} \left( a g y - \frac{1}{N} \sum_{k=1}^{N} \varphi(x; \xi^k) \varphi(\tilde{x}; \xi^k) \right) \nabla \xi \left( \varphi(x; \xi^k) \varphi(\tilde{x}; \xi^k) \right).
$$

**end for**

Sample a balanced minibatch of labels $\{y_i\}_{i=1}^n \sim \text{i.i.d. Uniform}\{-1, +1\}$.

Sample the minibatch $\{x_i\}_{i=1}^n$ such that $x_i|y_i = +1 \sim P_X$, and $x_i|y_i = -1 \sim P_X$ for all $i = 1, 2, \cdots, n$.

**Update the generator**

$$
(3.11a) \quad g_\omega \leftarrow \nabla_\omega \hat{D}_{\mu N}^\alpha [P_V, P_X], \quad \hat{\mu}^N = \frac{1}{N} \sum_{k=1}^{N} \delta(\xi - \xi^k).
$$

$$
(3.11b) \quad w \leftarrow w - \bar{\eta} \text{RMSprop}(g_\omega, \omega).
$$

**end while**

3.4. Related works. The mean-field description of SGD dynamics has been studied in several prior works for different information processing tasks. Wang et al. [50] consider the problem of online learning for the principal component analysis (PCA), and analyze the scaling limits of different online learning algorithms based on the notion of finite exchangeability. In their seminal papers, Montanari and co-authors [30, 17, 29] consider the scaling limits of SGD for training a two-layer neural network, and characterize the related McKean-Vlasov PDE for the limiting distribution of the empirical measure associated with the weights of the input layer. They also establish the uniqueness and existence of the solution for the PDE using the connection between McKean-Vlasov type PDEs and the gradient flows on the Wasserstein manifolds established by Otto [32], and Jordan, Kinderlehrer, and Otto [18]. Similar mean-field type results for two-layer neural networks are also studied recently in [39, 44].

The present application of the mean-field theory to the kernel optimization in MMD GANs is partly motivated by the work of Sinha and Duchi [43], which studied a distributional optimization for kernel learning with random features in the context of classification problems. Therein, the authors have proposed a robust optimization framework for the importance sampling of the random features. In contrast to the work of [43] that assign a weight (importance) to each sample and optimizes the weights, in this paper we directly optimize the weights. Kernel learning for MMD GAN has been studied in several prior works [25, 51, 14]. Nevertheless, those methods are either based on heuristics, or are difficult to characterize theoretically. Our work is also related to the unpublished work of Wang, et al. [49], which proposes a solvable model of GAN and analyzes the scaling limits. However, our GAN model is significantly different from [49] and is based on the notion of the kernel MMD.

Our work is also closely related to the recent work of Li, et al. [25] which proposes an implicit kernel learning method based on the following kernel definition

$$
(3.12) \quad K_h(\iota(x), \iota(y)) = \mathbb{E}_{\xi \sim \mu_0} \left[ e^{ih(\iota(x) - \iota(y))} \right],
$$

where $\mu_0$ is a user defined base distribution, and $h \in \mathcal{H}$ is a functions that transforms the base distribution $\mu_0$ into a distribution $\mu$ that provides a better kernel. Therefore, the work
of Li, et al [25] implicitly optimizes the distribution of random features through a function. In contrast, the proposed distributional optimization framework in this paper optimizes the distribution of random feature explicitly, via optimizing their empirical measures. Perhaps more importantly from a practical perspective is the fact that our kernel learning approach does not require the user-defined function class \( \mathcal{H} \). Moreover, our particle SGD method in (3.8) obviates tuning hyper-parameters related to the implicit kernel learning method such as the gradient penalty factor and the variance constraint factor (denoted by \( \lambda_{GP} \) and \( \lambda_{h} \), respectively, in Algorithm 1 of [25]).

4. Main Results: Consistency and Mean-Field Analysis

In this section, we provide theoretical guarantees for the consistency of various approximations we made to optimize the population MMD loss function in (3.2). We defer the proofs of the following theoretical results to Section A of Appendix.

4.1. Assumptions. Before we delve into technical results, we state the main assumptions underlying them:

(A.1) The feature space \( \mathcal{X} = \mathcal{V} \cup \mathcal{W} \subset \mathbb{R}^{d} \) is compact with a finite diameter \( \text{diam}(\mathcal{X}) < \infty \), where \( \mathcal{V} = \text{support}(P_{V}) \) and \( \mathcal{W} = \text{support}(P_{W}) \) are the supports of the distributions \( P_{V} \) and \( P_{W} \) respectively.

(A.2) The feature maps are bounded and Lipschitz almost everywhere (a.e.) \( \xi \in \mathbb{R}^{p} \). In particular, \( \sup_{x \in \mathcal{X}} |\varphi(x; \xi)| < L_{0}, \sup_{x \in \mathcal{X}} \|\nabla_{x} \varphi(x; \xi)\|_{2} \leq L_{1}, \text{ and } \sup_{\xi \in \mathbb{R}^{D}} \|\nabla_{x} \varphi(x; \xi)\| < L_{2}. \) Let \( L \overset{\text{def}}{=} \max\{L_{0}, L_{1}, L_{2}\} < +\infty \).

(A.3) Let \( \hat{\mu}_{0}^{N}(\xi) \overset{\text{def}}{=} \frac{1}{N} \sum_{k=1}^{N} \delta(\xi - \xi_{0}^{k}) \) denotes the empirical measure for the initial particles \( \xi_{0}^{1}, \cdots, \xi_{0}^{N} \). We assume that \( \hat{\mu}_{0}^{N}(\xi) \) converges (weakly) to a deterministic measure \( \mu_{0} \in \mathcal{M}(\mathbb{R}^{p}) \). Furthermore, we assume the limiting measure \( \mu_{0} \) is absolutely continuous w.r.t. Lebesgue measure and has a compact support \( \text{support}(\mu_{0}) = \Xi \subset \mathbb{R}^{p} \).

4.2. Consistency of finite-sample estimate. In this part, we prove that the solution to finite sample optimization problem in (3.7) approaches its population optimum in (3.2) as the number of data points as well as the number of random feature samples tends to infinity.

To establish the proof, we consider the \( d_{W}^{p}(\cdot, \cdot) \) is the \( p \)-Wasserstein (a.k.a. Kantorovich-Rubinstein metric) distance defined as below

\[
(4.1) \quad d_{W}^{p}(\mu_{1}, \mu_{2}) \overset{\text{def}}{=} \left( \inf_{\pi \in \Pi(\mu_{1}, \mu_{2})} \int_{\mathbb{R}^{D} \times \mathbb{R}^{D}} \|\xi_{1} - \xi_{2}\|_{p}^{2} d\pi(\xi_{1}, \xi_{2}) \right)^{\frac{1}{2}}.
\]

where the infimum is taken with respect to all couplings \( \pi \) of the measures \( \mu, \mu_{0} \in \mathcal{M}(\mathbb{R}^{D}) \), and \( \Pi(\mu, \mu_{0}) \) is the set of all such couplings.

**Theorem 4.1.** (Non-asymptotic Consistency of Finite-Sample Estimator) Suppose conditions (A.1)-(A.3) of Appendix A are satisfied. Consider the distribution balls \( \mathcal{P} \) and \( \mathcal{P}_{N} \) that are defined with respect to the \( 2 \)-Wasserstein distance. Furthermore, consider the optimal MMD values of the population optimization and its finite sample estimate

\[
(4.2a) \quad (W_{*}, \mu_{*}) \overset{\text{def}}{=} \arg \min_{W \in \mathcal{W}} \arg \sup_{\mu \in \mathcal{P}} \text{MMD}_{\mu}[P_{V}, P_{W}],
\]

\[
(4.2b) \quad (W_{*}^{N}, \hat{\mu}_{*}^{N}) \overset{\text{def}}{=} \arg \min_{W \in \mathcal{W}} \arg \inf_{\hat{\mu}^{N} \in \mathcal{P}_{N}} \text{MMD}_{\hat{\mu}^{N}}[P_{V}, P_{W}],
\]

where \( \text{MMD}_{\mu} \) and \( \text{MMD}_{\hat{\mu}^{N}} \) denote the MMDs with respect to the measures \( \mu \) and \( \hat{\mu}^{N} \) respectively.
respective. Then, with the probability of (at least) $1 - 3q$ over the training data samples $(\{(x_i, y_i)\}_{i=1}^n)$ and the random feature samples $(\{\xi^k\}_{k=1}^N)$, the following non-asymptotic bound holds

\begin{equation}
|MMD_{\mu_*}[P_{\mathcal{X}}, P_{\mathcal{W}_*}] - MMD_{\bar{\mu}^N_*}[P_{\mathcal{X}}, P_{\bar{W}^N_*}]| \\
\leq \sqrt{\frac{L^2 (d+2)}{N}} \ln \frac{1}{\delta} \left( \frac{2^d N \text{diam}^2(\mathcal{X})}{\theta} \right) + 2 \max \left\{ \frac{c_1 L^2}{n} \ln \frac{1}{\delta}, \frac{c_2 RL^4}{n^2} \ln \left( \frac{4e^{1/4}k}{\theta} \right) \right\} + \frac{8L^2}{\alpha},
\end{equation}

where $c_1 = 3^4 \times 2^4$, and $c_2 = 9 \times 2^{11}$.

The proof of Theorem 4.1 is presented in Appendix A.1.

Notice that there are three key parameters involved in the upper bound of Theorem 4.1. Namely, the number of training samples $n$, the number of random feature samples $N$, and the regularization parameter $\alpha$. The upper bound in Eq. (4.3) thus shows that when $n, N, \alpha \to +\infty$, the solution obtained from solving the empirical risk minimization in Eq. (3.5) yield the same MMD population value as that of the distributional optimization in Eq. (3.2).

### 4.3. Consistency of particle SGD algorithm

The consistency result of Theorem 4.1 is concerned with the MMD value of the optimal empirical measure $\hat{\mu}^N_* (\xi) = \frac{1}{N} \sum_{k=1}^N \delta(\xi - \xi^k)$ of the empirical risk minimization (3.7). In practice, the particle SGD is executed for a few iterations and its values are returned as an estimate for $(\xi^*_1, \cdots, \xi^*_N)$. Consequently, it is desirable to establish a consistency type result for the particle SGD estimates $(\xi^*_m, \cdots, \xi^*_N)$ at the $m$-th iteration, where the notion of consistency will be made precise shortly. To prove such a consistency result, we define the scaled empirical measure as follows

\begin{equation}
\mu^N_t = \hat{\mu}^N_{[Nt]} = \frac{1}{N} \sum_{k=1}^N \delta(\xi - \xi_{[Nt]}), \quad 0 \leq t \leq T.
\end{equation}

At any time $t$, the scaled empirical measure $\mu^N_t$ is a random element, and thus $(\mu^N_t)_{0 \leq t \leq T}$ is a measured-valued stochastic process. Therefore, we characterize the evolution of its Lebesgue density $p^N_t(\xi) \equiv \mu^N_t(d\xi)/d\xi$ in the following theorem:

**Theorem 4.2. (McKean-Vlasov Mean-Field PDE)** Suppose conditions (A.1)-(A.3) in Section A are satisfied. Further, suppose that the Radon-Nikodym derivative $q_0(\xi) = \mu_0(d\xi)/d\xi$ exists. Then, there exists a unique solution $(p^*_t(\xi))_{0 \leq t \leq T}$ to the following nonlinear partial differential equation

\begin{equation}
\begin{cases}
\frac{\partial p_t(\xi)}{\partial t} = -\frac{q_0}{2} \int_{x \times y} \left( \int_{\mathcal{X}^2} \varphi(x, \xi) \varphi(y, \xi) p_t(\xi) d\xi - q_0 \gamma \right) \nabla_{\xi} p_t(\xi) \nabla_{\xi} \varphi(x; \xi) \varphi(y; \xi) dP_{x \times y}, \\
p_0(\xi) = q_0(\xi).
\end{cases}
\end{equation}

Moreover, the measure-valued process $(\{\mu^N_t\}_{0 \leq t \leq T})_{N \in \mathbb{N}}$ defined in Eq. (4.4) converges (weakly) to the unique solution $\mu^*_t(\xi) = p^*_t(\xi) d\xi$ as the number of particles tends to infinity $N \to \infty$. 

\footnote{The notion of the weak convergence of a sequence of empirical measures is formally defined in Appendix.}
The proof of Theorem 4.2 is presented in Appendix A.1.

As a by product of the mean-field analysis of Theorem 4.2, we can prove that the empirical measure \( \hat{\mu}_m^N \) of the particles in SGD dynamic (3.8) remains inside the feasible distribution ball \( \mathcal{P}_N \) defined with respect to the Radon distance

\[
(4.6) \quad d_R(\mu_1, \mu_2) = \sup \left\{ \int_{\mathbb{R}^D} f(\xi)d(\mu_1 - \mu_2)(\xi) : f \in \mathcal{F}_c \right\},
\]

where \( \mathcal{F}_c \) is the class of continuous functions \( f : \mathbb{R}^D \to [-1, 1] \subset \mathbb{R} \).

**Corollary 4.2.1.** Consider the learning rate \( \eta = O(\frac{R}{T + T\sqrt{NT\log(2/\delta)}}) \) for the SGD in (3.8). Then, the empirical measure \( \hat{\mu}_m^N \) of the particles remains inside the distributional ball \( \hat{\mu}_0^N \in \mathcal{P}_N \) for all \( m \in [0, NT] \cap \mathbb{N} \), with the probability of (at least) \( 1 - \delta \).

Let us make two remarks about the PDE in Eq. (4.5).

First, the seminal works of Otto [32], and Jordan, et al. [18] establishes an intriguing connection between the McKean-Vlasov type PDEs specified in (4.5) and the gradient flow on the Wasserstein manifolds. More specifically, the PDE equation in Eq. (4.5) can be thought of as the minimization of the energy functional

\[
(4.7a) \quad \inf_{\mu \in \mathcal{M}(\mathbb{R}^p)} E_\alpha(p_\xi(\xi)) = \frac{1}{\alpha} \int_{\mathbb{R}^p} R_\alpha(\xi, p_\xi(\xi))p_\xi(\xi)d\xi
\]

\[
(4.7b) \quad R_\alpha(\xi, p_\xi(\xi)) = -\alpha(\mathbb{E}_{\mu^N, \nu[y\varphi(x; \xi)]}^2 + \mathbb{E}_{\xi \sim \mu^N}[\left( \mathbb{E}_{\mu^N}[\varphi(x; \xi)\varphi(x; \tilde{\xi})] \right)^2],
\]

using the following gradient flow dynamics

\[
(4.8) \quad \frac{dp_\xi(\xi)}{dt} = -\eta \cdot \nabla p_\xi E_\alpha(p_\xi(\xi)), \quad p_\xi(0) = q_\xi(\xi),
\]

where \( \nabla p_\xi E(p_\xi(\xi)) = \nabla \mu^N(p_\xi(\xi)) \nabla R_\alpha(\mu^N(p_\xi(\xi))) \) is the Riemannian gradient of \( R_\alpha(p_\xi(\xi)) \) with respect to the metric of the Wasserstein manifold. This shows that when the number of particles in particle SGD (3.8) tends to infinity \( N \to +\infty \), their empirical distribution follows a gradient descent path for minimization of the population version (with respect to data samples) of the distributional optimization in Eq. (3.5). In this sense, the particle SGD is a ‘consistent’ approximation algorithm for solving the distributional optimization.

Second, notice that as the scaling parameter tends to infinity \( \alpha \to \infty \), the energy functional tends to the limit \( E_\alpha(p_\xi(\xi)) \to E_A(p_\xi(\xi)) \) of \( (\mathbb{E}_{\mu^N, \nu[y\varphi(x; \xi)]}^2 \). Interestingly, this limiting energy functional is precisely the kernel polarization of Baram [2], measuring the correlation between the class labels and the random feature maps.

4.4. Kernel Learning by Solving A PDE. The kernel selection methods in the literature focuses on optimization methods based on kernel alignment optimization problem; see, e.g., [6, 22, 7].

The PDE in Eq. (4.5) puts forth an alternative method to finding good kernel functions. Namely, a good kernel function can be computed in two stages: first compute the stationary solution \( p^*(\xi) \) corresponding to a solution of the PDE in Eq. (4.5) with \( \partial p^*_\xi(\xi)/\partial t = 0 \), where the expectation with respect to the unknown data distribution is replaced by its finite-sample average. Then approximating the kernel via a Monte-Carlo sampling \( \xi^1, \cdots, \xi^N \sim \text{i.i.d.} \), \( p^*(\xi) \),

\[
(4.9) \quad K_{p^*}(x, y) = \int_{\mathbb{R}^p} \varphi(x; \xi)\varphi(y; \xi)p^*(\xi)d\xi = \frac{1}{N} \sum_{i=1}^N \varphi(x; \xi^i)\varphi(y; \xi^i).
\]
While such a method can be successful when the random variable in random features \( \xi \in \mathbb{R}^D \) is low dimensional, in the high-dimensional settings \( (D \gg 1) \), solving the PDE in Eq. (4.5) numerically and sampling from a high dimensional density function \( p^*(\xi) \) appears to be computationally more expensive than solving the kernel-target alignment optimization via SGD.

4.5. Propagation of Chaos. We now establish the so called ‘propagation of chaos’ property of particle SGD. At a high level, the propagation of chaos means that when the number of samples \( \{\xi^k\}_{k=1}^N \) tends to infinity \( (N \to +\infty) \), their dynamics are decoupled.

**Definition 4.3.** (Exchangability) Let \( \nu \) be a probability measure on a Polish space \( \mathcal{S} \) and. For \( N \in \mathbb{N} \), we say that \( \nu^\otimes N \) is an exchangeable probability measure on the product space \( \mathcal{S}^n \) if it is invariant under the permutation \( \pi \overset{\text{def}}{=} (\pi(1), \cdots, \pi(N)) \) of indices. In particular,

\[
\nu^\otimes N(\pi \cdot B) = \nu^\otimes N(B),
\]

for all Borel subsets \( B \in \mathcal{B}(\mathcal{S}^n) \).

An interpretation of the exchangability condition (4.10) can be provided via De Finetti’s representation theorem which states that the joint distribution of an infinitely exchangeable sequence of random variables is as if a random parameter were drawn from some distribution and then the random variables in question were independent and identically distributed, conditioned on that parameter.

Next, we review the mathematical definition of chaoticity, as well as the propagation of chaos in the product measure spaces:

**Definition 4.4.** (Chaoticity) Suppose \( \nu^\otimes N \) is exchangeable. Then, the sequence \( \{\nu^\otimes k\}_{N \in \mathbb{N}} \) is \( \nu \)-chaotic if, for any natural number \( \ell \in \mathbb{N} \) and any test function \( f_1, f_2, \cdots, f_k \in C_0^2(\mathcal{S}) \), we have

\[
\lim_{N \to \infty} \left\langle \prod_{k=1}^\ell f_k(s^k), \nu^\otimes N(ds^1, \cdots, ds^N) \right\rangle = \prod_{k=1}^\ell \langle f_k, \nu \rangle.
\]

According to Eq. (4.11) of Definition 4.4, a sequence of probability measures on the product spaces \( \mathcal{S} \) is \( \nu \)-chaotic if, for fixed \( k \) the joint probability measures for the first \( k \) coordinates tend to the product measure \( \nu(ds_1)\nu(ds_2)\cdots \nu(ds_k) = \nu^\otimes k \) on \( \mathcal{S}^k \). If the measures \( \nu^\otimes N \) are thought of as giving the joint distribution of \( N \) particles residing in the space \( \mathcal{S} \), then \( \{\nu^\otimes N\} \) is \( \nu \)-chaotic if \( k \) particles out of \( N \) become more and more independent as \( N \) tends to infinity, and each particles distribution tends to \( \nu \). A sequence of symmetric probability measures on \( \mathcal{S}^N \) is chaotic if it is \( \nu \)-chaotic for some probability measure \( \nu \) on \( \mathcal{S} \).

If a Markov process on \( \mathcal{S}^N \) begins in a random state with the distribution \( \nu^\otimes N \), the distribution of the state after \( t \) seconds of Markovian random motion can be expressed in terms of the transition function \( K^N \) for the Markov process. The distribution at time \( t > 0 \) is the probability measure \( U_t^N \nu^\otimes N \) is defined by the kernel

\[
U_t^N \nu^\otimes N(B) \overset{\text{def}}{=} \int_{\mathcal{S}^N} K^N(s, B, t) \nu^\otimes N(ds).
\]

**Definition 4.5.** (Propagation of Chaos) A sequence functions

\[
\left\{K^N(s, B, t)\right\}_{N \in \mathbb{N}}
\]
whose $N$-th term is a Markov transition function on $\mathcal{S}^N$ that satisfies the permutation condition
\begin{equation}
K^N(s, B, t) = K^N(\pi \cdot s, \pi \cdot B, t),
\end{equation}
propagates chaos if whenever $\{\nu^N\} \in \mathcal{H}_N$ is chaotic, so is $\{U^N_t\}$ for any $t \geq 0$, where $U^N_t$ is defined in Eq. (4.12).

