Journal of Machine Learning Research 25 (2024) 1-51 Submitted 6/23; Revised 4/24; Published 7/24

FedCBO: Reaching Group Consensus in Clustered Federated Learning through Consensus-based Optimization

José A. Carrillo CARRILLO@MATHS.OX.AC.UK
Mathematical Institute
University of Oxford
Oxford OX2 6GG, UK

Nicolás García Trillos GARCIATRILLO@WISC.EDU
Department of Statistics
University of Wisconsin-Madison
1300 University Avenue, Madison, Wisconsin 53706, USA

Sixu Li SLI739@WISC.EDU
Department of Statistics
University of Wisconsin-Madison
1300 University Avenue, Madison, Wisconsin 53706, USA

Yuhua Zhu YUHUA.ZHU@STAT.UCLA.EDU
Department of Statistics and Data Science,
University of California, Los Angeles
Los Angeles, California 90095-1554, USA

Editor: Qiang Liu

Abstract

Federated learning is an important framework in modern machine learning that seeks to integrate the training of learning models from multiple users, each user having their own local data set, in a way that is sensitive to data privacy and to communication loss constraints. In clustered federated learning, one assumes an additional unknown group structure among users, and the goal is to train models that are useful for each group, rather than simply training a single global model for all users. In this paper, we propose a novel solution to the problem of clustered federated learning that is inspired by ideas in consensus-based optimization (CBO). Our new CBO-type method is based on a system of interacting particles that is oblivious to group memberships. Our model is motivated by rigorous mathematical reasoning, which includes a mean-field analysis describing the large number of particles limit of our particle system, as well as convergence guarantees for the simultaneous global optimization of general non-convex objective functions (corresponding to the loss functions of each cluster of users) in the mean-field regime. Experimental results demonstrate the efficacy of our FedCBO algorithm compared to other state-of-the-art methods and help validate our methodological and theoretical work.

Keywords: consensus-based optimization, clustered federated learning, interacting particle system, mean-field limit, asymptotic convergence analysis.

1. Introduction

The wide use of *internet of things* (IoT) devices in various applications such as home automation, personal health monitoring, and vehicle-to-vehicle communications has led to the generation of vast amounts of data across a collective of users. However, concerns around data privacy and security, as well as limitations on communication costs and bandwidth, have made it challenging for an

individual user to take advantage of this large amount of stored information. This has motivated the design and development of federated learning (FL) strategies, which aim at pooling information from learning models trained on local devices to build models *without* relying on the collection of local data (McMahan et al., 2017; Kairouz et al., 2021).

Standard FL approaches aim to learn one global model for all local clients/users (McMahan et al., 2017; Li et al., 2020; Mohri et al., 2019; Karimireddy et al., 2020). However, data heterogeneity, also known as non-i.i.d. data setting, naturally arises in FL applications since data are usually generated from users' personal devices. Thus, it is expected that *no single* global model can perform well across all clients (Sattler et al., 2020). On the other hand, it is reasonable to expect that users with similar backgrounds are likely to make similar decisions and thus generate data following similar distributions. This paper studies one formulation of federated learning with non-i.i.d. data, namely Clustered Federated Learning (CFL) (Sattler et al., 2020; Ghosh et al., 2020; Ruan and Joe-Wong, 2022; Long et al., 2023; Ma et al., 2022). In CFL, users are partitioned into different clusters, and the objective is to train a distinct model for each cluster of users. These clusters may represent, for example, groups of users with preferences in different categories of movies and TV series. Our focus in this work is on the mathematical modeling and analysis of CFL methods and on exploring CFL's effectiveness in improving the performance of FL when dealing with non-i.i.d. data. Specifically, we investigate how CFL can create personalized models for clusters of users with similar preferences. Our research is motivated by previous studies of CFL that have shown promising results in enhancing the performance of FL in the non-i.i.d. data setting (Sattler et al., 2020; Ghosh et al., 2020; Ruan and Joe-Wong, 2022).