We note that for finite systems size $N$, the states of the particles are not independent of each other. However, as we prove in the following result, in the limiting system $N \to +\infty$, the particles are mutually independent. This phenomena is known as the propagation of chaos (a.k.a. asymptotic independence):

**Theorem 4.6.** (Chaoticity in Particle SGD) Consider Assumptions (A.1) – (A.3). Furthermore, suppose that $\{\xi^N_k\}_{1 \leq k \leq N} \sim \text{i.i.d.}$ $\mu_0$ is exchangeable in the sense that the joint law is invariant under the permutation of indices. Then, at each time instant $t \in (0, T]$, the scaled empirical measure $\mu^N_t \in \mathcal{M}(\mathbb{R}^p)$ defined via scaling
\begin{equation}
\mu_N^t(d\xi^1, \cdots, d\xi^N) \overset{\text{def}}{=} \frac{1}{N} \mathbb{P}\{|\xi^1|_{NT} \in d\xi^1, \cdots, \xi^N|_{NT} \in d\xi^N\},
\end{equation}
is $\mu^*_t$-chaotic, where $\mu^*_t$ is mean-field solution of (A.67).

5. **Empirical evaluation**

We now turn to empirical evaluations. We test the performance of Algorithm 1 on synthetic data-set, as well as on benchmark data-sets.

6. **Experimental Results on the Synthetic Data-Set**

The synthetic data-set we consider is as follows:

- The distribution of training data is $P_V = \mathcal{N}(0, (1 + \lambda)I_{d \times d})$.
- The distribution of generated data is $P_W = \mathcal{N}(0, (1 - \lambda)I_{d \times d})$.

To reduce the dimensionality of data, we consider the embedding $\iota: \mathbb{R}^d \to \mathbb{R}^p, x \to \iota(x) = \Sigma x$, where $\Sigma \in \mathbb{R}^{p \times d}$ and $p < d$. In this case, the distribution of the embedded features are $P_{X|Y=+1} = \mathcal{N}(0, (1 + \lambda)\Sigma \Sigma^T)$, and $P_{X|Y=-1} = \mathcal{N}(0, (1 - \lambda)\Sigma \Sigma^T)$.

Note that $\lambda \in [0, 1]$ is a parameter that determines the separation of distributions. In particular, the Kullback-Leibler divergence of the two multi-variate Gaussian distributions is controlled by $\lambda \in [0, 1]$,
\begin{equation}
D_{KL}(P_{X|Y=+1}, P_{X|Y=-1}) = \frac{1}{2} \left[ \log \left( \frac{1 - \lambda}{1 + \lambda} \right) - p + p(1 - \lambda^2) \right].
\end{equation}

In Figure 1, we show the distributions of i.i.d. samples from the distributions $P_V$ and $P_W$ for different choices of variance parameter of $\lambda = 0.1, 0.5$, and $\lambda = 0.9$. Notice that for larger $\lambda$ the divergence is reduced and thus performing the two-sample test is more difficult. From Figure 1, we clearly observe that for large values of $\lambda$, the data-points from the two distributions $P_V$ and $P_W$ have a large overlap and conducting a statistical test to distinguish between these two distributions is more challenging.
Figure 1. Visualization of data-points from the synthetic data-set $P_V = N(0, (1 + \lambda) I_d \times d)$ and $P_W = N(0, (1 - \lambda) I_d \times d)$ for $d = 2$. Panel (a): $\lambda = 0.1$, Panel (b): $\lambda = 0.5$, and Panel (c): $\lambda = 0.9$.

Figure 2. The evolution of the empirical measure $\mu_{nm}^N(\xi) = \frac{1}{N} \sum_{k=1}^{N} \delta(\xi - \xi_{km})$ of the SGD particles $\xi_{1m}, \cdots, \xi_{Nm} \in \mathbb{R}^2$ at different iterations $m$. The empirical measure of random feature maps seemingly converges to a Gaussian stationary measure corresponding to a Gaussian RBF kernel. Panel (a): $m = 0$, Panel (b): $m = 300$, Panel (c): $m = 1000$, and Panel (d): $m = 2500$. 
Figure 3. The statistical power versus the threshold \( \tau \) for the binary hypothesis testing via the unbiased estimator of the kernel MMD. The parameters for this simulations are \( \lambda \in \{0.1, 0.5, 0.9\} \), \( d = 100 \), \( n + m = 100 \), \( p = 50 \). Panel (a): Trained kernel using the two-phase procedure with the particle SGD in (3.8) and an auto-encoder. Panel (b): Trained kernel with an auto-encoder and a fixed Gaussian kernel with the bandwidth \( \sigma = 1 \). Panel (c): Untrained kernel without an auto-encoder.

6.0.1. Kernel Learning Approach. Figure 3 depicts our two-phase kernel learning procedure which we also employed in our implementations of Algorithm 1 on benchmark data-sets of Section 6.1 in the main text. The kernel learning approach consists of training the auto-encoder and the kernel optimization sequentially, i.e.,

\[
\sup_{\hat{\mu} \in \mathcal{P}} \sup_{\iota} \text{MMD}^\alpha_{K,\hat{\mu}}[\mathcal{P}_V, \mathcal{P}_W].
\]

(6.2)

where the function class is defined \( \mathcal{Q} \xlongequal{\text{def}} \{\iota(z) = \Sigma z, \Sigma \in \mathbb{R}^{p \times d}\} \), and \( (K_{\hat{\mu} \iota} \circ \iota)(x_1, x_2) = K_{\hat{\mu} \iota}(\iota(x_1), \iota(x_2)) \). Now, we consider a two-phase optimization procedure:

- **Phase (I):** we fix the kernel function, and optimize the auto-encoder to compute a co-variance matrix \( \Sigma \) for dimensionality reduction
- **Phase (II):** we optimize the kernel based on the learned embedded features.

This two-phase procedure significantly improves the computational complexity of SGD as it reduces the dimensionality of random feature samples \( \xi \in \mathbb{R}^D \), \( D \ll d \). When the kernel function \( K \) is fixed, optimizing the auto-encoder is equivalent to the kernel learning step of [24].

6.0.2. Statistical Hypothesis Testing with the Kernel MMD. Let \( V_1, \ldots, V_m \xsim \text{i.i.d.} \mathcal{P}_V = \mathcal{N}(0, (1 + \lambda)I_{d \times d}) \), and \( W_1, \ldots, W_n \xsim \text{i.i.d.} \mathcal{P}_W = \mathcal{N}(0, (1 - \lambda)I_{d \times d}) \). Given these i.i.d. samples, the statistical test \( \mathcal{T}(\{V_i\}_{i=1}^m, \{W_j\}_{j=1}^n) : \mathcal{Y}^m \times \mathcal{W}^n \to \{0, 1\} \) is used to distinguish between these hypotheses:

- **Null hypothesis** \( H_0 : \mathcal{P}_V = \mathcal{P}_W \) (thus \( \lambda = 0 \)),
- **Alternative hypothesis** \( H_1 : \mathcal{P}_V \neq \mathcal{P}_W \) (thus \( \lambda > 0 \)).

To perform hypothesis testing via the kernel MMD, we require that \( \mathcal{H}_X \) is a universal RKHS, defined on a compact metric space \( \mathcal{X} \). Universality requires that the kernel \( K(\cdot, \cdot) \) be continuous and, \( \mathcal{H}_X \) be dense in \( C(\mathcal{X}) \). Under these conditions, the following theorem establishes that the kernel MMD is indeed a metric:

**Theorem 6.1.** (Metrizability of the RKHS) Let \( \mathcal{F} \) denotes a unit ball in a universal RKHS \( \mathcal{H}_X \) defined on a compact metric space \( \mathcal{X} \) with the associated continuous kernel \( K(\cdot, \cdot) \).
Then, the kernel MMD is a metric in the sense that \( \text{MMD}_K[P_V, P_W] = 0 \) if and only if \( P_V = P_W \).

To design a test, let \( \hat{\mu}_m^N(\xi) = \frac{1}{N} \sum_{k=1}^N \delta(\xi - \xi_m^k) \) denotes the solution of SGD in (3.8) for solving the optimization problem. Consider the following MMD estimator consisting of two \( U \)-statistics and an empirical function

\[
\overline{\text{MMD}}_{K, m}N \left[ \{ V_i \}_{i=1}^m; \{ W_i \}_{i=1}^n \right] = \frac{1}{m(m - 1)} \sum_{k=1}^N \sum_{i \neq j} \varphi(i(V_i), \xi_m^k) \varphi(i(V_j), \xi_m^k) \\
+ \frac{1}{n(n - 1)} \sum_{k=1}^N \sum_{i \neq j} \varphi(i(W_i), \xi_m^k) \varphi(i(W_j), \xi_m^k) \\
- \frac{1}{nm} \sum_{k=1}^N \sum_{i=1}^m \sum_{j=1}^n \varphi(i(W_i), \xi_m^k) \varphi(i(V_j), \xi_m^k).
\]

(6.3)

Given the samples \( \{ V_i \}_{i=1}^m \) and \( \{ W_i \}_{i=1}^n \), we design a test statistic as below

\[
\mathcal{T}(\{ V_i \}_{i=1}^m, \{ W_i \}_{i=1}^n) \overset{\text{def}}{=} \begin{cases} 
H_0 & \text{if } \overline{\text{MMD}}_{K, m}N \left[ \{ V_i \}_{i=1}^m; \{ W_i \}_{i=1}^n \right] \leq \tau \\
H_1 & \text{if } \overline{\text{MMD}}_{K, m}N \left[ \{ V_i \}_{i=1}^m; \{ W_i \}_{i=1}^n \right] > \tau .
\end{cases}
\]

(6.4)

where \( \tau \in \mathbb{R} \) is a threshold. Notice that the unbiased MMD estimator of (6.3) can be negative despite the fact that the population MMD is non-negative. Consequently, negative values for the statistical threshold \( \tau \) (6.4) are admissible. In the following simulations, we only consider non-negative values for the threshold \( \tau \).

A Type I error is made when \( H_0 \) is rejected based on the observed samples, despite the null hypothesis having generated the data. Conversely, a Type II error occurs when \( H_0 \) is accepted despite the alternative hypothesis \( H_1 \) being true. The significance level \( \alpha \) of a test is an upper bound on the probability of a Type I error: this is a design parameter of the test which must be set in advance, and is used to determine the threshold to which we compare the test statistic. The power of a test is the probability of rejecting the null hypothesis \( H_0 \) when it is indeed incorrect. In particular,

\[
\text{Power} = \mathbb{P}(\text{reject } H_0 | H_1 \text{ is true}).
\]

In this sense, the statistical power controls the probability of making Type II errors.

6.0.3. Empirical Results. In Figure 2, we show the evolution of the empirical measure \( \mu_m^N(\xi) \) of SGD particles by plotting the 2D histogram of the particles \( \xi_m^1, \ldots, \xi_m^N \in \mathbb{R}^d \) at different iterations of SGD. Clearly, starting with a uniform distribution in 2(a), the empirical measure seemingly evolves into a Gaussian measure in Figure 2(d). The evolution to a Gaussian distribution demonstrates that the RBF Gaussian kernel corresponding to a Gaussian distribution for the random features indeed provides a good kernel function for the underlying hypothesis test with Gaussian distributions.

In Figure 3, we evaluate the power of the test for 100 trials of hypothesis test using the test statistics of (6.4). To obtain the result, we used an autoencoder to reduce the dimension from \( d = 100 \) to \( p = 50 \). Clearly, for the trained kernel in Panel (a) of Figure 3, the threshold \( \tau \) for which Power = 1 increases after learning the kernel via the two phase procedure described earlier. In comparison, in Panel (b), we observe that training an autoencoder only with a fixed standard Gaussian kernel \( K(x, y) = \exp(-\|x - y\|^2_2) \) attains lower thresholds compared to our two-phase procedure. In Panel (c), we demonstrate the case of
Figure 4. Sample generated images using CIFAR-10 (top row), and MNIST (bottom row) data-sets. Panels (a)-(e): Proposed MMD GAN with an automatic kernel selection via the particle SGD (Algorithm 1), Panels (b)-(f): MMD GAN [24] with an auto-encoder for dimensionality reduction in conjunction with a mixed RBF Gaussian kernel whose bandwidths are manually tuned, Panels (c)-(g): MMD GAN in [24] with a single RBF Gaussian kernel with an auto-encoder for dimensionality reduction in conjunction with a single RBF Gaussian kernel whose bandwidth is manually tuned, and Panel (d)-(g): GMMN without an auto-encoder [26].

6.1. Performance on benchmark datasets. We evaluate our kernel learning approach on large-scale benchmark data-sets. We train our MMD GAN model on two distinct types of data-sets, namely on MNIST [23] and CIFAR-10 [23], where the size of training instances are $60 \times 10^3$ and $50 \times 10^3$, respectively. All the generated samples are from a fixed noise random vectors and are not singled out.

6.1.1. Implementation and hyper-parameters. We implement Algorithm 1 as well as MMD GAN [24] in Pytorch using NVIDIA Titan V100 32GB graphics processing units (GPUs). The source code of Algorithm 1 is built upon the code of [24], and retains the auto-encoder implementation. In particular, we use a sequential training of the auto-encoder and kernel as explained in the synthetic data-set. For a fair comparison, our hyper-parameters are adjusted as in [24], i.e., the learning rate of 0.00005 is considered for RMSProp [46]. Moreover,
the batch-size for training the generator and auto-encoder is \( n = 64 \). The learning rate of particle SGD is tuned to \( \eta = 10 \).

6.1.2. Random feature maps. To approximate the kernel, we use the random feature model of Rahimi and Recht [36, 37], where \( \varphi(x; \xi) = \sqrt{2} \cos(x^T \xi + b) \). Here \( b \sim \text{Uniform}\{-1, +1\} \) is a random bias term.

6.1.3. Practical considerations. When data-samples \( \{V_i\} \in \mathbb{R}^d \) are high dimensional (e.g. CIFAR-10), the particles \( \xi_1, \ldots, \xi_N \in \mathbb{R}^p, p = d \) in SGD (3.8) are also high-dimensional. To reduce the dimensionality of the particles, we apply an auto-encoder architecture similar to [24], and train our kernel on top of learned embedded features. More specifically, in our simulations, we train an auto-encoder where the dimensionality of the latent space is \( h = 10 \) for MNIST, and \( h = 128 \) (thus \( p = d = 128 \)) for CIFAR-10. Therefore, the particles \( \xi_1, \ldots, \xi_N \) in subsequent kernel training phase have the dimension of \( D = 10 \), and \( D = 128 \), respectively.

6.1.4. Choice of the scaling parameter \( \alpha \). There is a trade-off in the choice of \( \alpha \). While for large values of \( \alpha \), the kernel is better able to separate data-samples from generated samples, in practice, a large value of \( \alpha \) slows down the convergence of particle SGD. This is due to the fact that the coupling strength between the particles in Eq. (3.8) decrease as \( \alpha \) increase. The scaling factor is set to be \( \alpha = 1 \) in all the following experiments.

6.1.5. Qualitative comparison. We now show that without the bandwidth tuning for Gaussian kernels and using the particle SGD to learn the kernel, we can attain better visual results on benchmark data-sets. In Figure 4, we show the generated samples on CIFAR-10 and MNIST data-sets, using our Algorithm 1, MMD GAN [24] with a mixed and homogeneous Gaussian RBF kernels, and GMMN [26].

Figure 4(a) shows the samples from Algorithm 1, Figure 4(b) shows the samples from MMD GAN [24] with the RBF Gaussian mixture kernel of Eq. , where \( \sigma_k \in \{1, 2, 4, 8, 16\} \) are the bandwidths of the Gaussian kernels that are fine tuned and optimized. We observe that our MMD GAN with automatic kernel learning visually attains similar results to MMD GAN [24] which requires manual tuning of the hyper-parameters. In Figure 4(c), we show the MMD GAN result with a single kernel RBF Gaussian kernel whose bandwidth is manually selected at \( \sigma = 16 \). Lastly, in Figure 4(d), we show the samples from GMMN [26] which does not exploit an auto-encoder or kernel training. Clearly, GMMN yield a poor results compared to other methods due to high dimensionality of features, as well as the lack of an efficient method to train the kernel.

On MNIST data-set in Figure 4(e)-(h), the difference between our method and MMD GAN [24] is visually more pronounced. We observe that without a manual tuning of the kernel bandwidth and by using the particle SGD (3.8) to optimize the kernel, we attain better generated images in Figure 4(e), compared to MMD GAN with mixed RBF Gaussian kernel and manual bandwidth tuning in Figure 4(f). Moreover, using a single RBF Gaussian kernel yields a poor result regardless of the choice of its bandwidth. The generated images from GMMN is also shown in Figure 4(h).

6.1.6. Quantitative comparison. To quantitatively measure the quality and diversity of generated samples, we compute the inception score (IS) [42] as well as Fréchet Inception Distance (FID) [15] on CIFAR-10 images.
| Method                                      | FID (↓)       | IS (↑)    |
|--------------------------------------------|---------------|-----------|
| MMD GAN (Gaussian) [24]                    | 67.244 ± 0.134| 5.608 ± 0.051 |
| MMD GAN (Mixture Gaussian) [24]            | 67.129 ± 0.148| 5.850 ± 0.055 |
| SGD Alg. 1                                 | **65.059 ± 0.153** | **5.97 ± 0.046** |
| Benchmark                                  | -             | 11.237 ± 0.116  |

Table 1. Comparison of the quantitative performance measures of MMD GANs with different kernel learning approaches.

Intuitively, the inception score is used for GANs to measure samples quality and diversity. This score is based on the Inception-v3 Network [45] which is a deep convolutional architecture designed for classification tasks on ImageNet [8], a dataset consisting of 1.2 million RGB images from 1000 classes. Given an image \( x \), the task of the network is to output a class label \( y \) in the form of a vector of probabilities. The inception score uses an Inception-v3 Network pre-trained on ImageNet and calculates a statistic of the networks outputs when applied to generated images. More precisely, the inception score of a generative model is

\[
\text{IS} \overset{\text{def}}{=} \exp \left( \mathbb{E}_{X \sim P_W} D_{\text{KL}}(P_{Y|X}||P_Y) \right),
\]

where \( D_{\text{KL}}(\cdot||\cdot) \) is the Kullback-Leibler divergence. The definition of the inception score is motivated by the following two observations:

(i) The images generated should contain meaningful objects, i.e., \( P_{Y|X} \) should be low entropy. In other words, images with meaningful objects are supposed to have low label (output) entropy, that is, they belong to few object classes.

(ii) The generative algorithm should generate diverse images from all the different classes in ImageNet, i.e., the distribution of labels \( P_Y \) should have a high entropy.

The FID improves on IS by actually comparing the statistics of generated samples to real samples, instead of evaluating generated samples independently. In particular, Heusel, et al. [15] propose to use the Fréchet distance between two multivariate Gaussians \( N(\mu_1, \Sigma_1) \) and \( N(\mu_2, \Sigma_2) \) as follows

\[
\text{FID} \overset{\text{def}}{=} \|\mu_1 - \mu_2\|_2^2 + \text{Tr} \left( \Sigma_1 + \Sigma_2 - 2(\Sigma_1 \Sigma_2)^{\frac{1}{2}} \right).
\]

In Table 1, we report the quantitative measures for different MMD GAN model using different scoring metric. Note that in Table 1 lower FID scores and higher IS scores indicate a better performance. We observe from Table 1 that our approach attain lower FID score, and higher IS score compared to MMD GAN with single Gaussian kernel (bandwidth \( \sigma = 16 \)), and a mixture Gaussian kernel (bandwidths \{1, 2, 4, 8, 16\}).

7. Acknowledgement

M.B.K. would like to thank his colleague Dr. CHUANG WANG from Harvard University for his intriguing theoretical work on the scaling limits of the online algorithms [50] which motivated this work. M.B.K. would also like to thank PROF. RAVI MAZUMDAR from The University of Waterloo for his theoretical paper [48] which inspired the proof of tightness of Skhorhod spaces in this paper. We also thank Dr. MAXIME BASSENE for useful discussion regarding the MacKean-Vlasov PDE.

The research of MBK, LS, and LX is supported by NIH under the grant 1R01 CA176553 and R01E0116777. The content of this article are solely the responsibility of the authors and do not necessarily reflect the official NIH views.
Appendix

Appendix A. Proofs of Main Theoretical Results

Notation: We denote vectors by lower case bold letters, e.g. \( \mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n \), and matrices by the upper case bold letters, e.g., \( \mathbf{M} = [M_{ij}] \in \mathbb{R}^{n \times m} \). The Frobenius norm of a matrix is denoted by \( \|\mathbf{M}\|_F = \sum_{i=1}^n \sum_{j=1}^m |M_{ij}|^2 \). Let \( \mathbb{B}_r(x) \equiv \{ y \in \mathbb{R}^d : \|y - x\|_2 \leq r \} \) denote the Euclidean ball of radius \( r \) centered at \( x \). For a given metric space \( \mathcal{X} \), let \( \mathcal{M}(\mathcal{X}) \) denote the class of the integrable functions \( f \). We denote the Borel \( \sigma \)-algebra by \( \mathcal{B}(\mathcal{X}) \). We assume that \( \mathcal{X} \) is a Polish space. The Dirac measure with the unit mass at \( x \in \mathcal{X} \) is denoted by \( \delta(x) \). Further, \( C_b^k(\mathcal{X}) \) denotes the space of functions whose partial derivatives up to order \( k \) are bounded and continuous, and \( C_b^k(\mathcal{X}) \) the space of functions whose partial derivatives up to order \( k \) are continuous with compact support.

We use asymptotic notations throughout the paper. We use the standard asymptotic notation for sequences. If \( \{a_n\}_{n \in \mathbb{N}} \) and \( \{b_n\}_{n \in \mathbb{N}} \) are positive sequences, then \( a_n = \Theta(b_n) \) means that \( \lim_{n \to \infty} a_n/b_n < \infty \), whereas \( a_n = \Omega(b_n) \) means that \( \lim_{n \to \infty} a_n/b_n > 0 \). Furthermore, \( a_n = \tilde{O}(b_n) \) implies \( a_n = O(b_n \text{poly}(b_n)) \). Moreover \( a_n = o(b_n) \) means that \( \lim_{n \to \infty} a_n/b_n = 0 \) and \( a_n = \omega(b_n) \) means that \( \lim_{n \to \infty} a_n/b_n = \infty \). Lastly, we have \( a_n = \Theta(b_n) \) if \( a_n = O(b_n) \) and \( a_n = \Omega(b_n) \). Finally, for positive \( a, b > 0 \), denote \( a \lesssim b \) if \( a/b \) is at most some universal constant.

A.1. Proof of Theorem 4.1. By the triangle inequality, we have that

\[
\text{MMD}_{\mu_n}[P_V, P_{\tilde{W}^n_{\nu}}] - \text{MMD}_{\mu^N_{i=1} \equiv n}[P_V, P_{\tilde{W}^n_{\nu}}] \leq A_1 + A_2 + A_3 + A_4,
\]

\[\text{A.1.} \quad \|f\|_\infty \equiv \sup_{x \in \mathcal{X}} |f(x)|.\]

Further, \( C_b^k(\mathcal{X}) \) the space of all functions in \( C_b(\mathcal{X}) \) whose partial derivatives up to order \( k \) are bounded and continuous, and \( C_b^k(\mathcal{X}) \) the space of functions whose partial derivatives up to order \( k \) are continuous with compact support.

We denote the class of the integrable functions \( f \) with \( f(t) \geq 0 \) a.e., on \( 0 \leq t \leq T \) by \( L^1[0, T] \). Similarly, \( L^\infty[0, T] \) will denote the essentially bounded functions with \( f(t) \geq 0 \) almost everywhere. For a given metric space \( \mathcal{X} \), we denote the Borel \( \sigma \)-algebra by \( \mathcal{B}(\mathcal{X}) \). For a Borel set \( B \in \mathcal{B}(\mathcal{X}) \), the measure value of the set \( B \) with respect to the measure is given by \( \mu(B) \). The space of finite non-negative measures defined on \( \mathcal{X} \) is denoted by \( \mathcal{M}(\mathcal{X}) \). The Dirac measure with the unit mass at \( x \in \mathcal{X} \) is denoted by \( \delta(x) \). For any measure \( \mu \in \mathcal{M}(\mathcal{X}) \) and any bounded function \( f \in C_b(\mathcal{X}) \), we define

\[
\langle \mu, f \rangle \equiv \int_{\mathcal{X}} f(x) \mu(dx).
\]

The space \( \mathcal{M}(\mathcal{X}) \) is equipped with the weak topology, i.e., a (random) sequence \( \{\mu^N\}_{N \in \mathbb{N}} \) converges weakly to a deterministic measure \( \mu \in \mathcal{M}(\mathcal{X}) \) if and only if \( \langle \mu^N, f \rangle \to \langle \mu, f \rangle \) for all \( f \in C_b(\mathcal{X}) \). We denote the weak convergence by \( \mu^N \rightharpoonup \mu \). Notice that when \( \mathcal{X} \) is Polish, then \( \mathcal{M}(\mathcal{X}) \) equipped with the weak topology is also Polish.\(^4\) For a Polish space \( \mathcal{X} \), let \( \mathcal{D}_c([0, T]) \) denotes the Skorokhod space of the càdlàg functions that take values in \( \mathcal{X} \) defined on \([0, T]\). We assume that \( \mathcal{D}_c([0, T]) \) is equipped with the Skorokhod’s \( J_1 \)-topology \([3]\), which in that case \( \mathcal{D}_c([0, T]) \) is also a Polish space.

We use asymptotic notations throughout the paper. We use the standard asymptotic notation for sequences. If \( a_n \) and \( b_n \) are positive sequences, then \( a_n = \Theta(b_n) \) means that \( \limsup_{n \to \infty} a_n/b_n < \infty \), whereas \( a_n = \Omega(b_n) \) means that \( \liminf_{n \to \infty} a_n/b_n > 0 \). Furthermore, \( a_n = \tilde{O}(b_n) \) implies \( a_n = O(b_n \text{poly}(b_n)) \). Moreover \( a_n = o(b_n) \) means that \( \lim_{n \to \infty} a_n/b_n = 0 \) and \( a_n = \omega(b_n) \) means that \( \lim_{n \to \infty} a_n/b_n = \infty \). Lastly, we have \( a_n = \Theta(b_n) \) if \( a_n = O(b_n) \) and \( a_n = \Omega(b_n) \). Finally, for positive \( a, b > 0 \), denote \( a \lesssim b \) if \( a/b \) is at most some universal constant.

\(^4\)A topological space is Polish if it is homeomorphic to a complete, separable metric space.
where the terms $A_i, i = 1, 2, 3, 4$ are defined as follows

\[
A_1 \overset{\text{def}}{=} |\text{MMD}_\mu[P_V, P_{W_t}] - \sup_{W \in \mathcal{W}} \text{MMD}_\mu[P_V, P_W]|
\]

\[
A_2 \overset{\text{def}}{=} \left| \min_{W \in \mathcal{W}} \text{MMD}_\mu[P_V, P_W] - \min_{W \in \mathcal{W}} \text{MMD}_{\mu_N}[P_V, P_W] \right|
\]

\[
A_3 \overset{\text{def}}{=} \left| \min_{W \in \mathcal{W}} \text{MMD}_{\mu_N}[P_V, P_W] - \text{MMD}_{\mu_N}[P_V, P_{W_N}] \right|
\]

\[
A_4 \overset{\text{def}}{=} \left| \text{MMD}_{\mu_N}[P_V, P_{W_N}] - \text{MMD}_{\mu_N}[P_V, P_{W_N}] \right|
\]

In the sequel, we compute an upper bound for each term on the right hand side of Eq. (A.3):

**Upper bound on $A_1$:**

First, notice that the squared kernel MMD loss in Eq. (2.3) can be characterized in terms of class labels and features defined in Section 3.1 as follows

\[
(\text{A.4}) \quad \text{MMD}_\mu[P_V, P_W] = 4E_{P_{x,y}^\otimes 2}[y\hat{y}K_\mu(x, \hat{x})].
\]

To see this equivalence, we first rewrite the right hand side of Eq. (A.4) as follows

\[
E_{P_{x,y}^\otimes 2}[y\hat{y}K_\mu(x, \hat{x})] = \mathbb{P}\{y = +1\} \mathbb{P}\{\hat{y} = +1\} E_{x,\hat{x} \sim P_{x\mid y=+1}^\otimes 2}[K_\mu(x, \hat{x})]
\]

\[+ \mathbb{P}\{y = -1\} \mathbb{P}\{\hat{y} = -1\} E_{x,\hat{x} \sim P_{x\mid y=-1}^\otimes 2}[K_\mu(x, \hat{x})]
\]

\[+ \mathbb{P}\{y = +1\} \mathbb{P}\{\hat{y} = +1\} E_{x,\hat{x} \sim P_{x\mid y=-1}^\otimes 2}[K_\mu(x, \hat{x})]
\]

\[+ \mathbb{P}\{y = +1\} \mathbb{P}\{\hat{y} = -1\} E_{x,\hat{x} \sim P_{x\mid y=+1}^\otimes 2}[K_\mu(x, \hat{x})].
\]

Now, recall from Section 3.1 that $P_{x\mid y=+1} = P_V$, and $P_{x\mid y=-1} = P_W$ by construction of the labels and random features. Moreover, $y, \hat{y} \sim \text{i.i.d.}$ Uniform\{-1, +1\}, and thus $\mathbb{P}\{y = -1\} = \mathbb{P}\{y = +1\} = \frac{1}{2}$. Therefore, from Eq. (A.5), we derive

\[
E_{P_{x,y}^\otimes 2}[y\hat{y}K_\mu(x, \hat{x})] = \frac{1}{4} E_{P_{x\mid y=+1}^\otimes 2}[K_\mu(x, \hat{x})] + \frac{1}{4} E_{P_{x\mid y=-1}^\otimes 2}[K_\mu(x, \hat{x})] - \frac{1}{2} E_{P_V, P_W}[K_\mu(x, \hat{x})]
\]

\[= \frac{1}{4} \text{MMD}_\mu[P_V, P_W].
\]

For any given $W \in \mathcal{W}$, we have that

\[
\left| \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] - \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] \right|
\]

\[\leq \sup_{\mu \in \mathcal{P}} \left| \text{MMD}_\mu[P_V, P_W] - \text{MMD}_\mu[P_V, P_W] \right|
\]

\[= 4 \sup_{\mu \in \mathcal{P}} \left| \frac{1}{n(n-1)} \sum_{i \neq j} y_i y_j K_\mu(x_i, x_j) - E_{P_{x,y}^\otimes 2}[y\hat{y}K_\mu(x, \hat{x})] \right|
\]

\[= 4 \sup_{\mu \in \mathcal{P}} \left| E_{\mu}[E_n(\xi)] \right|
\]

\[\leq 4 \sup_{\mu \in \mathcal{P}} \left| E_{\mu}[E_n(\xi)] \right|
\]
where the error term is defined using the random features
\begin{equation}
E_n(\xi) \overset{\text{def}}{=} \frac{1}{n(n-1)} \sum_{i \neq j} y_i y_j \varphi(x_i; \xi) \varphi(x_j; \xi) - \mathbb{E}_{P_{x,y}^{\otimes 2}}[y_i y_j \varphi(x; \xi) \varphi(\hat{x}; \xi)].
\end{equation}

Now, we invoke the following strong duality theorem [10]:

**Theorem A.1. (Strong Duality for Robust Optimization, [10, Theorem 1])** Consider the general metric space \((\Xi, d)\), and any normal distribution \(\nu \in \mathcal{M}(\Xi)\), where \(\mathcal{M}(\Xi)\) is the set of Borel probability measures on \(\Xi\). Then,
\begin{equation}
\sup_{\mu \in \mathcal{M}(\Xi)} \{ \mathbb{E}_\mu[\Psi(\xi)] : W_p(\mu, \nu) \leq R \} = \min_{\lambda \geq 0} \left\{ \lambda R^p - \int_{\Xi} \inf_{\xi \in \Xi} \lambda d^p(\xi, \zeta) - \Psi(\xi) \nu(d\zeta) \right\},
\end{equation}
providing that \(\Psi\) is upper semi-continuous in \(\xi\).

Under the strong duality of Theorem A.1, we obtain that
\begin{equation}
\left| \sup_{\mu \in \mathcal{P}} \tilde{\text{MMD}}_\mu[P_V, P_W] - \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] \right| \leq 4 \left| \min_{\lambda \geq 0} \left\{ \lambda R^p - \int_{\mathbb{R}^d} \inf_{\xi \in \mathbb{R}^d} \lambda \|\xi - \zeta\|_2^p - E_n(\zeta) \right\} \mu_0(d\xi) \right|.
\end{equation}

In the sequel, let \(p = 2\). The Moreau’s envelope [33] of a function \(f : \mathcal{X} \to \mathbb{R}\) is defined as follows
\begin{equation}
M^\beta_f(y) \overset{\text{def}}{=} \inf_{x \in \mathcal{X}} \left\{ \frac{1}{2\beta} \|x - y\|_2^2 + f(x) \right\}, \quad \forall y \in \mathcal{X},
\end{equation}
where \(\beta > 0\) is the regularization parameter. When the function \(f\) is differentiable, the following lemma can be established:

**Lemma A.2. (Moreau’s envelope of differentiable functions)** Suppose the function \(f : \mathcal{X} \to \mathbb{R}\) is differentiable. Then, the Moreau’s envelope defined in Eq. (A.9) has the following upper bound and lower bounds
\begin{equation}
f(y) - \frac{\beta}{2} \int_0^1 \sup_{x \in \mathcal{X}} \|\nabla f(y + s(x - y))\|_2^2 ds \leq M^\beta_f(y) \leq f(y).
\end{equation}

In particular, when \(f\) is \(L_f\)-Lipschitz, we have
\begin{equation}
f(y) - \frac{\beta L_f^2}{2} \leq M^\beta_f(y) \leq f(y).
\end{equation}

The proof is presented in Appendix B.1.
Now, we return to Equation (A.8). We leverage the lower bound on Moreau’s envelope in Eq. (A.10) of Lemma A.2 as follows

\[
\begin{align*}
&\sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] - \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] \\
&\leq 4 \min_{\lambda \geq 0} \left\{ \lambda R^2 - \int_{\mathbb{R}^d} M_{\lambda}^2 E_n(\xi) \mu_0(d\xi) \right\} \\
&\leq 4 \min_{\lambda \geq 0} \left\{ \lambda R^2 + \mathbb{E}\mu_0[E_n(\xi)] + \frac{1}{4\lambda} \mathbb{E}\mu_0 \left[ \int_0^1 \sup_{\xi \in \mathbb{R}^d} \| \nabla E_n((1-s)\xi + s\zeta) \|^2 ds \right] \right\} \\
&\leq 4 \| \mathbb{E}\mu_0[E_n(\xi)] \| + 4 \mathbb{E}\mu_0 \left[ \int_0^1 \sup_{\xi \in \mathbb{R}^d} \| \nabla E_n((1-s)\xi + s\zeta) \|^2 ds \right].
\end{align*}
\]

Let \( \zeta_* = \zeta_*(\xi, s) = \arg \sup_{\xi \in \mathbb{R}^d} \| \nabla E_n((1-s)\xi + s\zeta) \| \). Then, applying the union bound in conjunction with Inequality (A.12) yields

\[
\begin{align*}
&\mathbb{P}\left( \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] - \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] \geq \delta \right) \\
&\leq \mathbb{P}\left( \int_{\mathbb{R}^d} E_n(\xi) \mu_0(d\xi) \geq \frac{\delta}{8} \right) + \mathbb{P}\left( \int_{\mathbb{R}^d} \int_0^1 \| \nabla E_n((1-s)\xi + s\zeta_*) \|^2 ds \mu_0(d\xi) \geq \frac{\delta}{8R} \right).
\end{align*}
\]

To proceed from Eq. (A.13), we require a few definitions in the sequel:

**Definition A.3. (Orlicz Norm)** The Young-Orlicz modulus is a convex non-decreasing function \( \psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) such that \( \psi(0) = 0 \) and \( \psi(x) \rightarrow \infty \) when \( x \rightarrow \infty \). Accordingly, the Orlicz norm of an integrable random variable \( X \) with respect to the modulus \( \psi \) is defined as

\[
\| X \|_\psi \overset{\text{def}}{=} \inf \{ \beta > 0 : \mathbb{E}[\psi(|X| - \mathbb{E}[|X|]) / \beta] \leq 1 \}.
\]

In the sequel, we consider the Orlicz modulus \( \psi_\nu(x) \overset{\text{def}}{=} \exp(x^\nu) - 1 \). Accordingly, the cases of \( \| \cdot \|_\psi_2 \) and \( \| \cdot \|_\psi_1 \) norms are called the sub-Gaussian and the sub-exponential norms and have the following alternative definitions:

**Definition A.4. (Sub-Gaussian Norm)** The sub-Gaussian norm of a random variable \( Z \), denoted by \( \| Z \|_\psi_2 \), is defined as

\[
\| Z \|_\psi_2 = \sup_{q \geq 1} q^{-1/2}(\mathbb{E}[Z^q])^{1/q}.
\]

For a random vector \( Z \in \mathbb{R}^n \), its sub-Gaussian norm is defined as follows

\[
\| Z \|_\psi_2 = \sup_{x \in S^{n-1}} \| (x, Z) \|_\psi_2.
\]

**Definition A.5. (Sub-exponential Norm)** The sub-exponential norm of a random variable \( Z \), denoted by \( \| Z \|_\psi_1 \), is defined as follows

\[
\| Z \|_\psi_1 = \sup_{q \geq 1} q^{-1}(\mathbb{E}[|Z|^q])^{1/q}.
\]
For a random vector $Z \in \mathbb{R}^n$, its sub-exponential norm is defined below

\begin{equation}
\|Z\|_{\psi_1} = \sup_{x \in S^{n-1}} \|\langle Z, x \rangle\|_{\psi_1}.
\end{equation}

Now, we state the following lemma:

**Lemma A.6. (Tail Bounds for the Finite Sample Estimation Error)** Consider the estimation error $E_n$ defined in Eq. (A.6). Then, the following statements hold:

- $Z = \|\nabla E_n(\xi)\|_2^2$ is a sub-exponential random variable with the Orlicz norm of $\|Z\|_{\psi_1} \leq \frac{9 \times 2^\frac{2}{3} \times L^4}{n^2}$ for every $\xi \in \mathbb{R}^D$. Moreover,

\begin{equation}
\mathbb{P}\left( \int_{\mathbb{R}^D} \|\nabla E_n((1-s)\xi + s\zeta_*)\|_2^2 \, ds \mu_0(d\xi) \geq \delta \right) \leq 2e^{-\frac{n^2}{9 \times 2^{\frac{2}{3}} \times L^4} + \frac{L^4}{n}},
\end{equation}

- $E_n(\xi)$ is zero-mean sub-Gaussian random variable with the Orlicz norm of $\|E_n(\xi)\|_{\psi_2} \leq 16 \sqrt{\frac{3L^4}{n}}$ for every $\xi \in \mathbb{R}^D$. Moreover,

\begin{equation}
\mathbb{P}\left( \left| \int_{\mathbb{R}^D} E_n(\xi) \mu_0(d\xi) \right| \geq \delta \right) \geq 2e^{-\frac{n^2}{16 \sqrt{3L^4}}},
\end{equation}

The proof of Lemma A.6 is presented in Appendix A.2.

Now, we leverage the probability bounds (A.19) and (A.20) of Lemma A.6 to upper bound the terms on the right hand side of Eq. (A.13) as below

\begin{equation}
\mathbb{P}\left( \sup_{\mu \in \mathcal{P}} \widehat{\text{MMD}}_\mu[P_V, P_W] - \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] \geq \delta \right) \leq 2e^{-\frac{n^2}{9 \times 2^{\frac{2}{3}} \times L^4} + \frac{L^4}{n}} + 2e^{-\frac{n^2}{9 \times 2^{\frac{2}{3}} \times RL^4} + \frac{L^4}{n}} \leq 4 \max\left\{ e^{-\frac{n^2}{9 \times 2^{\frac{2}{3}} \times L^4}}, e^{-\frac{n^2}{9 \times 2^{\frac{2}{3}} \times RL^4} + \frac{L^4}{n}} \right\},
\end{equation}

where the last inequality comes from the basic inequality $a + b \leq 2 \max\{a, b\}$. Therefore, with the probability of at least $1 - \varrho$, we have that

\begin{equation}
\left| \sup_{\mu \in \mathcal{P}} \widehat{\text{MMD}}_\mu[P_V, P_W] - \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] \right| \leq \min\left\{ \frac{3^4 \times 2^\frac{11}{2} \times L^2}{n} \ln^\frac{1}{2} \left( \frac{4}{\varrho} \right), \frac{9 \times 2^{\frac{12}{3}} \times RL^4}{n^2} \ln \left( \frac{4e \frac{L^4}{n}}{\varrho} \right) \right\},
\end{equation}

for all $W \in \mathcal{W}$.

**Lemma A.7. (Distance between minima of Adjacent Functions)** Let $\Psi(W) : \mathcal{W} \to \mathbb{R}$ and $\Phi(W) : \mathcal{W} \to \mathbb{R}$. Further, suppose $\|\Psi(W) - \Phi(W)\|_\infty \leq \delta$ for some $\delta > 0$. Then,

\begin{equation}
\left| \min_{W \in \mathcal{W}} \Psi(W) - \min_{W \in \mathcal{W}} \Phi(W) \right| \leq \delta.
\end{equation}

See Appendix B.4 for the proof.
Let $\Psi(W) \overset{\text{def}}{=} \sup_{\mu \in \mathcal{P}} \widehat{\text{MMD}}_\mu[P_V, P_W]$, and $\Phi(W) \overset{\text{def}}{=} \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W]$. Then, from Inequality (A.23), we have the following upper bound on $A_1$

$$A_1 = \left| \text{MMD}_\mu[P_V, P_W] - \min_{w \in \mathcal{W}} \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] \right|
= \left| \min_{w \in \mathcal{W}} \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] - \min_{w \in \mathcal{W}} \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] \right|
\leq \min \left\{ \frac{3^4}{n} \times \left( 4 \sqrt{\frac{L^2}{\rho}} \right) \right\}
\leq \frac{9 \times 2^{11} \times RL^4}{n^2} \ln \left( \frac{4e \sqrt{\text{L}}}{\rho} \right)
\leq \frac{9 \times 2^{11} \times RL^4}{n^2} \ln \left( \frac{4e \sqrt{\text{L}}}{\rho} \right)
$$

(A.24)

with the probability of (at least) $1 - \varrho$.

**Upper bound on $A_2$:**

To establish the upper bound on $A_2$, we recall that

\begin{align*}
\text{(A.25a)} \quad \widehat{\text{MMD}}_{\tilde{\mu}^N}[P_V, P_W] &= \frac{1}{n(n-1)} \sum_{i \neq j}^{N} y_i y_j \varphi(x_i; \xi^j) \varphi(x_j; \xi^i) \\
\text{(A.25b)} \quad \text{MMD}_\mu[P_V, P_W] &= \frac{1}{n(n-1)} \sum_{i \neq j}^{N} y_i y_j \mathbb{E}_\mu[\varphi(x_i; \xi) \varphi(x_j; \xi)].
\end{align*}

Therefore,

$$\left| \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] - \sup_{\tilde{\mu}^N \in \mathcal{P}_N} \widehat{\text{MMD}}_{\tilde{\mu}^N}[P_V, P_W] \right| \leq \sup_{\mu \in \mathcal{P}} \left| \mathbb{E}_\mu[\varphi(x_i; \xi) \varphi(x_j; \xi)] - \sup_{\tilde{\mu}^N \in \mathcal{P}_N} \mathbb{E}_{\tilde{\mu}^N}[\varphi(x_i; \xi) \varphi(x_j; \xi)] \right|.
$$

(A.26)

Here, the last inequality is due to Theorem A.1 and the following duality results hold

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}_\mu[\varphi(x; \xi) \varphi(x; \xi)] = \inf_{\lambda \geq 0} \left\{ \lambda R^2 - \int_{\mathbb{R}_D^D} \inf_{\xi \in \mathbb{R}_D} \{ \lambda \|\xi - \xi\|^2 - \varphi(x; \xi) \varphi(x; \xi) \} \mu_0(d\xi) \right\}
$$

\begin{align*}
\sup_{\tilde{\mu}^N \in \mathcal{P}_N} \mathbb{E}_{\tilde{\mu}^N}[\varphi(x_i; \xi) \varphi(x_j; \xi)] &= \inf_{\lambda \geq 0} \left\{ \lambda R^2 - \frac{1}{N} \sum_{k=1}^{N} \inf_{\xi^k \in \mathbb{R}_D} \{ \lambda \|\xi^k - \xi\|^2 - \varphi(x_i; \xi) \varphi(x_j; \xi) \} \right\}
\end{align*}

Now, in the sequel, we establish a uniform concentration result for the following function

$$T^\lambda_{(\mathbf{x}, \tilde{\mathbf{x}})} : \mathbb{R}^{n \times D} \rightarrow \mathbb{R}
$$

$$(\xi_1^0, \cdots, \xi_N^0) \mapsto T^\lambda_{(\mathbf{x}, \tilde{\mathbf{x}})}(\xi_1^0, \cdots, \xi_N^0) = \frac{1}{N} \sum_{k=1}^{N} M^\lambda_{-\varphi(x_i; \varphi(\xi^k))} \cdots \varphi^\lambda_{-\varphi(x_i; \varphi(\xi^k))} \mu_0(d\xi).
$$

Then, from Eq. (A.26) we have

$$\left| \sup_{\mu \in \mathcal{P}} \text{MMD}_\mu[P_V, P_W] - \sup_{\tilde{\mu}^N \in \mathcal{P}_N} \widehat{\text{MMD}}_{\tilde{\mu}^N}[P_V, P_W] \right| \leq \sup_{\lambda \geq 0, \mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}} \left| T^\lambda_{(\mathbf{x}, \tilde{\mathbf{x}})}(\xi_1^0, \cdots, \xi_N^0) \right|.
$$

We now closely follow the argument of [36] to establish a uniform concentration result with respect to the data points $\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}$. In particular, consider an $\epsilon$-net cover of $\mathcal{X} \subset \mathbb{R}^d$. Then, we require $N_\epsilon = \left( \frac{4 \text{diam}(\mathcal{X})}{\epsilon} \right)^d$ balls of the radius $\epsilon > 0$, e.g., see [34, Lemma 4.1, Section 4]. Let $\mathcal{Z} = \{ z_1, \cdots, z_{N_\epsilon} \} \subset \mathcal{X}$ denotes the center of the covering net. Now, let
(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k) \in \mathbb{R}^{N \times D}$ and $(\tilde{\xi}_0^k, \ldots, \tilde{\xi}_{N-1}^k, \ldots, \tilde{\xi}_N^k) \in \mathbb{R}^{N \times D}$ be two sequences that differ in the $k$-th coordinate for $1 \leq k \leq N$. Then,

$$
|T_{(z_i, \lambda z_i)}(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k) - T_{(z_i, \lambda z_i)}(\tilde{\xi}_0^k, \ldots, \tilde{\xi}_{N-1}^k, \ldots, \tilde{\xi}_N^k)|
= \frac{1}{N} \left| M_{\lambda^{-1} \varphi(z_i) \varphi(z_j)}(\xi_0) - M_{\lambda^{-1} \varphi(z_i) \varphi(z_j)}(\tilde{\xi}_0) \right|
$$

(A.28)

Without loss of generality suppose $M_{\lambda^{-1} \varphi(z_i) \varphi(z_j)}(\xi_0) \geq M_{\lambda^{-1} \varphi(z_i) \varphi(z_j)}(\tilde{\xi}_0)$. Then,

$$
M_{\lambda^{-1} \varphi(z_i) \varphi(z_j)}(\xi_0) - M_{\lambda^{-1} \varphi(z_i) \varphi(z_j)}(\tilde{\xi}_0)
= \inf_{\zeta \in \mathbb{R}^D} \left\{ \lambda \| \zeta - \xi_0 \|^2 - \varphi(z_i) \varphi(z_j) \zeta \right\} - \inf_{\zeta \in \mathbb{R}^D} \left\{ \lambda \| \zeta - \tilde{\xi}_0 \|^2 - \varphi(z_i) \varphi(z_j) \zeta \right\}
\leq -\varphi(z_i; \xi_0^k) \varphi(z_j; \xi_0^k) - \inf_{\zeta \in \mathbb{R}^D} \left\{ \lambda \| \zeta - \tilde{\xi}_0 \|^2 - \varphi(z_i; \zeta) \varphi(z_j; \zeta) \right\}
\leq -\varphi(z_i; \xi_0^k) \varphi(z_j; \xi_0^k) + \sup_{\zeta \in \mathbb{R}^D} \{ \varphi(z_i; \zeta) \varphi(z_j; \zeta) \}
\leq 2L^2
$$

(A.29)  

(c)

where (a) follows by letting $\zeta = \xi_0^k$ in the first optimization problem, (b) follows by using the fact that $-\lambda \| \zeta - \tilde{\xi}_0 \|^2$ is non-positive for any $\zeta \in \mathbb{R}^D$ and can be dropped, and (c) follows from Assumption (A.2).

Now, plugging the upper bound in Eq. (A.29) into Eq. (A.28) yields

$$
|T_{(z_i, \lambda z_i)}(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k) - T_{(z_i, \lambda z_i)}(\tilde{\xi}_0^k, \ldots, \tilde{\xi}_{N-1}^k, \ldots, \tilde{\xi}_N^k)| \leq \frac{2L^2}{N}.
$$

From McDiarmid’s Martingale inequality [28] and the union bound, we obtain that

$$
\Pi \left( \bigcup_{z_i, z_j \in Z} |T_{(z_i, \lambda z_i)}(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k)| \geq 6 \right) \leq \left( \frac{4\text{diam}(\mathcal{X})}{\epsilon} \right)^d \exp \left( -\frac{N \delta^2}{L^2} \right),
$$

(A.30)

for all $\lambda \geq 0$. Now, consider arbitrary points $(\tilde{x}, \tilde{\alpha}) \in \mathcal{X} \times \mathcal{X}$. Let the center of the balls containing those points be $z_i, z_j \in Z$, i.e., $x \in \mathcal{B}_x(z_i)$ and $\tilde{x} \in \mathcal{B}_x(z_j)$ for some $z_i, z_j \in Z$.

Then, by the triangle inequality, we have that

$$
|T_{(\tilde{x}, \tilde{\alpha})}(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k) - T_{(z_i, \lambda z_i)}(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k)|
\leq |T_{(\tilde{x}, \tilde{\alpha})}(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k)| - |T_{(z_i, \lambda z_i)}(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k)|
+ |T_{(\tilde{x}, \tilde{\alpha})}(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k)|
\leq \| \nabla_\tilde{x} T_{(\tilde{x}, \tilde{\alpha})}(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k) \| \| \tilde{x} - z_i \| + \| \nabla_\tilde{\alpha} T_{(\tilde{x}, \tilde{\alpha})}(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k) \| \| \tilde{\alpha} - z_j \| + L_T \epsilon,
$$

(A.31)

where $L_T = L_T(\xi_0^k, \ldots, \xi_N^k) \defeq \sup_{x, \tilde{x} \in \mathcal{X}} \| \nabla_\tilde{x} T_{(\tilde{x}, \tilde{\alpha})}(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k) \|$ is the Lipschitz constant of the mapping $T$. Note that the Lipschitz constant $L_T$ is a random variable with respect to the random feature samples $\xi_0, \ldots, \xi_N$. Let $(x_*, \tilde{x}_*) \defeq \arg \sup_{x, \tilde{x} \in \mathcal{X}} \| \nabla_\tilde{x} T_{(\tilde{x}, \tilde{\alpha})}(\xi_0^k, \ldots, \xi_{N-1}^k, \ldots, \xi_N^k) \|_2$. 
We further proceed using the triangle inequality as well as the basic inequality \( (a_1 + a_2 + \cdots + a_N)^2 \leq N(a_1^2 + a_2^2 + \cdots + a_N^2) \),

\[
\mathbb{E}_{\mu_0} \left[ L_T^2 \right] = \mathbb{E}_{\mu_0} \left[ \left\| \nabla_x T(x_\star, x_\star) (\xi_0^1, \cdots, \xi_0^N) \right\|^2 \right] \\
= \mathbb{E}_{\mu_0} \left[ \left\| \frac{1}{N} \sum_{k=1}^N \nabla_x M^\lambda_{\varphi(x_\star)} \varphi(\hat{x}_\star; \cdot) (\xi_0^k) - \int_{\mathbb{R}_D} \nabla_x M^\lambda_{-\varphi(x_\star)} \varphi(\hat{x}_\star; \cdot) (\xi) \mu_0 (d\xi) \right\|^2 \right] \\
= \mathbb{E}_{\mu_0} \left[ \left\| \frac{1}{N} \sum_{k=1}^N \nabla_x M^\lambda_{\varphi(x_\star)} \varphi(\hat{x}_\star; \cdot) (\xi_0^k) \right\|^2 \right] - \mathbb{E}_{\mu_0} \left[ \left\| \int_{\mathbb{R}_D} \nabla_x M^\lambda_{-\varphi(x_\star)} \varphi(\hat{x}_\star; \cdot) (\xi) \mu_0 (d\xi) \right\|^2 \right] \\
\leq \frac{1}{N^2} \mathbb{E}_{\mu_0} \left[ \sum_{k=1}^N \left\| \nabla_x M^\lambda_{\varphi(x_\star)} \varphi(\hat{x}_\star; \cdot) (\xi_0^k) \right\|^2 \right].
\] (A.32)

We further proceed using the triangle inequality as well as the basic inequality \( (a_1 + a_2 + \cdots + a_N)^2 \leq N(a_1^2 + a_2^2 + \cdots + a_N^2) \),

\[
\mathbb{E}_{\mu_0} \left[ L_T^2 \right] = \frac{1}{N^2} \mathbb{E}_{\mu_0} \left[ \left\| \sum_{k=1}^N \nabla_x M^\lambda_{\varphi(x_\star)} \varphi(\hat{x}_\star; \cdot) (\xi_0^k) \right\|^2 \right] \\
\leq \frac{1}{N^2} \mathbb{E}_{\mu_0} \left[ \left\| \sum_{k=1}^N \nabla_x M^\lambda_{\varphi(x_\star)} \varphi(\hat{x}_\star; \cdot) (\xi_0^k) \right\|^2 \right] \\
\leq \frac{1}{N} \sum_{k=1}^N \mathbb{E}_{\mu_0} \left[ \left\| \nabla_x M^\lambda_{\varphi(x_\star)} \varphi(\hat{x}_\star; \cdot) (\xi_0^k) \right\|^2 \right].
\] (A.32)

To proceed from (A.32), we leverage the following lemma:

**Lemma A.8. (Moreau’s Envelope of Parametric Functions)** Consider the parametric function \( f : \mathcal{X} \times \Theta \to \mathbb{R} \) and the associated Moreau’s envelope for a given \( \theta \in \Theta \subset \mathbb{R}^d \):

\[
M^\beta_{f(\cdot; \theta)}(x) = \inf_{y \in \mathcal{X}} \left\{ \frac{1}{2\beta} \left\| x - y \right\|^2 + f(y; \theta) \right\}.
\] (A.33)

Furthermore, define the proximal operator as follows

\[
\text{Prox}^\beta_{f(\cdot; \theta)}(x) = \arg \inf_{y \in \mathcal{X}} \left\{ \frac{1}{2\beta} \left\| x - y \right\|^2 + f(y; \theta) \right\}.
\] (A.34)

Then, Moreau’s envelope has the following upper bound

\[
\left\| \nabla_\theta M^\beta_{f(\cdot; \theta)}(x) \right\|^2 \leq \left\| \nabla_\theta f \left( \text{Prox}^\beta_{f(\cdot; \theta)}(x); \theta \right) \right\|^2.
\] (A.35)

The proof is presented in Appendix B.2.

Equipped with Inequality (A.35) of Lemma A.8, we now compute an upper bound on the right hand side of Eq. (A.32) as follows

\[
\mathbb{E}_{\mu_0} \left[ L_T^2 \right] \leq \frac{1}{N} \sum_{k=1}^N \mathbb{E}_{\mu_0} \left[ \left\| \nabla_x \varphi(x_\star; \xi_0^k) \right\|^2 \right] \leq L^4,
\] (A.36)

where the last inequality is due to (A.2).
Invoking Markov’s inequality now yields
\[
\mathbb{P}\left( \left| T_{(x, \hat{x})}^\lambda (\xi_0^1, \ldots, \xi_0^N) - T_{(x, z)}^\lambda (\xi_0^1, \ldots, \xi_0^N) \right| \geq \delta \right) = \mathbb{P}\left( L_T \geq \frac{\delta}{2\epsilon} \right) \\
\leq \left( \frac{2\epsilon}{\delta} \right)^2 \mathbb{E}_{\nu_0}[L_T^2] \\
\leq \left( \frac{2\epsilon}{\delta} \right)^2 L^4.
\]

(A.37)

Now, using the union bound, for every arbitrary pair of data points \((x, \hat{x}) \in \mathcal{X} \times \mathcal{X}\) the following inequality holds
\[
\mathbb{P}\left( \left| T_{(x, \hat{x})}^\lambda (\xi_0^1, \ldots, \xi_0^N) \right| \geq \delta \right) \leq \mathbb{P}\left( \left| T_{(x, z)}^\lambda (\xi_0^1, \ldots, \xi_0^N) \right| \geq \delta/2 \right) \\
+ \mathbb{P}\left( \left| T_{(x, \hat{x})}^\lambda (\xi_0^1, \ldots, \xi_0^N) - T_{(x, z)}^\lambda (\xi_0^1, \ldots, \xi_0^N) \right| \geq \delta/2 \right) \\
\leq \left( \frac{2\epsilon}{\delta} \right)^2 \left( 4diam(\mathcal{X})^d \right) \cdot \exp\left( \frac{-N\delta^2}{L^2} \right),
\]

(A.38)

Following the proposal of [36], we choose \(\epsilon = (\kappa_1/\kappa_2)^{1/2} \) and \(\kappa_1 \overset{\text{def}}{=} (4diam(\mathcal{X}))^d \cdot e^{-2\kappa_2 M^2} \) and \(\kappa_2 \overset{\text{def}}{=} (2/\delta)^2 L^4 \). Then,
\[
\mathbb{P}\left( \sup_{x, \hat{x} \in \mathcal{X}} \left| T_{(x, \hat{x})}^\lambda (\xi_0^1, \ldots, \xi_0^N) \right| \geq \delta \right) \leq 2^8 \left( \frac{L^2diam(\mathcal{X})}{\delta} \right)^2 \cdot \exp\left( \frac{-N\delta^2}{L^2(d + 2)} \right).
\]

Thus, with the probability of at least \(1 - \varrho\), the following inequality holds
\[
\sup_{\lambda \geq 0} \sup_{x, \hat{x} \in \mathcal{X}} \left| T_{(x, \hat{x})}^\lambda (\xi_0^1, \ldots, \xi_0^N) \right| \leq \left( \frac{L^2(d + 2)}{N} \right)^{1/2} W\left( \frac{2^8N\text{diam}^2(\mathcal{X})}{\varrho} \right),
\]
where \(W(\cdot)\) is the Lambert \(W\)-function.\(^5\) Since \(W(x) \leq \ln(x)\) for \(x > e\), we can rewrite the upper bound in terms of elementary functions
\[
\sup_{\lambda \geq 0} \sup_{x, \hat{x} \in \mathcal{X}} \left| T_{(x, \hat{x})}^\lambda (\xi_0^1, \ldots, \xi_0^N) \right| \leq \sqrt{\frac{L^2(d + 2)}{N} \ln^2 \left( \frac{2^8N\text{diam}^2(\mathcal{X})}{\varrho} \right)},
\]
provided that \(N\) is sufficiently large and/or \(\varrho\) is sufficiently small so that \(2^8N\text{diam}^2(\mathcal{X}) \geq e\).

Plugging Inequality (A.40) in Eq. (A.27) now results in the following inequality
\[
\left| \sup_{\mu \in P} \overrightarrow{\text{MMD}}_{\mu}[P_V, P_W] - \sup_{\hat{\mu}^N \in P_N} \overrightarrow{\text{MMD}}_{\hat{\mu}^N}[P_V, P_W] \right| \leq \sqrt{\frac{L^2(d + 2)}{N} \ln^2 \left( \frac{2^8N\text{diam}^2(\mathcal{X})}{\varrho} \right)},
\]
for all \(W \in W\). Employing (A.23) from Lemma A.7 now yields the following upper bound
\[
A_2 = \left| \min_{W \in W} \sup_{\mu \in P} \overrightarrow{\text{MMD}}_{\mu}[P_V, P_W] - \min_{W \in \hat{W}} \sup_{\hat{\mu}^N \in P_N} \overrightarrow{\text{MMD}}_{\hat{\mu}^N}[P_V, P_W] \right| \\
\leq \sqrt{\frac{L^2(d + 2)}{N} \ln^2 \left( \frac{2^8N\text{diam}^2(\mathcal{X})}{\varrho} \right)}.
\]

(A.42)

Upper bound on \(A_3\):

\(^5\)Recall that the lambert \(W\)-function is the inverse of the function \(f(W) = We^W\).
Recall that the solution of the empirical risk function of Eq. (3.7) is denoted by
\[
(\hat{W}^N_\ast, \hat{\mu}^N_\ast) \overset{\text{def}}{=} \arg\min_{\hat{W} \in W} \arg\min_{\hat{\mu}^N \in P_N} \inf_{\hat{\mu}^N \in P_N} \hat{\text{MMD}}^\alpha_{\hat{\mu}^N} [P_V, P_W]
\]
\[
= \arg\min_{\hat{W} \in W} \arg\sup_{\hat{\mu}^N \in P_N} \frac{8}{n(n-1)} \sum_{1 \leq i < j \leq n} y_i y_j E_{\hat{\mu}^N} [\varphi(x_i; \xi) \varphi(x_j; \xi)]
\]
\[
- \frac{8}{n(n-1)\alpha} \sum_{1 \leq i < j \leq n} (E_{\hat{\mu}^N} [\varphi(x_i; \xi) \varphi(x_j; \xi)])^2.
\]
(A.43)

We also define the solution of the empirical kernel alignment as follows
\[
(\hat{W}^N_\ast, \hat{\mu}^N_\ast) \overset{\text{def}}{=} \arg\min_{\hat{W} \in W} \arg\sup_{\hat{\mu}^N \in P_N} \hat{\text{MMD}}^\alpha_{\hat{\mu}^N} [P_V, P_W]
\]
\[
= \frac{8}{n(n-1)} \sum_{1 \leq i < j \leq n} y_i y_j E_{\hat{\mu}^N} [\varphi(x_i; \xi) \varphi(x_j; \xi)].
\]
(A.44)

Due to the optimality of the empirical measure \(\hat{\mu}^N_\ast\) for the inner optimization in Eq. (A.43), the following inequality holds
\[
\hat{\text{MMD}}^\alpha_{\hat{\mu}^N_\ast} [P_V, P_{\hat{W}^N_\ast}] \leq \hat{\text{MMD}}^\alpha_{\hat{\mu}^N_\ast} [P_V, P_{\hat{W}^N_\ast}]
\]
\[
\leq \frac{8}{n(n-1)} \sum_{1 \leq i < j \leq n} y_i y_j E_{\hat{\mu}^N_\ast} [\varphi(x_i; \xi) \varphi(x_j; \xi)].
\]
(A.45)

Upon expansion of \(\hat{\text{MMD}}^\alpha_{\hat{\mu}^N_\ast} [P_V, P_{\hat{W}^N_\ast}]\), and after rearranging the terms in Eq. (A.45), we arrive at
\[
\hat{\text{MMD}}^\alpha_{\hat{\mu}^N_\ast} [P_V, P_{\hat{W}^N_\ast}] - \hat{\text{MMD}}^\alpha_{\hat{\mu}^N_\ast} [P_V, P_{\hat{W}^N_\ast}]
\]
\[
= \frac{8}{n(n-1)} \sum_{1 \leq i < j \leq n} y_i y_j (E_{\hat{\mu}^N_\ast} [\varphi(x_i; \xi) \varphi(x_j; \xi)] - E_{\hat{\mu}^N_\ast} [\varphi(x_i; \xi) \varphi(x_j; \xi)])
\]
\[
\leq \frac{8}{n(n-1)\alpha} \sum_{1 \leq i < j \leq n} (E_{\hat{\mu}^N_\ast} [\varphi(x_i; \xi) \varphi(x_j; \xi)])^2
\]
\[
\leq \frac{8L^4}{\alpha},
\]
(A.46)

where the last inequality is due to the fact that \(|\varphi|_\infty < L\) by (A.1). Now, due to optimality of \(\hat{W}^N_\ast\) for the outer optimization problem in Eq. (A.44), we have
\[
\hat{\text{MMD}}^\alpha_{\hat{\mu}^N_\ast} [P_V, P_{\hat{W}^N_\ast}] \leq \hat{\text{MMD}}^\alpha_{\hat{\mu}^N_\ast} [P_V, P_{\hat{W}^N_\ast}].
\]
(A.47)

Putting together Inequalities (A.46) and (A.47) yields
\[
\hat{\text{MMD}}^\alpha_{\hat{\mu}^N_\ast} [P_V, P_{\hat{W}^N_\ast}] - \hat{\text{MMD}}^\alpha_{\hat{\mu}^N_\ast} [P_V, P_{\hat{W}^N_\ast}] \leq \frac{8L^4}{\alpha}.
\]
(A.48)

Similarly, due to the optimality of the empirical measure \(\hat{\mu}^N_\ast\) for the optimization in Eq. (A.44) we have that
\[
\hat{\text{MMD}}^\alpha_{\hat{\mu}^N_\ast} [P_V, P_{\hat{W}^N_\ast}] \leq \hat{\text{MMD}}^\alpha_{\hat{\mu}^N_\ast} [P_V, P_{\hat{W}^N_\ast}]
\]
\[
\leq \hat{\text{MMD}}^\alpha_{\hat{\mu}^N_\ast} [P_V, P_{\hat{W}^N_\ast}].
\]
(A.49)
Combining Eqs. (A.48) and (A.49) then yields

\[(A.50) \quad A_3 = \left| \tilde{\text{MMD}}_{\tilde{\mu}_N} \left[ P_V, P_{\tilde{W}_N} \right] - \text{MMD}_{\tilde{\mu}_N} \left[ P_V, P_{\tilde{W}_N} \right] \right| \leq \frac{8L^4}{\alpha}. \]

**Upper bound on \( A_4 \):**

The upper bound on \( A_4 \) can be obtained exactly the same way as \( A_1 \). Indeed, from Eq. (A.22) it follows directly that

\[(A.51) \quad A_4 = \left| \tilde{\text{MMD}}_{\tilde{\mu}_N} \left[ P_V, P_{\tilde{W}_N} \right] - \text{MMD}_{\tilde{\mu}_N} \left[ P_V, P_{\tilde{W}_N} \right] \right| \leq \sup_{\tilde{\mu}_N \in P_N} \left| \tilde{\text{MMD}}_{\tilde{\mu}_N} \left[ P_V, P_{\tilde{W}_N} \right] - \text{MMD}_{\tilde{\mu}_N} \left[ P_V, P_{\tilde{W}_N} \right] \right| \leq \min \left\{ \frac{3^{\frac{3}{2}} \times 2^{\frac{1}{2}} \times L^2}{n}, \frac{9 \times 2^{\frac{1}{2}} \times RL^4}{n^2} \right\}. \]

Now, plugging the derived upper bounds in \( A_1 \) through \( A_4 \) in Eq. (A.3) and employing the union bound completes the proof.

---

**A.2. Proof of Theorem 4.2.** The proof has three main ingredients and follows the standard procedure in the literature, see, e.g., [50, 27]. In the first step, we identify the mean-field limit of the particle SGD in Eq. (3.8). In the second step, we prove the convergence of the measured-valued process \( \{ (\mu_N^t)_{0 \leq t \leq T} \} \) to the mean-field solution by establishing the pre-compactness of Sokhorhod space. Lastly, we prove the uniqueness of the mean-field solution of the particle SGD.

**Step 1-Identification of the scaling limit:** First, we identify the weak limit of converging sub-sequences via the action of the empirical measure \( \tilde{\mu}_m^N (\xi) = \frac{1}{N} \sum_{k=1}^{N} \delta (\xi - \xi_m^k) \) on a test function \( f \in C_b^3 (\mathbb{R}^d) \). In particular, we use the standard techniques of computing the scaling limits from [27].

Recall that the action of an empirical measure on a bounded function is defined as follows

\[(A.53) \quad \langle f, \tilde{\mu}_m^N \rangle \overset{\text{def}}{=} \frac{1}{N} \sum_{k=1}^{N} f (\xi_m^k). \]

We analyze the evolution of the empirical measure \( \tilde{\mu}_m^N \) via its action on a test function \( f \in C_b^3 (\mathbb{R}^d) \). Using Taylor’s expansion, we obtain

\[
\langle f, \tilde{\mu}_m^{N+1} \rangle - \langle f, \tilde{\mu}_m^{N} \rangle = \langle f, \tilde{\mu}_m^{N+1} \rangle - \langle f, \tilde{\nu}_m^{N+1} \rangle \\
= \frac{1}{N} \sum_{k=1}^{N} f (\xi_m^{k+1}) - f (\xi_m^k) \\
= \frac{1}{N} \sum_{k=1}^{N} \nabla f (\xi_m^k) (\xi_m^{k+1} - \xi_m^k)^T + R_m^N,
\]

where \( R_m^N \) is a remainder term defined as follows

\[(A.54) \quad R_m^N \overset{\text{def}}{=} \frac{1}{N} \sum_{k=1}^{N} (\xi_m^{k+1} - \xi_m^k)^T \nabla^2 f (\tilde{\xi}_m^k) (\xi_m^{k+1} - \xi_m^k), \]

xxix
where $\xi_k \triangleq (\xi^k(1), \ldots, \xi^k(p))$, and $\tilde{\xi}^k(i) \in [x^k(i), x^k_{m+1}(i)]$, for $i = 1, 2, \ldots, p$.

Plugging the difference term $(\xi_{m+1}^k - \xi_m^k)$ from the SGD equation in Eq. (3.8) results in

\begin{align*}
(D.55) \\
(f, \widehat{\mu}_{m+1}^N) - (f, \widehat{\mu}_m^N) \\
= \frac{\eta}{N^2\alpha} \sum_{k=1}^{N} \triangledown f(x_m^k) \cdot \left( \left( \frac{1}{N} \sum_{t=1}^{N} \phi(x_m^k; \xi_m^k) - \alpha y_m \bar{y}_m \right) \triangledown \phi \left( \phi(x_m^k; \xi_m^k), \psi(x_m^k; \xi_m^k) \right) \right) + R_m^N.
\end{align*}

Now, we define the drift and Martingale terms as follows

\begin{align*}
(\text{A.56a}) & \quad D_m^N \overset{\text{def}}{=} \frac{\eta}{N^2\alpha} \int_{X \times Y} \left( \langle \phi(x, \xi) \phi(\bar{x}, \xi), \mu_m \rangle - \alpha y_m \bar{y}_m \right) \\
& \quad \times \langle \triangledown f(x) \phi(\bar{x}, \xi) \triangledown \phi(x, \xi) + \phi(x, \xi) \triangledown \phi(\bar{x}, \xi) \rangle dP_{x, y} \left( (x, y), (\bar{x}, \bar{y}) \right) \\
(\text{A.56b}) & \quad M_m^N \overset{\text{def}}{=} \frac{\eta}{N^2\alpha} \left( \langle \phi(x, \xi) \phi(\bar{x}, \xi), \mu_m \rangle - \alpha y_m \bar{y}_m \right) \\
& \quad \times \langle \triangledown f(x) \phi(\bar{x}, \xi) \triangledown \phi(x, \xi) + \phi(x, \xi) \triangledown \phi(\bar{x}, \xi) \rangle dP_{x, y} \left( (x, y), (\bar{x}, \bar{y}) \right).
\end{align*}

respectively. Using the definitions of $D_m^N$ and $M_m^N$ in Eqs. (A.56a)-(A.56b), we recast Equation (A.55) as follows

\begin{align*}
(\text{A.57}) & \quad (f, \widehat{\mu}_{m+1}^N) - (f, \widehat{\mu}_m^N) = D_m^N + M_m^N + R_m^N.
\end{align*}

Summation over $\ell = 0, 1, 2, \ldots, m - 1$ and using the telescopic sum yields

\begin{align*}
(\text{A.58}) & \quad (f, \widehat{\mu}_{m+1}^N) - (f, \widehat{\mu}_m^N) = \sum_{\ell=0}^{m-1} D_{\ell}^N + \sum_{\ell=0}^{m-1} M_{\ell}^N + \sum_{\ell=0}^{m-1} R_{\ell}^N.
\end{align*}

We also define the following continuous embedding of the drift, martingale, and the remainder terms as follows

\begin{align*}
(\text{A.59a}) & \quad D_t^N \overset{\text{def}}{=} \sum_{\ell=0}^{\lfloor Nt \rfloor} D_{\ell}^N \\
(\text{A.59b}) & \quad M_t^N \overset{\text{def}}{=} \sum_{\ell=0}^{\lfloor Nt \rfloor} M_{\ell}^N \\
(\text{A.59c}) & \quad R_t^N \overset{\text{def}}{=} \sum_{\ell=0}^{\lfloor Nt \rfloor} R_{\ell}^N, \quad t \in (0, T].
\end{align*}

The scaled empirical measure $\mu_t^N \overset{\text{def}}{=} \hat{\mu}_{\lfloor Nt \rfloor}^N$ then can be written as follows

\begin{align*}
(\text{A.60}) & \quad (f, \mu_t^N) - (f, \mu_0^N) = D_t^N + M_t^N + R_t^N.
\end{align*}

Since the drift process $(D_t^N)_{0 \leq t \leq T}$ is a piecewise c"adl"ag process, we have

\begin{align*}
(\text{A.61}) & \quad D_t^N = \int_0^t \int_{\mathbb{R}} R[\mu_s] ds,
\end{align*}
where the functional $R[\mu_s]$ is defined as follows
\begin{equation}
R[\mu_s] \overset{\text{def}}{=} \frac{\eta}{\alpha} \int \int_{X \times Y} \left( (\varphi(x, \xi) \varphi(\bar{x}, \xi), \mu_s) - \alpha y \bar{y} \right) \times \langle \nabla f(\xi)(\varphi(\bar{x}; \xi) \nabla_{\xi} \varphi(x; \xi) + \varphi(x; \xi) \nabla_{\xi} \varphi(\bar{x}; \xi)) \rangle, \mu_s \rangle P_{x,y}^{\otimes 2}((dx, d\bar{x}), (dy, d\bar{y})).
\end{equation}

Therefore, the expression in Eq. (A.60) can be rewritten as follows
\begin{equation}
\langle f, \mu_t^N \rangle - \langle f, \mu_0^N \rangle = \int_0^T R[\mu_s] ds + \mathcal{M}_t^N + \mathcal{R}_t^N.
\end{equation}

In the following lemma, we prove that the remainder term $\sup_{0 \leq t \leq T} |\mathcal{R}_t^N|$ vanishes in probabilistic sense as the number of particles tends to infinity $N \to \infty$:

**Lemma A.9.** (Large $N$-Limit of the Remainder Process) Consider the remainder process $(\mathcal{R}_t^N)_{0 \leq t \leq T}$ defined via scaling in Eqs. (A.54)-(A.59c). Then, there exists a constant $C_0 > 0$ such that
\begin{equation}
\sup_{0 \leq t \leq T} |\mathcal{R}_t^N| \leq \frac{C_0 T}{N} \left( \eta L^2 + \frac{2\eta L^4}{\alpha} \right).
\end{equation}
and thus $\lim \sup_{N \to \infty} \sup_{0 \leq t \leq T} |\mathcal{R}_t^N| = 0$ almost surely.

**Proof.** The proof is relegated to Appendix B.5. \hfill \square

We can also prove a similar result for the process defined by the remainder term:

**Lemma A.10.** (Large $N$-Limit of the Martingale Process) Consider the Martingale process $(\mathcal{M}_t^N)_{0 \leq t \leq T}$ defined via scaling in Eqs. (A.56b)-(A.59b). Then, for some constant $C_1 > 0$, the following inequality holds
\begin{equation}
\mathbb{P} \left( \sup_{0 \leq t \leq T} |\mathcal{M}_t^N| \geq \varepsilon \right) \leq \frac{1}{N \alpha \varepsilon} 4\sqrt{2L^2 \sqrt{[NT]} \eta C_1 (L^2 + \alpha)^2}.
\end{equation}

In particular, with the probability of at least $1 - \rho$, we have
\begin{equation}
\sup_{0 \leq t \leq T} |\mathcal{M}_t^N| \leq \frac{1}{N \alpha \rho} 4\sqrt{2L^2 \sqrt{[NT]} \eta C_1 (L^2 + \alpha)^2}.
\end{equation}
and thus $\lim \sup_{N \to \infty} \sup_{0 \leq t \leq T} |\mathcal{M}_t^N| = 0$ almost surely.

**Proof.** The proof is deferred to Appendix B.6. \hfill \square

Now, using the results of Lemmata A.10 and A.9 in conjunction with Eq. (A.63) yields the following mean-field equation as $N \to \infty$,
\begin{equation}
\langle \mu_t, f \rangle = \langle \mu_0, f \rangle + \frac{\eta}{\alpha} \int_0^T \int \int_{X \times Y} \left( (\varphi(x, \xi) \varphi(\bar{x}, \xi), \mu_s) - \alpha y \bar{y} \right) \times \langle \nabla f(\xi)(\varphi(\bar{x}; \xi) \nabla_{\xi} \varphi(x; \xi) + \varphi(x; \xi) \nabla_{\xi} \varphi(\bar{x}; \xi)) \rangle, \mu_s \rangle P_{x,y}^{\otimes 2}((dx, d\bar{x}), (dy, d\bar{y})).
\end{equation}

Notice that he mean-field equation in Eq. (A.67) is in the weak form. When the Lebesgue density $p_t(\xi) = d\mu_t/d\xi$ exists, the McKean-Vlasov PDE in Eq. (4.5) can be readily obtained from Eq. (A.67).
Step 2: Pre-compactness of the Skorkhod space: To establish our results in this part of the proof, we need a definition and a theorem:

**Definition A.11. (Tightness)** A set $\mathcal{A}$ of probability measures on a metric space $S$ is tight if there exists a compact subset $S_0 \subset S$ such that

$$\nu(S_0) \geq 1 - \varepsilon, \quad \text{for all } \nu \in \mathcal{A},$$

(A.68) for all $\varepsilon > 0$. A sequence $\{X^N\}_{N \in \mathbb{N}}$ of random elements of the metric space $S$ is tight if there exists a compact subset $S_0 \subset S$ such that

$$\nu(X^N \in S_0) > 1 - \varepsilon,$$

(A.69) for all $\varepsilon > 0$, and all $N \in \mathbb{N}$.

Now, to show the tightness of the measured valued process $(\mu^N_t)_{0 \leq t \leq T}$, we must verify Jakubowski’s criterion [16, Thm. 1]:

**Theorem A.12. (Jakubowski’s criterion [16, Thm. 1])** A sequence of measured-valued process $\{(\zeta^N_t)_{0 \leq t \leq T}\}_{N \in \mathbb{N}}$ is tight in $\mathcal{D}(\mathbb{R}^p([0,T]))$ if and only if the following two conditions are satisfied:

- **(J1)** For each $T > 0$ and $\gamma > 0$, there exists a compact set $\mathcal{U}_{T,\gamma}$ such that
  $$\lim_{N \to \infty} \inf \mathbb{P} \left( \zeta^N_t \in \mathcal{U}_{T,\gamma}, \forall t \in (0,T) \right) > 1 - \gamma.$$  

This condition is referred to as the compact-containment condition.

- **(J2)** There exists a family $\mathcal{H}$ of real-valued functions $H : \mathcal{M}(\mathbb{R}^p) \mapsto \mathbb{R}$ that separates points in $\mathcal{M}(\mathbb{R}^p)$ and is closed under addition such that for every $H \in \mathcal{H}$, the sequence $\{(H(\zeta^N_t))_{0 \leq t \leq T}\}_{N \in \mathbb{N}}$ is tight in $\mathcal{D}(\mathbb{R})([0,T])$.

To establish (J1), we closely follow the proof of [11, Lemma 6.1.]. In particular, for each $L > 0$, we define $S_L = [0,B]^p$. Then, $S_B \subset \mathbb{R}^p$ is compact, and for each $t \geq 0$, and $N \in \mathbb{N}$, we have

$$\mathbb{E}[\mu^N_t(\mathbb{R}^p/S_B)] = \frac{1}{N} \sum_{k=1}^{N} \mathbb{P}(\|\xi^k_{[Nt]}\|_2 \geq B)$$

(A.71)

$$\leq \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}[\|\xi^k_{[Nt]}\|_2]$$

(a)

$$\leq \frac{c_0 + \eta L^2 T + 2\eta L^4 T}{B},$$

(b)

(A.73)

where (a) follows from Markov’s inequality, and (b) follows from the upper bound on the norm of the particles in Eq. (B.22) of Appendix B. We now define the following set

$$\mathcal{U}_B = \left\{ \mu \in \mathcal{M}(\mathbb{R}^p) : \mu(\mathbb{R}^p/\mathcal{S}(B+j)^2) < \frac{1}{\sqrt{B+j}} \text{ for all } j \in \mathbb{N} \right\}.$$  

(A.74)
We let $\mathcal{U}_{T, \gamma} = \overline{\mathcal{U}}_B$, where $\overline{\mathcal{U}}_B$ is the completion of the set $\mathcal{U}_B$. By definition, $\mathcal{U}_{T, \gamma}$ is a compact subset of $\mathcal{M}(\mathbb{R}^p)$. Now, we have

$$\mathbb{P}\left(\mu^N_t \notin \mathcal{U}_{T, \gamma}\right) \leq \sum_{j=1}^{\infty} \mathbb{P}\left(\mu^N_t(\mathbb{R}^p / S(B + j^2)) > 1/ \sqrt{B + j}\right)$$

$$\leq \sum_{j=1}^{\infty} \mathbb{E}[\mu^N_t(\mathbb{R}^p / S(B + j^2))] \left(1/ \sqrt{B + j}\right)$$

$$\leq \sum_{j=1}^{\infty} \frac{c_0 + \eta L^2 T + 2(\eta/\alpha)L^4 T}{(B + j)^2 / \sqrt{B + j}}$$

$$= \sum_{j=1}^{\infty} \frac{c_0 + \eta L^2 T + 2(\eta/\alpha)L^4 T}{(B + j)^{3/2}}.$$  
(A.75)

Now, since

$$\lim_{B \to \infty} \sum_{j=1}^{\infty} \frac{c_0 + \eta L^2 T + 2(\eta/\alpha)L^4 T}{(B + j)^{3/2}} = 0,$$  
(A.76)

this implies that for any $\gamma > 0$, there exists a $B > 0$, such that

$$\lim_{N \to \infty} \inf \mathbb{P}(\mu^N_t \in \overline{\mathcal{U}}_B, \forall t \in (0, T]) > 1 - \gamma.$$  
(A.77)

This completes the proof of (J.1). To verify (J.2), we consider the following class of functions

$$\mathcal{H} \triangleq \{ H : \exists f \in C^3_b(\mathbb{R}^p) \text{ such that } H(\mu) = \langle \mu, f \rangle, \forall \mu \in \mathcal{M}(\mathbb{R}^p) \}.$$  
(A.78)

By definition, every function $H \in \mathcal{H}$ is continuous with respect to the weak topology of $\mathcal{M}(\mathbb{R}^p)$ and further the class of functions $\mathcal{H}$ separate points in $\mathcal{M}(\mathbb{R}^p)$ and is closed under addition. Now, we state the following sufficient conditions to establish (J.2). The statement of the theorem is due to [38, Thm. C.9]:

**Theorem A.13.** ( Tightness in $\mathcal{D}_\mathbb{R}([0, T])$, [38, Thm. C.9]) A sequence $\{(Z^N_t)_{0 \leq t \leq T}\}_{N \in \mathbb{N}}$ is tight in $\mathcal{D}_\mathbb{R}([0, T])$ iff for any $\delta > 0$, we have

(T.1) There exists $\epsilon > 0$, such that

$$\mathbb{P}(|Z^N_0| > \epsilon) \leq \delta,$$  
(A.79)

for all $N \in \mathbb{N}$.

(T.2) For any $\rho > 0$, there exists $\sigma > 0$ such that

$$\mathbb{P}\left(\sup_{t_1, t_2 \leq T, |t_1 - t_2| \leq \rho} |Z^N_{t_1} - Z^N_{t_2}| > \sigma\right) \leq \delta,$$  
(A.80)

This completes the tightness proof of the of the laws of the measured-valued process $\{(\mu^N_t)_{0 \leq t \leq T}\}_{N \in \mathbb{N}}$. Now, we verify the condition (J.2) by showing that the sufficient conditions (T.1) and (T.2) hold for function values $\{(H(\mu^N_t))_{0 \leq t \leq T}\}_{N \in \mathbb{N}}$, where $H \in \mathcal{H}$ and
\( \mathcal{H} \) is defined in Eq. (A.78). Now, condition (T.1) is readily verified since

\begin{equation}
H(\mu_0^N) = \langle \mu_0^N, f \rangle = \int_{\mathbb{R}^p} f(\xi) \mu_0^N(d\xi) 
\end{equation}

\begin{equation}
\leq \|f\|_\infty \int_{\mathbb{R}^p} \mu_0^N(d\xi)
\end{equation}

\begin{equation}
\leq b,
\end{equation}

where in the last step, we used the fact that \( f \in C_b^3(\mathbb{R}^p) \), and hence, \( \|f\|_\infty \leq b \). Thus, \( \mathbb{P}(H(\mu_0^N) \geq b) = 0 \) for all \( N \in \mathbb{N} \), and the condition (T.1) is satisfied. Now, consider the condition (T.2). From Equation (A.63), and with \( 0 \leq t_1 < t_2 \leq T \) we have

\begin{equation}
|H(\mu_{t_1}^N) - H(\mu_{t_2}^N)| = |\langle f, \mu_{t_1}^N \rangle - \langle f, \mu_{t_2}^N \rangle| 
\end{equation}

\begin{equation}
\leq \int_{t_1}^{t_2} |R[\mu_s]| ds + |\mathcal{M}_{t_1}^N - \mathcal{M}_{t_2}^N| + |\mathcal{R}_{t_1}^N - \mathcal{R}_{t_2}^N|.
\end{equation}

To bound the first term, recall the definition of \( R[\mu_s] \) from Eq. (A.62). The following chain of inequalities holds,

\begin{equation}
|R[\mu_s]| \leq \frac{\eta}{\alpha} \mathbb{E}_f \mathbb{E}_{x,y} [\|\xi\|_\infty \cdot |\langle I, \mu_s \rangle| + \alpha \|f(\xi)\|_\infty \cdot |\langle f, \mu_s \rangle|]
\end{equation}

\begin{equation}
\leq \frac{\eta}{\alpha} \mathbb{E}_f \mathbb{E}_{x,y} [\|\xi\|_\infty + \alpha] \cdot \|f(\xi)\|_\infty \cdot |\langle f, \mu_s \rangle|]
\end{equation}

\begin{equation}
\leq \frac{2\eta}{\alpha} (L^2 + \alpha) L^2 C_1,
\end{equation}

where the last inequality is due to (A.1). Therefore,

\begin{equation}
\int_{t_1}^{t_2} |R[\mu_s]| ds \leq s_0 |t_2 - t_1|,
\end{equation}

where \( s_0 \defeq \frac{2\eta}{\alpha} (L^2 + \alpha) L^2 C_1 \).

Consider the middle term of (A.84). Using the definition of the martingale term in Eq. (A.59b), we obtain that

\begin{equation}
|\mathcal{M}_{t_1}^N - \mathcal{M}_{t_2}^N| = \left| \sum_{l=0}^{N_{t_1}} M_l^N - \sum_{l=0}^{N_{t_2}} M_l^N \right|
\end{equation}

\begin{equation}
\leq \left| \sum_{l=0}^{N_{t_2}} M_l^N \right|
\end{equation}

In Equation of Section B, we have proved the following concentration bound

\begin{equation}
\mathbb{P}(|\mathcal{M}_m^N| \geq \varepsilon) \leq 2 \exp \left( - \frac{N^2 \alpha^2 \varepsilon^2}{8m L^4 \eta^2 C_1^2 (L^2 + \alpha)^2} \right), \quad \forall m \in [0, NT] \cap \mathbb{N}.
\end{equation}
Now, recall the alternative definition of the sub-Gaussian random variables:

**Definition A.14. (Sub-Gaussian Random Variables [4])** A random variable $X$ is $\sigma^2$-sub-Gaussian if

$$
\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right).
$$

We enumerate a few standard consequences of sub-Gaussianity [4]. If $X_i$ are independent and $\sigma_i^2$-sub-Gaussian, then $\sum_{i=1}^n X_i$ is $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian. Moreover, $X$ is $\sigma^2$-sub-Gaussian if and only if

$$
\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right).
$$

Now, it is clear from (A.89) and that $M^N_m$ is sub-Gaussian random variable with a zero mean, and with the parameter $\sigma^2_m = \frac{4mL^2\eta^2C_1^2(L^2 + \alpha)^2}{N^2G^2}$. Therefore, $\sum_{\ell=|Nt_1|}^{[Nt_2]} M^N_{\ell}$ is sub-Gaussian with the parameter $\sigma^2(t_1, t_2) \overset{\text{def}}{=} 2L^4\eta^2C_1^2(L^2 + \alpha)^2\frac{([Nt_1] - [Nt_2] + 1)\left([Nt_1] + [Nt_2]\right)}{N^2G^2}$. Consequently, from Inequality (A.88) and the concentration inequality in Eq. (A.91), we have

$$
\mathbb{P}\left(\sup_{t_1, t_2 \leq T, |t_1 - t_2| \leq \rho} |M^N_{t_1} - M^N_{t_2}| \geq \varepsilon\right) \leq 2 \mathbb{P}\left(\left|\sum_{\ell=|Nt_1|}^{[Nt_2]} M^N_{\ell}\right| \geq \varepsilon\right)
$$

$$
= 2 \exp\left(-\frac{\varepsilon^2}{2\sigma^2(t_1, t_2)}\right)
$$

where $(t_1^*, t_2^*) \overset{\text{def}}{=} \arg\sup_{t_1, t_2 \leq T, |t_1 - t_2| \leq \rho} \left|\sum_{\ell=|Nt_1|}^{[Nt_2]} M^N_{\ell}\right|$.

We first compute a bound for the last term of (A.84) using the definition of the scaled term $\mathcal{R}^N_t$ from (A.59c). We have

$$
|\mathcal{R}^N_{t_1} - \mathcal{R}^N_{t_2}| = \left|\sum_{\ell=0}^{[Nt_1]} R^N_{\ell} - \sum_{\ell=0}^{[Nt_2]} R^N_{\ell}\right|
$$

$$
= \left|\sum_{\ell=|Nt_1|}^{[Nt_2]} R^N_{\ell}\right|
$$

$$
\leq \sum_{\ell=|Nt_1|}^{[Nt_2]} |R^N_{\ell}|
$$

$$
\overset{(a)}{\leq} \left|[Nt_2] - [Nt_1]\right| \frac{C_0}{N^2} (\eta L^2 + (L^4/\alpha))
$$

$$
\overset{(b)}{\leq} s |t_2 - t_1|.
$$
where (a) follows from the upper bound in Eq. (B.23) of Section B, and in (b) we define
\[ s_1 = \frac{C_0}{N}(\eta L^2 + (L^4/\alpha)). \]

Putting together (A.87), (A.92), and (A.96), we conclude from Inequality (A.84) that
\[
\mathbb{P}\left( \sup_{t_1, t_2 \leq T, |t_1 - t_2| \leq \rho} |H(\mu^N_{t_1}) - H(\mu^N_{t_2})| \geq \sigma \right) \leq \mathbb{P}\left( \sup_{t_1, t_2 \leq T, |t_1 - t_2| \leq \rho} |\mathcal{M}^N_{t_1} - \mathcal{M}^N_{t_2}| + (s_0 + s_1)\rho \geq \sigma \right) 
\leq 2 \exp\left( -\frac{\alpha^2(\sigma - (s_0 + s_1)\rho)^2}{4L^3\eta^2C_1^2(L^2 + \alpha)^2(\rho + 1)T} \right).
\]

Therefore, condition (T.2) is also satisfied. Since the sufficient conditions (T.1) and (T.2) are satisfied, the condition (J.2) is satisfied. This completes the tightness proof of the measured-valued sequence \( \{\mu^N_t\}_{N \in \mathbb{N}} \).

Now, we prove its convergence to a mean-field solution \( \mu^*_t \). The following theorem proved in [35, Proposition 4.2] claims that the limiting measure \( \pi^\infty \) is a Dirac’s delta function concentrated at a mean-field solution \( \mu^*_t \). We define the following functional
\[
F_t : \mathcal{D}_{[0,T]}(\mathcal{M}(\mathbb{R}^p)) \to \mathbb{R},
\]
\[
\mu \mapsto F_t[\mu] = |\langle \mu_t, f \rangle - \langle \mu_0, f \rangle - \int_0^t R[\mu_s]ds|.
\]

We compute the expectation of the functional \( F_t \) with respect to \( \pi^N \). We then have
\[
\mathbb{E}_{\pi^N}[F_t(\mu)] = \mathbb{E}[F_t[\mu^N_t]]
\]
\[
= \mathbb{E}\left[ \left| \langle \mu^N_t, f \rangle - \langle \mu^N_0, f \rangle - \int_0^t R[\mu^N_s]ds \right| \right].
\]

Now, from Equation (A.63), we have that
\[
\langle \mu^N_t, f \rangle - \langle \mu^N_0, f \rangle - \int_0^t R[\mu^N_s]ds = \mathcal{M}^N_t + \mathcal{R}^N_t.
\]

Plugging (A.99) in Eq. (A.98) gives
\[
\mathbb{E}_{\pi^N}[F_t(\mu)] = \mathbb{E}[F_t[\mu^N_t]]
\]
\[
= \mathbb{E}\left[ \left| \mathcal{M}^N_t + \mathcal{R}^N_t \right| \right]
\leq \mathbb{E}\left[ \sup_{0 \leq t \leq T} |\mathcal{M}^N_t| \right] + \mathbb{E}\left[ \sup_{0 \leq t \leq T} |\mathcal{R}^N_t| \right]
\leq \frac{1}{N\alpha^2} 4\sqrt{2L^2(\lceil NT \rceil \eta C_1(L^2 + \alpha)^2 + \frac{C_0 T}{N}(\eta L^2 T + 2\eta L^4 T)},
\]

xxxvi
where the last equality is due to the bounds in Eqs. (A.64) and (A.65) of Lemma A.9 and A.10, respectively. Taking the limit of $N \to \infty$ from Eq. (A.100) yields
\begin{equation}
\lim_{N_k \to \infty} \mathbb{E}_{\pi_{N_k}}[|F_t[\mu]|] = 0.
\end{equation}
It can be shown that the functional $F_t[\cdot]$ is continuous and bounded. Therefore, due to the weak convergence of the sequence $\{\pi_{N_k}\}_{N_k \in \mathbb{N}}$ to $\pi^\infty$, Eq. (A.101) implies that
\begin{equation}
\mathbb{E}_{\pi^\infty}[|F_t[\mu]|] = 0.
\end{equation}
Since the identity (A.102) holds for all bounded test functions $f \in C^0_b(\mathbb{R}^p)$ and for all $t \in (0, T)$, it follows that $\pi^\infty$ is a Dirac’s delta function concentrated at a solution $(\mu^*_t)_{0 \leq t \leq T}$ of the mean-field equation.

**Step 3: Uniqueness of a mean-field solution:** Before we establish the uniqueness result we make two remarks:

First, we make it clear that from the compact-containment condition (J.1) of Jakubowski’s criterion in Theorem A.12, the support of the measured-valued process $(\mu^N_t)_{0 \leq t \leq T} = (\hat{\mu}^N_{(Nt)})_{0 \leq t \leq T}$ is compact for all $0 \leq t \leq T$. Moreover, in Step 2 of the proof, we established that the measure valued process $(\mu^N_t)_{0 \leq t \leq T}$ converges weakly to a mean-field solution as the number of particles tends to infinity (i.e., $N \to \infty$). Thus, all the possible solutions of the mean-field equation also have compact supports. Let $\mathbb{E} \subset \mathbb{R}^p$ denotes a compact set containing the supports of all such solutions at $0 \leq t \leq T$. In the sequel, it suffices to establish the uniqueness of the mean-field solution for the test functions with a compact domain, i.e., let $f \in C^0_b(\mathbb{E})$.

Second, for all bounded continuous test functions $f \in C^0_b(\mathbb{E})$, the operator $f \to \langle \mu_t, f \rangle$ is a linear operator with $\mu_1(\mathbb{R}^p) = 1$. Hence, from Riesz-Markov-Kakutani representation theorem [40, 47] by assuming $\mu_t \in \mathcal{M}(\mathbb{R}^p)$, existence of unique operator implies $f \mapsto \langle f, \mu_t \rangle$ implies the existence of the unique probability measure $\mu_t$. Now, we equip the measure space $\mathcal{M}(\mathbb{R}^p)$ with the following norm
\begin{equation}
\|\mu\| \overset{\text{def}}{=} \sup_{f \in C^0_b(\mathbb{E}), \|f\| \neq 0} \frac{|\langle f, \mu \rangle|}{\|f\|_\infty}.
\end{equation}

Given an initial measure $\mu_0$, we next prove that there exists at most one mean-field model solution by showing that there exists at most one real valued process $(\mu_t, f)$ corresponding to the mean-field model. Suppose $(\mu^{1*}_t)_{0 \leq t \leq T}$, $(\mu^{2*}_t)_{0 \leq t \leq T}$ are two solutions satisfying the mean-field equations (A.67) with the initial distributions $\mu^{1}_0, \mu^{2}_0 \in \mathcal{M}(\mathbb{R}^p)$, respectively. For any test function $f \in C^0_b(\mathbb{E})$ we have that
\begin{equation}
\langle \mu^{1*}_t - \mu^{2*}_t, f \rangle = \langle \mu^{1}_0 - \mu^{2}_0, f \rangle + \eta \alpha \int_0^t \left( \int_{\mathcal{X} \times \mathcal{Y}} (\langle \varphi(x, \xi) \varphi(\hat{x}, \xi), \mu^{1*}_s - \mu^{2*}_s \rangle - \alpha y_\xi) \right.
\end{equation}
\begin{equation*}
\times \langle \nabla f(\xi)(\nabla_\xi(\varphi(\hat{x}, \xi) \varphi(x; \xi)))^T, \mu^{1*}_s - \mu^{2*}_s \rangle P^{\otimes 2}_x dy_\xi((dx, d\hat{x})) ds.
\end{equation*}

We bound the first term on the right side of Equation (A.104) as follows
\begin{equation}
\langle \mu^{1}_0 - \mu^{2}_0, f \rangle \leq \|\mu^{1}_0 - \mu^{2}_0\| \cdot \|f\|_\infty
\end{equation}
\begin{equation}
\leq b \|\mu^{1}_0 - \mu^{2}_0\|,
\end{equation}
\begin{equation}
\text{xxxvii}
\end{equation}
where used the definition of the norm $\| \cdot \|$ on the measure space $\mathcal{M}(\mathbb{R}^p)$ from Eq. (A.103).

Furthermore, let
\[
\int_{X \times Y} \alpha y \gamma (\nabla f(\xi)(\nabla \xi (\varphi(\bar{x}; \xi)\varphi(x; \xi)))^T, \mu^{2n+1}_s - \mu^{*2}_s) P^{\otimes 2}_x(d(x, y), d(\bar{x}, \bar{y})) \\
\leq \int_{X \times Y} \alpha |y| \gamma (|\nabla f(\xi)(\nabla \xi (\varphi(\bar{x}; \xi)\varphi(x; \xi)))^T, \mu^{2n+1}_s - \mu^{*2}_s)| P^{\otimes 2}_x(d(x, y), d(\bar{x}, \bar{y})) \\
\leq \alpha \|\mu^{2n+1}_s - \mu^{*2}_s\| \int_X |\nabla f(\xi)(\nabla \xi (\varphi(\bar{x}; \xi)\varphi(x; \xi)))^T| P^{\otimes 2}_x(dx, d\bar{x}) \\
\leq \alpha \|\mu^{2n+1}_s - \mu^{*2}_s\| \int_X \|\nabla f(\xi)(\nabla \xi (\varphi(\bar{x}; \xi)\varphi(x; \xi)))\| P^{\otimes 2}_x(dx, d\bar{x})
\]
(A.107) \[
\leq \alpha L^2 C_1 \|\mu^{2n+1}_s - \mu^{*2}_s\|.
\]
where in the last inequality, we used the fact that $\|\nabla f(\xi)\| \leq C_1$ since the test function is three-times continuously differentiable $f \in C_b^3(\Xi)$ on a compact support.

Similarly, we have
\[
\int_X (\varphi(x, \xi) \varphi(\bar{x}, \xi), \mu^{2n+1}_s - \mu^{*2}_s) \langle \nabla f(\xi)(\nabla \xi (\varphi(\bar{x}; \xi)\varphi(x; \xi)))^T, \mu^{2n+1}_s - \mu^{*2}_s \rangle P^{\otimes 2}_x(dx, d\bar{x}) \\
\leq \|\mu^{2n+1}_s - \mu^{*2}_s\|^2 \int_X \|\varphi(x, \xi) \varphi(\bar{x}, \xi)\| P^{\otimes 2}_x(dx, d\bar{x}) \\
\leq \|\mu^{2n+1}_s - \mu^{*2}_s\|^2
\]
(A.108) \[
\leq L^4 C_1 \|\mu^{2n+1}_s - \mu^{*2}_s\|^2.
\]

Putting together the inequalities in Eqs. (A.106),(A.107), and (A.108) yield
\[
(\mu^{2n+1}_s - \mu^{*2}_s, f) \leq b\|\mu^{2n+1}_s - \mu^{*2}_s\| + L^4 C_1 \eta \int_0^t \|\mu^{2n+1}_s - \mu^{*2}_s\| ds + \frac{\eta L^4 C_1}{\alpha} \int_0^t \|\mu^{2n+1}_s - \mu^{*2}_s\|^2 ds.
\]
(A.109) The above inequality holds for all bounded functions $f \in C_b^3(\Xi)$. Thus, by taking the supremum with respect to $f$ we obtain
\[
\|\mu^{2n+1}_s - \mu^{*2}_s\| = \sup_{f \in C_b^3(\Xi)} (\mu^{2n+1}_s - \mu^{*2}_s, f)
\]
(A.110) \[
\leq b\|\mu^{2n+1}_s - \mu^{*2}_s\| + L^4 C_1 \eta \int_0^t \|\mu^{2n+1}_s - \mu^{*2}_s\| ds + \frac{L^4 C_1}{\alpha} \int_0^t \|\mu^{2n+1}_s - \mu^{*2}_s\|^2 ds.
\]
(A.111) Now, we employ the following result which generalizes Gronwall’s inequality when higher order terms are involved:

**Lemma A.16.** (Extended Gronwall’s inequality, [52, Thm 2.1.]) Let $p \in \mathbb{N}$ and suppose that for a.e. $t \in [0, T]$, $u \in L^\infty([0, T])$ satisfies
\[
u_t \leq c_0(t) + \int_0^t (c_1(s)u_s + c_2(s)u^2_s + \cdots + c_{p+1}(s)u^{p+1}_s) ds,
\]
(A.112) where $c_0 \in L_\infty[0, T]$ is non-decreasing, and $c_j \in L^1_\infty[0, T]$ for $j \in \{1, \cdots, p+1\}$. Then, if
\[
\int_0^T c_{j+1}(s)u^2_s ds \leq M_j, \quad j \in \{1, 2, \cdots, p\}.
\]
(A.113)
It follows that for a.e. \( t \in [0,T] \)

\[
\text{(A.114)} \quad u_t \leq c_0(t) \exp \left( \int_0^t c_1(s) \, ds \right) \exp(M_1 + \cdots + M_p).
\]

We now apply the extended Gronewall’s Inequality (A.112) with \( p = 1, c_0(t) = b\|\mu_0^1 - \mu_0^2\|, 0 \leq t \leq T, c_1(t) = \eta L^2 C_1, 0 \leq t \leq T, \) and \( u_s = \|\mu_s^1 - \mu_s^2\| \). In this case, it is easy to see that \( M_1 = \frac{2bT\eta L^4 C_1}{\alpha} \). Hence, from Eqs. (A.110) and (A.114), we obtain that

\[
\text{(A.115)} \quad \|\mu_t^1 - \mu_t^2\| \leq b\|\mu_0^1 - \mu_0^2\| \cdot \exp \left( \eta L^2 C_1 t + \frac{2bT\eta L^4 C_1}{\alpha} \right), \quad 0 \leq t \leq T.
\]

Thus, starting from an initial measure \( \mu_0^1 = \mu_0^2 = \mu_0 \), there exists at most one solution for the mean-field model equations (A.67).

\[\blacksquare\]

A.3. Proof of Corollary 4.2.1. To establish the proof, we recall from (A.58) that

\[
\text{(A.116)} \quad \langle f, \hat{\mu}_m^N \rangle - \langle f, \hat{\mu}_0^N \rangle = \sum_{\ell=0}^{m-1} D_\ell^N + \sum_{\ell=0}^{m-1} M_\ell^N + \sum_{\ell=0}^{m-1} R_\ell^N,
\]

for all \( f \in C_b(\mathbb{R}^p) \). Recall the definition of the Radon distance \( d_R(\cdot, \cdot) \) from (4.6). By taking the supremum of (A.116) with respect to the functions from the function class \( \mathcal{F}_c \text{ def } \{ f \in C_1(\mathbb{R}^p) \} \), we obtain the following upper bound on the Radon distance

\[
\text{(A.117)} \quad d_R(\hat{\mu}_m^N, \hat{\mu}_0^N) \leq \frac{1}{2} \sum_{\ell=0}^{m-1} |D_\ell^N| + \frac{1}{2} \sum_{\ell=0}^{m-1} |M_\ell^N| + \frac{1}{2} \sum_{\ell=0}^{m-1} |R_\ell^N|.
\]

Based on the upper bound (B.23) on the remainder term, we have

\[
\text{(A.118)} \quad |R_\ell^N| \leq \frac{C_0}{N^2} \left( \eta L^2 + 2\frac{\eta}{\alpha} L^4 \right), \quad \ell \in [0, m - 1],
\]

for some constant \( C_0 > 0 \). Moreover, from the concentration inequality (B.34), we also have that with the probability of at least \( 1 - \delta \), the following inequality holds

\[
\text{(A.119)} \quad |M_\ell^N| \leq \frac{8\sqrt{\ell}\eta L^2 C_1 (L^2 + \alpha)}{N\alpha} \log \left( \frac{2}{\delta} \right).
\]

Lastly, recall the definition of the drift term in (A.56a). By carrying out a similar bounding method leading to (A.86), it can be shown that

\[
\text{(A.120)} \quad |D_\ell^N| \leq \frac{2m\eta C_0}{\alpha} (L^2 + \alpha)L^2 C_1.
\]

By plugging (A.118), (A.119), and (A.120) into (A.117), we derive that

\[
\text{(A.121)} \quad d_R(\hat{\mu}_m^N, \hat{\mu}_0^N) \leq \frac{mnC_0}{2N^2} \left( L^2 + 2\frac{L^4}{\alpha} \right) + \frac{8m\sqrt{\eta} L^2 C_1 (L^2 + \alpha)}{N\alpha} \log \left( \frac{2}{\delta} \right) + \frac{mn}{N\alpha} (L^2 + \alpha)L^2 C_1,
\]

with the probability of \( 1 - \delta \).
A.4. **Proof of Theorem 4.6.** The proof is built upon [48]. It suffices to show that for every integer \( \ell \in \mathbb{N} \), and for all the test functions \( f_1, \cdots, f_\ell \in C^3_b(\mathbb{R}^p) \), we have

\[
\limsup_{N \to \infty} \mathbb{E} \left[ \prod_{k=1}^{\ell} f_k(\xi_{\ell Nt}^k) \right] - \prod_{k=1}^{\ell} \langle \mu^*_t, f_k \rangle = 0. \tag{A.122}
\]

Using the triangle inequality, we now have that

\[
\left| \mathbb{E} \left[ \prod_{k=1}^{\ell} f_k(\xi_{\ell Nt}^k) \right] - \prod_{k=1}^{\ell} \langle \mu^*_t, f_k \rangle \right| \leq \left| \mathbb{E} \left[ \prod_{k=1}^{\ell} (\hat{\mu}_t^N, f_k) \right] - \prod_{k=1}^{\ell} \langle \mu^*_t, f_k \rangle \right| + \left| \mathbb{E} \left[ \prod_{k=1}^{\ell} \langle \hat{\mu}_t^N, f_k \rangle \right] - \mathbb{E} \left[ \prod_{k=1}^{\ell} f_k(\xi_{\ell Nt}^k) \right] \right|. \tag{A.123}
\]

For the first term on the right hand side of Eq. (A.123) we have

\[
\limsup_{N \to \infty} \left| \mathbb{E} \left[ \prod_{k=1}^{\ell} (\hat{G}_t^N, f_k) \right] - \prod_{k=1}^{\ell} \langle \mu^*_t, f_k \rangle \right| \leq \limsup_{N \to \infty} \mathbb{E} \left[ \prod_{k=1}^{\ell} (\hat{\mu}_t^N, f_k) - \prod_{k=1}^{\ell} \langle \mu^*_t, f_k \rangle \right] \leq b^\ell \mathbb{E} \left[ \prod_{k=1}^{\ell} (\hat{\mu}_t^N, f_k) - \prod_{k=1}^{\ell} \langle \mu^*_t, f_k \rangle \right] \leq b^\ell \prod_{k=1}^{\ell} \limsup_{N \to \infty} \left| \langle \hat{\mu}_t^N, f_k \rangle - \langle \mu^*_t, f_k \rangle \right| \leq 0, \tag{A.124}
\]

where (a) is by Jensen’s inequality, (b) is by Fatou’s lemma, (c) follows from the basic inequality \( \left| \prod_{i=1}^{N} a_i - \prod_{i=1}^{N} b_i \right| \leq \sum_{i=1}^{N} |a_i - b_i| \) for \( |a_i|, |b_i| \leq 1, 1 \leq i \leq N \), as well as the fact that \( \langle \mu^*_t, f_k \rangle \leq b \) and \( \langle \hat{\mu}_t^N, f_k \rangle \leq b \) for all \( k = 1, 2, \cdots, N \) due to the boundedness of the test functions \( f_1, \cdots, f_\ell \in C^3_b(\mathbb{R}^p) \), and (d) follows from the weak convergence \( \hat{\mu}_t^N \xrightarrow{\text{weak}} \mu^*_t \) to the mean-field solution (A.67).

Now, consider the second term on the right hand side of Eq. (A.123). Due to the exchangability of the initial states \( (\xi_{\ell}^k)_{1 \leq k \leq N} \), the law of the random variables \( (\xi_{\ell}^k)_{1 \leq k \leq N} \) is also exchangible. Therefore, we obtain that

\[
\mathbb{E} \left[ \prod_{k=1}^{\ell} f_k(\xi_{\ell Nt}^k) \right] = \frac{\ell!}{N^\ell} \mathbb{E} \left[ \prod_{\tau \in \Pi(\ell, N)} \prod_{k=1}^{\ell} f_k(\xi_{\tau Nt}^{\pi(k)}) \right], \tag{A.125}
\]

where \( \Pi(\ell, N) \) is the set of all permutations of \( \ell \) numbers selected from \( \{1, 2, \cdots, N\} \). Notice that the right hand side of Eq. (A.125) is the symmetrized version of the left hand side (A.126).

Further, by the definition of the empirical measure \( \hat{\mu}_t^N \) we obtain that

\[
\mathbb{E} \left[ \prod_{k=1}^{\ell} (\hat{\mu}_t^N, f_k) \right] = \frac{1}{N^\ell} \mathbb{E} \left[ \prod_{k=1}^{\ell} \left( \sum_{m=1}^{N} f_k(\xi_{\ell Nt}^{\pi(m)}) \right) \right] = \frac{1}{N^\ell} \mathbb{E} \left[ \sum_{\tau \in \Pi(\ell, N)} \left( \prod_{k=1}^{\ell} f_k(\xi_{\tau Nt}^{\pi(k)}) \right) \right]. \tag{A.126}
\]

\[
\mathbb{E} \left[ \prod_{k=1}^{\ell} (\hat{\mu}_t^N, f_k) \right] = \frac{1}{N^\ell} \mathbb{E} \left[ \sum_{\tau \in \Pi(\ell, N)} \left( \prod_{k=1}^{\ell} f_k(\xi_{\tau Nt}^{\pi(k)}) \right) \right]. \tag{A.127}
\]

\[\text{x1}\]
Therefore, subtracting (A.125) and (A.126) yields
\[
\text{(A.128) } \quad \mathbb{E} \left[ \prod_{k=1}^{\ell} \langle \hat{\mu}_t^N, f_k \rangle \right] - \mathbb{E} \left[ \prod_{k=1}^{\ell} f_k(\xi_{\lfloor Nt \rfloor}^k) \right] \leq b^\ell \left( 1 - \frac{N!}{\ell!N^\ell} \right).
\]

Hence,
\[
\text{(A.129) } \quad \lim_{N \to \infty} \sup_{N} \mathbb{E} \left[ \prod_{k=1}^{\ell} \langle \hat{\mu}_t^N, f_k \rangle \right] - \mathbb{E} \left[ \prod_{k=1}^{\ell} f_k(\xi_{\lfloor Nt \rfloor}^k) \right] = 0.
\]

Combining Eqs. (A.124)-(A.129) yields the desired result.

\[\square\]

APPENDIX B. PROOFS OF AUXILIARY RESULTS

B.1. Proof of Lemma A.2. The upper bound follows trivially by letting \( x = y \) in the optimization problem (A.9).

Now, consider the lower bound. Define the function \( g : [0, 1] \to \mathbb{R}, t \mapsto g(t) = f(y + t(x - y)) \). Then, when \( f \) is differentiable, we have \( g'(t) = \langle x - y, \nabla f(y + t(x - y)) \rangle \). In addition, \( g(0) = f(y) \), and \( g(1) = f(x) \). Based on the basic identity \( g(1) = g(0) + \int_0^1 g'(s)ds \), we derive
\[
f(x) = f(y) + \int_0^1 \langle x - y, \nabla f(y + s(x - y)) \rangle ds \geq f(y) - \|x - y\|_2 \int_0^1 \|\nabla f(y + s(x - y))\|_2 ds,
\]
where the last step is due to the Cauchy-Schwarz inequality. Using Inequality (B.1) yields the following lower bound on Moreau’s envelope
\[
M^\beta_f(y) \geq f(y) + \inf_{x \in \mathcal{X}} \left\{ \frac{1}{2\beta} \|x - y\|_2^2 - \|x - y\|_2 \int_0^1 \|\nabla f(y + s(x - y))\|_2 ds \right\}
\]
\[
= f(y) + \inf_{x \in \mathcal{X}} \left\{ \left( \frac{1}{\sqrt{2\beta}} \|x - y\|_2 - \sqrt{\frac{\beta}{2}} \int_0^1 \|\nabla f(y + s(x - y))\|_2 ds \right)^2 - \frac{\beta}{2} \left( \int_0^1 \|\nabla f(y + s(x - y))\|_2 ds \right)^2 \right\}
\]
\[
\geq f(y) - \frac{\beta}{2} \sup_{x \in \mathcal{X}} \left( \int_0^1 \|\nabla f(y + s(x - y))\|_2 ds \right)^2 \geq f(y) - \frac{\beta}{2} \sup_{x \in \mathcal{X}} \int_0^1 \|\nabla f(y + s(x - y))\|_2^2 ds \geq f(y) - \frac{\beta}{2} \int_0^1 \sup_{x \in \mathcal{X}} \|\nabla f(y + s(x - y))\|_2^2 ds,
\]
where (a) is due to Jensen’s inequality.

\[\square\]

B.2. Proof of Lemma A.8. Let \( u \in \mathbb{R}^d \) denotes an arbitrary unit vector \( \|u\|_2 = 1 \). From the definition of the gradient of a function, we have that
\[
\text{(B.3) } \quad \langle \nabla_M M^\beta_f(x), u \rangle = D_u[M_f(\cdot; \theta)(x)],
\]
where $D_u[M_f(\cdot; \theta)(x)]$ is the directional derivative

\[
D_u[M_f(\cdot; \theta)(x)] \overset{\text{def}}{=} \lim_{\delta \to 0} \frac{M_f(\cdot; \theta + \delta u)(x) - M_f(\cdot; \theta)(x)}{\delta}.
\]

We now have

\[
M_f(\cdot; \theta + \delta u)(x) = \inf_{y \in X} \left\{ \frac{1}{2\beta} \| x - y \|_2^2 + f(y; \theta + \delta u) \right\}
\]

\[
= \inf_{x \in X} \left\{ \frac{1}{2\beta} \| x - y \|_2^2 + f(y; \theta) + \delta \langle \nabla_\theta f(y; \theta), u \rangle \right\} + O(\delta^2)
\]

\[
\overset{(a)}{\leq} \frac{1}{2\beta} \| x - \text{Prox}_{f(\cdot; \theta)}(x) \|_2^2 + f(\text{Prox}_{f(\cdot; \theta)}(x); \theta)
\]

\[
+ \delta \langle \nabla_\theta f(\text{Prox}_{f(\cdot; \theta)}(x); \theta), u \rangle + O(\delta^2),
\]

where the inequality in (a) follows by letting $y = \text{Prox}_{f(\cdot; \theta)}(x)$ in the optimization problem.

Now, recall that

\[
M_f(\cdot; \theta)(x) = \inf_{y \in X} \left\{ \frac{1}{2\beta} \| x - y \|_2^2 + f(y; \theta) \right\},
\]

\[
\text{Prox}_{f(\cdot; \theta)}(x) = \arg \min_{y \in X} \left\{ \frac{1}{2\beta} \| x - y \|_2^2 + f(y; \theta) \right\}.
\]

Therefore,

\[
M_f(\cdot; \theta)(x) = \frac{1}{2\beta} \| x - \text{Prox}_{f(\cdot; \theta)}(x) \|_2^2 + f(\text{Prox}_{f(\cdot; \theta)}(x); \theta).
\]

Substitution of Eq. (B.6) in (B.5) yields

\[
M_f(\cdot; \theta + \delta u)(x) \leq M_f(\cdot; \theta)(x) + \delta \langle \nabla_\theta f(\text{Prox}_{f(\cdot; \theta)}(x); \theta), u \rangle + O(\delta^2).
\]

Hence, $D_u[M_f(\cdot; \theta)(x)] \leq \langle \nabla_\theta f(\text{Prox}_{f(\cdot; \theta)}(x); \theta), u \rangle$. From Eq. (B.3) and by using Cauchy-Schwarz inequality, we compute the following bound on the inner product of the gradient with the unit vectors $u \in \mathbb{R}^d, \|u\|_2 = 1$,

\[
\langle \nabla_\theta M_f^2(\cdot; \theta)(x), u \rangle \leq \|\nabla_\theta f(\text{Prox}_{f(\cdot; \theta)}(x); \theta)\|_2 \cdot \|u\|_2 = \|\nabla_\theta f(\text{Prox}_{f(\cdot; \theta)}(x); \theta)\|_2.
\]

Since the preceding upper bound holds for all the unit vectors $u \in \mathbb{R}^d$, we let $u = \frac{\nabla_\theta M_f(\cdot; \theta)(x)}{\| \nabla_\theta M_f(\cdot; \theta)(x) \|_2}$ to get Inequality (A.35). \[\Box\]

B.3. Proof of Lemma A.6. Let $z \in S^{d-1}$ denotes an arbitrary vector on the sphere. Define the random variable

\[
Q_{\mathbf{a}}(y_1, x_1, \ldots, y_n, x_n) \overset{\text{def}}{=} \langle z, \nabla E_{\mathbf{a}}(\xi) \rangle
\]

\[
= \frac{1}{n(n-1)} \sum_{i \neq j} y_i y_j \left( \varphi(x_i; \xi)(z, \nabla \varphi(x_j; \xi)) + \varphi(x_j; \xi)(z, \nabla \varphi(x_i; \xi)) \right)
\]

\[
- \mathbb{E}_{y \in S^2} \left[ \frac{\delta}{2} \left( \varphi(x_i; \xi)(z, \nabla \varphi(x_j; \xi)) + \varphi(x_j; \xi)(z, \nabla \varphi(x_i; \xi)) \right) \right].
\]

\[
(B.9)
\]
Clearly, \( \mathbb{E}_{P_x}[Q_z] = 0 \). Now, let \((\bar{y}_m, \bar{x}_m) \in \mathcal{Y} \times \mathcal{X}, 1 \leq m \leq n \). By repeated application of the triangle inequality, we obtain that
\[
\left| Q_z((y_1, x_1), \cdots , (y_m, x_m), \cdots , (y_n, x_n)) - Q_z((y_1, x_1), \cdots , (\bar{y}_m, \bar{x}_m), \cdots , (y_n, x_n)) \right|
\leq \frac{1}{n(n-1)} \left| \sum_{i \neq m} y_i \varphi(x_i; \xi) \langle z, y_m \nabla \varphi(x_m; \xi) - \bar{y}_m \nabla \varphi(\bar{x}_m; \xi) \rangle \right|
+ \frac{1}{n(n-1)} \left| \sum_{i \neq m} y_i \langle z, \nabla \varphi(x_i; \xi) \rangle (y_m \varphi(x_m; \xi) - \bar{y}_m \varphi(\bar{x}_m; \xi)) \right|
\leq \frac{1}{n(n-1)} \left| \sum_{i \neq m} |\varphi(x_i; \xi)| \cdot \|z\|_2 \cdot \|y_m \nabla \varphi(x_m; \xi) - \bar{y}_m \nabla \varphi(\bar{x}_m; \xi)\|_2 \right|
+ \frac{1}{n(n-1)} \left| \sum_{i \neq m} \|z\|_2 \cdot \|\nabla \varphi(x_i; \xi)\|_2 \cdot \|y_m \varphi(x_m; \xi) - \bar{y}_m \varphi(\bar{x}_m; \xi)\| \right|
\leq \frac{4L^2}{n},
\]
where the last inequality is due to assumption (A.2) and the fact that \( \|z\|_2 = 1 \) for \( z \in \mathbb{S}^{d-1} \).
In particular, to derive Inequality (B.10), we employed the following upper bounds
\[
|\varphi(x_i; \xi)| \leq L, \quad |y_m \varphi(x_m; \xi) - \bar{y}_m \varphi(\bar{x}_m; \xi)| \leq |\varphi(x_m; \xi)| + |\varphi(\bar{x}_m; \xi)| \leq 2L, \quad \|\nabla \varphi(x_i; \xi)\|_2 \leq L, \quad \|y_m \nabla \varphi(x_m; \xi) - \bar{y}_m \nabla \varphi(\bar{x}_m; \xi)\|_2 \leq \|\nabla \varphi(x_m; \xi)\|_2 + \|\nabla \varphi(\bar{x}_m; \xi)\|_2 \leq 2L.
\]
Using McDiarmid Martingale’s inequality [28] then gives us
\[
\mathbb{P} \left( \left| Q_z((y_1, x_1), \cdots , (y_n, x_n)) \right| \geq u \right) \leq 2 \exp \left( - \frac{nu^2}{16L^4} \right),
\]
for \( x \geq 0 \). Now, for every \( p \in \mathbb{N} \), the \( 2p \)-th moment of the random variable \( Q_z \) is given by
\[
\mathbb{E} \left[ Q_z^{2p}((y_1, x_1), \cdots , (y_n, x_n)) \right] = \int_{\mathbb{R}^+} 2pu^{2p-1} \mathbb{P}(Q_z((y_1, x_1), \cdots , (y_n, x_n)) \geq u)du
\leq \int_{\mathbb{R}^+} 4pu^{2p-1} \exp \left( - \frac{u^2}{16nL^4} \right) du
= 2(16L^4/n)^{2p},
\]
where (a) is due to the concentration bound in Eq. (B.11). Now Therefore,
\[
\mathbb{E} \left[ \exp \left( Q_z^2((y_1, x_1), \cdots , (y_n, x_n))/\gamma \right)^2 \right] = \sum_{p=0}^{\infty} \frac{1}{p! \gamma^{2p}} \mathbb{E} \left[ \phi_z^{2p}((y_1, x_1), \cdots , (y_n, x_n)) \right]
= 1 + \sum_{p \in \mathbb{N}} \left( \frac{16L^4}{n \gamma} \right)^{2p}
= \frac{2}{1 - (16L^4/n \gamma)^2} - 1.
\]
For \( \gamma = 16\sqrt{3}L^4/n \), we obtain \( \mathbb{E} \left[ \exp \left( Q_z^2((y_1, x_1), \cdots , (y_n, x_n))/\gamma \right)^2 \right] \leq 2 \). Therefore, \( \|Q_z\|_{\psi_2} = \|\langle z, \nabla E_n(\xi) \rangle\|_{\psi_2} \leq 16\sqrt{3}L^4/n \) for all \( z \in \mathbb{S}^{n-1} \) and \( \xi \in \mathbb{R}^D \). Consequently, by
the definition of the sub-Gaussian random vector in Eq. (A.16) of Definition A.4, we have \( \| \nabla E_n(\xi) \|_{\psi_2} \leq 16\sqrt{3L^4/n} \) for every \( \xi \in \mathbb{R}^D \). We invoke the following lemma proved by the first author in [19, Lemma 16]:

**Lemma B.1. (The Orlicz Norm of the Squared Vector Norms, [19, Lemma 16])**

Consider the zero-mean random vector \( \mathbf{Z} \) satisfying \( \| \mathbf{Z} \|_{\psi_2} \leq \beta \) for every \( \nu \geq 0 \). Then, \( \| \| \mathbf{Z} \|_{2,\psi_2} \|_{\psi_2} \leq 2 \cdot 3^2 \cdot \beta^2 \).

Using Lemma B.1, we now have that \( \| \| \nabla E_n(\xi) \|_{2,\psi_2} \|_{\psi_2} \leq 4608L^4/n^2 \) for every \( \xi \in \mathbb{R}^D \). Applying the exponential Chebyshev’s inequality with \( \beta = 4608L^4/n^2 \) yields

\[
\mathbb{P} \left( \int_{\mathbb{R}^D} \int_0^1 \| \nabla E_n((1-s)\xi + s\zeta_s) \|_{2,\psi_2}^2 - \mathbb{E}_{x,y}[\| \nabla E_n((1-s)\xi + s\zeta_s) \|_{2,\psi_2}^2] \right) \mu_0(d\xi) \geq \delta 
\leq e^{-\frac{\delta^2}{4608L^4}} \mathbb{E}_{x,y} \left[ e^{\frac{\delta^2}{4608L^4} \left( \| \nabla E_n((1-s)\xi + s\zeta_s) \|_{2,\psi_2}^2 - \mathbb{E}_{x,y}[\| \nabla E_n((1-s)\xi + s\zeta_s) \|_{2,\psi_2}^2] \right)} \right]
\leq e^{-\frac{\delta^2}{4608L^4}} \mathbb{E}_{x,y} \left[ e^{\frac{\delta^2}{4608L^4} \left( \| \nabla E_n((1-s)\xi + s\zeta_s) \|_{2,\psi_2}^2 - \mathbb{E}_{x,y}[\| \nabla E_n((1-s)\xi + s\zeta_s) \|_{2,\psi_2}^2] \right)} \right]
\leq 2e^{-\frac{\delta^2}{4608L^4}},
\]

where (a) follows by Jensen’s inequality, and (b) follows from the fact that

\[
\mathbb{E}_{x,y} \left( e^{\frac{\delta^2}{4608L^4} \left( \| \nabla E_n((1-s)\xi + s\zeta_s) \|_{2,\psi_2}^2 - \mathbb{E}_{x,y}[\| \nabla E_n((1-s)\xi + s\zeta_s) \|_{2,\psi_2}^2] \right)} \right) \leq 2,
\]
due to Definition A.3. Therefore,

\[
\mathbb{P} \left( \int_{\mathbb{R}^D} \int_0^1 \| \nabla E_n((1-s)\xi + s\zeta_s) \|_{2,\psi_2}^2 ds\mu(d\xi) \geq \delta \right) 
\leq 2 \exp \left( -\frac{n^2(\delta - \int_{\mathbb{R}^D} \mathbb{E}_{x,y}[\| \nabla E_n((1-s)\xi + s\zeta_s) \|_{2,\psi_2}^2] ds\mu_0(d\xi))}{4608L^4} \right).
\]

It now remains to compute an upper bound on the expectation \( \mathbb{E}_{x,y}[\| \nabla E_n((1-s)\xi + s\zeta_s) \|_{2,\psi_2}^2] \). But this readily follows from Eq. (B.12) by letting \( p = 1 \) and \( z = \frac{\nabla E_n((1-s)\xi + s\zeta_s)}{\| \nabla E_n((1-s)\xi + s\zeta_s) \|_2} \) as follows

\[
\mathbb{E}_{x,y}[\| \nabla E_n((1-s)\xi + s\zeta_s) \|_{2,\psi_2}^2] = \mathbb{E}_{x,y} \left[ \left( \frac{\nabla E_n((1-s)\xi + s\zeta_s)}{\| \nabla E_n((1-s)\xi + s\zeta_s) \|_2}, \nabla E_n((1-s)\xi + s\zeta_s) \right)^2 \right]
\leq \mathbb{E}_{x,y} \left[ E_{x,y} \left( Q^2 \mathbb{E}_{E_n((1-s)\xi + s\zeta_s)} \right) \right]
\leq 2^9 \frac{L^8}{n^2}.
\]

Plugging the expectation upper bound of Eq. (B.15) into Eq. (B.14) completes the proof of the first part of Lemma A.6.

The second part of Lemma A.6 follows by a similar approach and we thus omit the proof. ■
B.4. Proof of Lemma A.7. Let \( W_* \overset{\text{def}}{=} \arg \min_{W} \Psi(W) \) and \( W_0 \overset{\text{def}}{=} \arg \min_{W} \Phi(W) \). Then, since \( |\Psi(W) - \Phi(W)| \leq \delta \) for all \( W \in \mathcal{W} \), we have that
\[
|\Psi(W_*) - \Phi(W_*)| = \left| \min_{W \in \mathcal{W}} \Psi(W) - \Phi(W_*) \right| \leq \delta.
\]
Therefore,
\[
\min_{W \in \mathcal{W}} \Phi(W) \leq \Phi(W_*) \leq \min_{W \in \mathcal{W}} \Psi(W) + \delta.
\]
Similarly, it can be shown that
\[
\min_{W \in \mathcal{W}} \Psi(W) \leq \Psi(W_*) \leq \min_{W \in \mathcal{W}} \Phi(W) + \delta.
\]
Combining Eqs. (B.17) and (B.18) yields the desired inequality. 

B.5. Proof of Lemma A.10. We recall the expression of the remainder term \( \{R_m^N \}_{0 \leq m \leq NT} \) from Eq. (A.54). For each \( 0 \leq m \leq NT \), \( N \in \mathbb{N} \), we can bound the absolute value of the remainder term as follows
\[
|R_m^N| = \frac{1}{N} \left| \sum_{k=1}^{N} (\xi_{m+1}^{k} - \xi_{m}^{k}) \nabla^2 f(\tilde{x}^{k})(\xi_{m+1}^{k} - \xi_{m}^{k})^T \right|
\]
\[
\leq \frac{1}{N} \sum_{k=1}^{N} \left| \xi_{m+1}^{k} - \xi_{m}^{k} \right| \nabla^2 f(\tilde{x}^{k})(\xi_{m+1}^{k} - \xi_{m}^{k})^T
\]
\[
\leq \frac{1}{N} \sum_{k=1}^{N} \left\| \xi_{m+1}^{k} - \xi_{m}^{k} \right\|_2 \cdot \left\| \nabla^2 f(\tilde{x}^{k}) \right\|_F.
\]
Next, we characterize a bound on the difference term \( \| \xi_{m+1}^{k} - \xi_{m}^{k} \|_2 \). To attain this goal, we use the iterations of the particle SGD in Equation (3.8). We have that
\[
\| \xi_{m+1}^{k} - \xi_{m}^{k} \|_2 \leq \eta \left[ \frac{\eta L^2}{N} + 2 \frac{\eta L^4}{N^2} \right],
\]
where in (a), we used the fact that \( \| \varphi \|_{\infty} < L \) and \( \| \nabla \varphi(x; \xi) \|_2 < L \) due to (A.1), and \( y_m, \tilde{y}_m \in \{-1, 1\} \). Plugging the last inequality in Eq. (B.20) yields
\[
|R_m^N| \leq \frac{1}{N^3} \left( \eta L^2 + 2 \frac{\eta L^4}{N^2} \right) \sum_{k=1}^{N} \left\| \nabla^2 f(\tilde{x}^{k}) \right\|_F.
\]
We next compute an upper bound on the Frobenious norm of the Hessian matrix \( \nabla^2 f(\tilde{x}^{k}) \).
To this end, we first show that there exists a compact set \( C \subset \mathbb{R}^d \) such that \( \xi_{m}^{k} \in C \) for all
\[ k = 1, 2, \ldots, N \] and all \( m \in [0, NT] \cap \mathbb{N} \). For each \( k = 1, 2, \ldots, N \), from Inequality (B.20) we obtain that
\[
\| \xi_m^k \|_2 \leq \| \xi_{m-1}^k \|_2 + \frac{\eta L^2}{N} + \frac{2\eta L^4}{N \alpha} = \| \xi_0^k \|_2 + \frac{m \eta L^2}{N} + \frac{2m \eta L^4}{N \alpha} \leq \| \xi_0^k \|_2 + \eta L^2 T + 2(\eta/\alpha)L^4 T.
\]
(B.22)

Now, \( \| \xi_0^k \|_2 < c_0 \) for some constant \( c_0 > 0 \) since the initial samples \( \xi_0^1, \ldots, \xi_0^N \) are drawn from the measure \( \mu_0 \) whose support \( \text{supp}(\mu_0) = \Xi \) is assumed to be compact by (A.3). From upper bound in Eq. (B.22), it thus follows that \( \| \xi_m^k \|_2 < C \) for some constant \( C > 0 \), for all \( m \in [0, NT] \cap \mathbb{N} \). Now, recall that \( \xi^k = (\xi^k(1), \cdots, \xi^k(p)) \), where \( \xi^k(i) \in \xi^k(i), \xi_{m+1}(i) \), \( i = 1, 2, \cdots, m + 1 \), for \( i = 1, 2, \cdots, p \). Therefore, \( \xi^k \in \mathcal{C} \). Since all the test function \( f \in C^3_0(\mathbb{R}^3) \) are three-times continuously differentiable, it follows that there exists a constant \( C_0 \equiv C_0(T) > 0 \) such that \( \sup_{\xi \in \mathcal{C}} \| \nabla^2 f(\xi) \|_F < C_0 \). From Inequality (B.21), it follows that
\[
| R_m^N | \leq \frac{C_0}{N^2} \left( \eta L^2 + \frac{2 \eta L^4}{\alpha} \right), \quad m \in [0, NT] \cap \mathbb{N}.
\]
(B.23)

Now, recall the definition of the scaled term \( R_m^N \) from Eq. (A.59c). Using the Inequality (B.23) as well as the definition of \( R_m^N \), we obtain
\[
\sup_{0 \leq t \leq T} | R_m^N | \leq \frac{C_0 T}{N} \left( \eta L^2 + \frac{2 \eta L^4}{\alpha} \right).
\]
(B.24)

**B.6. Proof of Lemma A.10.** Let \( \mathcal{F}_{m-1} = \sigma((x_k, y_k)_{0 \leq k \leq m-1}, (\bar{x}_k, \bar{y}_k)_{0 \leq k \leq m-1}) \) denotes the \( \sigma \)-algebra generated by the samples up to time \( m - 1 \). We define \( \mathcal{F}_{m-1} \equiv \emptyset \). Further, define the following random variable
\[
\Delta_m^N \equiv (\langle \varphi(x_m, \xi) \varphi(\bar{x}_m, \xi), \hat{\mu}_m^N \rangle - \alpha y_m \bar{y}_m) \times (\nabla f(\xi) \nabla \varphi(x_m; \xi) \nabla \varphi(\bar{x}_m; \xi), \hat{\mu}_m^N).
\]
(B.25)

Notice that \( \frac{1}{N} \mathbb{E}[| R_m^N | \mathcal{F}_{m-1}] = D_m^N \). We now rewrite the martingale term in Eq. (A.56b) in term of \( \Delta_m^N \),
\[
M_m^N \equiv \frac{\eta}{N \alpha} \sum_{\ell=0}^{m} (\Delta_\ell^N - \mathbb{E}[\Delta_\ell^N | \mathcal{F}_{\ell-1}]),
\]
(B.26)

with \( M_0^N = 0 \).

By construction of \( M_m^N \) in Eq. (B.26), it is a Martingale \( \mathbb{E}[M_m^N | \mathcal{F}_{m-1}] = M_{m-1}^N \). We now prove that \( M_m^N \) has also bounded difference. To do so, we define the shorthand notations
\[
a_m^N \equiv (\varphi(x_m; \xi) \varphi(x_m, \xi), \hat{\mu}_m^N - \alpha y_m \bar{y}_m),
\]
(B.27)
\[
b_m^N \equiv (\nabla f(\xi) (\nabla \varphi(x_m; \xi) \nabla \varphi(\bar{x}_m; \xi)))^T, \hat{\mu}_m^N).
\]
(B.28)
Then, we compute
\[
|M_m^N - M_{m-1}^N| = \frac{\eta}{N\alpha} |\Delta_m^N - \mathbb{E}[\Delta_m^N|\mathcal{F}_{m-1}]|
\]
\[
\leq \frac{\eta}{N\alpha} |\Delta_m^N| + \frac{\eta}{N\alpha} \mathbb{E}[\|\Delta_m^N\||\mathcal{F}_{m-1}]
\]
\[
= \frac{\eta}{N\alpha} |a_m^N| \cdot |b_m^N| + \frac{\eta}{N\alpha} \mathbb{E}[|a_m^N| \cdot |b_m^N| |\mathcal{F}_{m-1}] .
\]
(B.29)

For the difference terms, we derive that
\[
|a_m^N| = \left| \langle \varphi(x_m, \xi) \varphi(\bar{x}_m, \xi), \bar{\mu}_m^N \rangle - \alpha y_m y_m \right|
\]
\[
\leq \frac{1}{N} \sum_{k=1}^N |\varphi(x_m, \xi^k_m) \varphi(\bar{x}_m, \xi^k_m)| + \alpha |y_m y_m|
\]
\[
\leq L^2 + \alpha,
\]
(B.30)
where the last step follows from the fact that \(\|\varphi\|_\infty \leq L\) due to (A.1). Similarly, we obtain that
\[
|b_m^N| = \left| \langle \nabla f(\xi)(\nabla \xi(\varphi(\bar{x}_m; \xi) \varphi(x_m; \xi)))^T, \bar{\mu}_m^N \rangle \right|
\]
\[
\leq \frac{1}{N} \sum_{k=1}^N |\varphi(\bar{x}_m, \xi^k_m)| \cdot |\nabla f(\xi^k_m)(\nabla \xi(\varphi(\bar{x}_m; \xi^k_m)))^T|
\]
\[
+ \frac{1}{N} \sum_{k=1}^N |\varphi(x_m, \xi^k_m)| \cdot |\nabla f(\xi^k_m)(\nabla \xi(\varphi(\bar{x}_m; \xi^k_m)))^T|
\]
\[
\leq \frac{L}{N} \sum_{k=1}^N |\nabla f(\xi_m^k)(\nabla \xi \varphi(\bar{x}_m; \xi^k_m))|^2 + \frac{L}{N} \sum_{k=1}^N |\nabla f(\xi_m^k)(\nabla \xi \varphi(\bar{x}_m; \xi^k_m))|^2
\]
\[
\leq \frac{L}{N} \sum_{k=1}^N \|\nabla f(\xi_m^k)\|^2 \cdot \|\nabla \xi \varphi(\bar{x}_m; \xi^k_m)\|^2 + \frac{L}{N} \sum_{k=1}^N \|\nabla f(\xi_m^k)\|^2 \cdot \|\nabla \xi \varphi(\bar{x}_m; \xi^k_m)\|^2
\]
\[
\leq \frac{2L^2}{N} \sum_{k=1}^N \|\nabla f(\xi_m^k)\|^2,
\]
(B.31)
where (a) and (c) follows from (A.1), and (b) follows from the Cauchy-Schwarz inequality. From Inequality (B.22) and the ensuing discussion in Appendix B.5, we recall that \(\|\xi_m^k\|_2 < C\) for some constant and for all \(m \in [0, NT] \cap \mathbb{N}\), and \(k = 1, 2, \ldots, N\). For the two times continuously test function \(f \in C^2_b(\mathbb{R}^p)\), it then follows that \(\|\nabla f(\xi_m^k)\|_2 \leq C_1\) for some constant \(C_1 > 0\). The following bound can now be computed from Eq. (B.31),
\[
|b_m^N| \leq \frac{2C_1L^2}{N}.
\]
(B.32)

Plugging the upper bounds on \(|a_m^N|\) and \(|b_m^N|\) from Eqs. (B.30)-(B.32) into Eq. (B.29) we obtain that
\[
|M_m^N - M_{m-1}^N| \leq \frac{4\eta C_1}{N\alpha} L^2 (L^2 + \alpha),
\]
(B.33)
Thus, $(M^N_m)_{m \in [0, NT] \cap \mathbb{N}}$ is a Martingale process with bounded difference. From the Azuma-Hoeffding inequality it follows that

$$\mathbb{P}(|M^N_m| \geq \varepsilon) = \mathbb{P}(|M^N_m - M^N_0| \geq \varepsilon) \leq 2 \exp \left(-\frac{N^2 \alpha^2 \varepsilon^2}{8mL^4 \eta^2 C_1^2 (L^2 + \alpha)^2}\right), \quad \forall m \in [0, NT] \cap \mathbb{N}. \tag{B.34}$$

Therefore, since $M^N_T = M^N_{\lfloor NT \rfloor}$, we have

$$\mathbb{P}(|M^N_T| \geq \varepsilon) \leq 2 \exp \left(-\frac{N^2 \alpha^2 \varepsilon^2}{8L^4 \lfloor NT \rfloor \eta^2 C_1^2 (L^2 + \alpha)^2}\right). \tag{B.35}$$

Then,

$$\mathbb{E} \left[ |M^N_T| \right] = \int_0^\infty \mathbb{P}(|M^N_T| \geq \varepsilon) d\varepsilon \leq 2 \int_0^\infty \exp \left(-\frac{N^2 \alpha^2 \varepsilon^2}{8L^4 \lfloor NT \rfloor \eta^2 C_1^2 (L^2 + \alpha)^2}\right) \, d\varepsilon$$

$$= \frac{1}{N\alpha} 4\sqrt{2L^2 \lfloor NT \rfloor} \eta C_1 (L^2 + \alpha)^2. \tag{B.36}$$

where the inequality follows from (B.35).

By Doob’s Martingale inequality [9], the following inequality holds

$$\mathbb{P} \left( \sup_{0 \leq t \leq T} |M^N_t| \geq \varepsilon \right) \leq \frac{\mathbb{E} [ |M^N_T| ]}{\varepsilon} \leq \frac{1}{N\alpha \varepsilon} 4\sqrt{2L^2 \lfloor NT \rfloor} \eta C_1 (L^2 + \alpha)^2. \tag{B.37}$$

In particular, with the probability of at least $1 - \rho$, we have

$$\sup_{0 \leq t \leq T} |M^N_t| \leq \frac{1}{N\alpha \rho} 4\sqrt{2L^2 \lfloor NT \rfloor} \eta C_1 (L^2 + \alpha)^2. \tag{B.38}$$

References

1. Martin Arjovsky, Soumith Chintala, and Léon Bottou, Wasserstein GAN, arXiv preprint arXiv:1701.07875 (2017).
2. Yoram Baram, Learning by kernel polarization, Neural Computation 17 (2005), no. 6, 1264–1275.
3. Patrick Billingsley, Convergence of probability measures, John Wiley & Sons, 2013.
4. Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, Concentration inequalities: A nonasymptotic theory of independence, Oxford university press, 2013.
5. Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li, Mode regularized generative adversarial networks, arXiv preprint arXiv:1612.02136 (2016).
6. Corinna Cortes, Mehryar Mohri, and Afsheen Rostamizadeh, Algorithms for learning kernels based on centered alignment, Journal of Machine Learning Research 13 (2012), no. Mar, 795–828.
7. Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz S Kandola, On kernel-target alignment, Advances in neural information processing systems, 2002, pp. 367–373.
8. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, Imagenet: A large-scale hierarchi- cal image database, 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.
9. Joseph Leo Doob, Stochastic processes, vol. 101, New York Wiley, 1953.
10. Rui Gao and Anton J Kleywegt, Distributionally robust stochastic optimization with wasserstein dis- tance, arXiv preprint arXiv:1604.02199 (2016).
11. Kay Giesecke, Konstantinos Spiliopoulos, Richard B Sowers, et al., Default clustering in large portfolios: Typical events, The Annals of Applied Probability 23 (2013), no. 1, 348–385.
12. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, Generative adversarial nets, Advances in neural information processing systems, 2014, pp. 2672–2680.

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

14. Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur, Optimal kernel choice for large-scale two-sample tests, Advances in neural information processing systems, 2012, pp. 1205–1213.

15. Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in Neural Information Processing Systems, 2017, pp. 6626–6637.

16. Adam Jakubowski, On the skorokhod topology, Annales de l’IHP Probabilités et statistiques, vol. 22, 1986, pp. 263–285.

17. Adel Javanmard, Marco Mondelli, and Andrea Montanari, Analysis of a two-layer neural network via displacement convexity, arXiv preprint arXiv:1901.01375 (2019).

18. Richard Jordan, David Kinderlehrer, and Felix Otto, The variational formulation of the Fokker–Planck equation, SIAM journal on mathematical analysis 29 (1998), no. 1, 1–17.

19. Masoud Badiee Khuzani and Na Li, Stochastic primal-dual method on riemannian manifolds of bounded sectional curvature, 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2017, pp. 133–140.

20. Diederik P Kingma and Max Welling, Auto-encoding variational Bayes, arXiv preprint arXiv:1312.6114 (2013).

21. Alex Krizhevsky and Geoffrey Hinton, Learning multiple layers of features from tiny images, Tech. report, Citeseer, 2009.

22. Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan, Learning the kernel matrix with semidefinite programming, Journal of Machine learning research 5 (2004), no. Jan, 27–72.

23. Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos, Mmd GAN: Towards deeper understanding of moment matching network, Advances in Neural Information Processing Systems, 2017, pp. 2203–2213.

24. Chun-Liang Li, Wei-Cheng Chang, Youssef Mroueh, Yiming Yang, and Barnabás Póczos, Implicit kernel learning, arXiv preprint arXiv:1902.10214 (2019).

25. Yu V Prokhorov, Convergence of random processes and limit theorems in probability theory, Theory of Probability & Its Applications 115 (2018), no. 33, E7665–E7671.

26. Ali Rahimi and Benjamin Recht, Random features for large-scale kernel machines, Advances in neural information processing systems, 2008, pp. 1177–1184.
37. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning, Advances in neural information processing systems, 2009, pp. 1313–1320.
38. Philippe Robert, Stochastic networks and queues, vol. 52, Springer Science & Business Media, 2013.
39. Grant M Rotskoff and Eric Vanden-Eijnden, Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error, arXiv preprint arXiv:1805.09915 (2018).
40. W Rudin, Real and complex analysis mcgraw-hill book co, New York 3rd ed., xiv (1987).
41. Ruslan Salakhutdinov and Geoffrey Hinton, Deep boltzmann machines, Artificial intelligence and statistics, 2009, pp. 448–455.
42. Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, Improved techniques for training GANs, Advances in neural information processing systems, 2016, pp. 2234–2242.
43. Aman Sinha and John C Duchi, Learning kernels with random features, Advances in Neural Information Processing Systems, 2016, pp. 1298–1306.
44. Justin Sirignano and Konstantinos Spiliopoulos, Mean field analysis of neural networks, arXiv preprint arXiv:1805.01053 (2018).
45. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, Rethinking the inception architecture for computer vision, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.
46. Tijmen Tieleman and Geoffrey Hinton, Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural networks for machine learning 4 (2012), no. 2, 26–31.
47. VS Varadarajan, On the theorem of f riesz concerning the form of linear functional, (1958).
48. Thirupathaiah Vasantam, Arpan Mukhopadhyay, and Ravi R Mazumdar, The mean-field behavior of processor sharing systems with general job lengths under the sq (d) policy, Performance Evaluation 127 (2018), 120–153.
49. Chuang Wang, Hong Hu, and Yue M Lu, A solvable high-dimensional model of GAN.
50. Chuang Wang, Jonathan Mattingly, and Yue Lu, Scaling limit: Exact and tractable analysis of online learning algorithms with applications to regularized regression and PCA, arXiv preprint arXiv:1712.04332 (2017).
51. Wei Wang, Yuan Sun, and Saman Halgamuge, Improving mmd-gan training with repulsive loss function, arXiv preprint arXiv:1812.09916 (2018).
52. Jeff Webb, Extensions of Gronwall’s inequality with quadratic growth terms and applications, Electronic Journal of Qualitative Theory of Differential Equations 2018 (2018), no. 61, 1–12.
53. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, Show, attend and tell: Neural image caption generation with visual attention, International conference on machine learning, 2015, pp. 2048–2057.

Stanford University, 450 Serra Mall, Stanford, CA 94305,
E-mail address: mbadieik, liyues, lei@stanford.edu, shahin@tamu.edu.