To start making our set-up more precise, let us consider the clustered federated learning setting with one global server and N different agents. We assume that each agent belongs to one of K non-overlapping groups denoted by $S_1^*, \dots, S_K^*$. We further assume that each agent belonging to group S_k^* owns data points generated from distribution $\mathcal{D}_k$, and the agent may use these points to train their own learning model. In an ideal scenario, an agent would further seek to communicate with other agents in their group to accelerate the training process of their own model. However, to satisfy data privacy constraints, the underlying partition $S_1^*, \dots, S_K^*$ is never revealed to the learning algorithm, and in particular a single agent will not know the other agents belonging to their group. In other words, no agent shares their local data with a global server or with other agents (see the discussion on privacy in Remark 7). Let $l(\cdot; z) : \Theta \to \mathbb{R}$ be a loss function associated with a data point z, where $\Theta \subset \mathbb{R}^d$ is the parameter space for the learning models. Our goal is to minimize the population loss function

$$L_k(\theta) := \mathbb{E}_{z\sim\mathcal{D}_k}[l(\theta; z)] \tag{1}$$

for all $k \in [K]$ simultaneously. In other words, the goal is to find minimizers θ_k^* for all loss functions

$$\theta_k^* \in \arg\min_{\theta\in\Theta} L_k(\theta), \quad k \in [K]. \tag{2}$$

A toy example illustrating the clustered federated learning framework is shown in Fig. 1.

As suggested by the discussion above, the main difficulty in CFL comes from the fact that *cluster identities of users are unknown.* A CFL algorithm must then be able to induce clustering among users and simultaneously train models in a distributed setting without relying on local data collection. In order to propose an algorithm that accomplishes this, in this paper we abstract the CFL problem and formulate it mathematically borrowing ideas from consensus-based optimization (CBO) (Pinnau et al., 2017; Carrillo et al., 2021; Totzeck, 2021)). CBO is a family of global optimization methodologies based on systems of interacting particles that seek consensus around

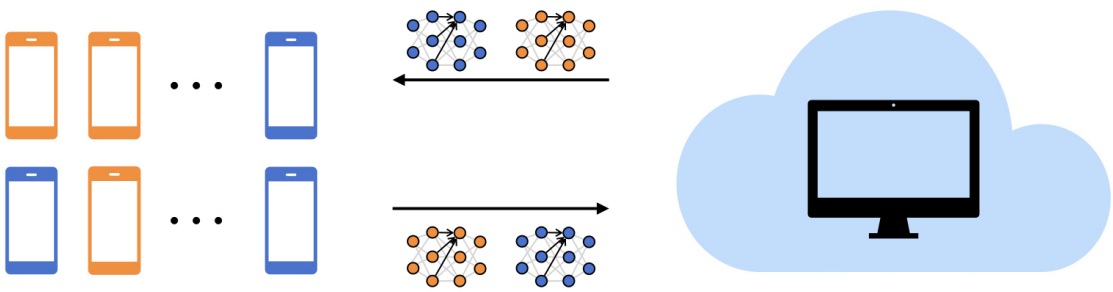

Figure 1: Toy example of a clustered federated learning problem. Each mobile phone user has an underlying cluster identity, here represented by the colors orange and blue. We aim to identify the group memberships of users while simultaneously training models for every cluster by communicating model parameters with the cloud global server.

global minimizers of target objectives. Precisely, consider

$$\min_{\theta\in\mathbb{R}^d} L(\theta),$$

where the target function L, which may be non-convex, is a continuous function with a unique global minimizer θ^*. In what follows we give a brief introduction to the standard CBO framework.

Standard CBO: For each $i \in [N]$, let $\theta^i \in \mathbb{R}^d$ represent the position of particle i and consider the following system of equations

$$d\theta_t^i = -\lambda\left(\theta_t^i - m_t\right)dt + \sigma\left|\theta_t^i - m_t\right|dB_t^i, \qquad \text{for } i = 1, 2, \dots, N, \tag{3}$$

where the $\{B^i\}_i$ are independent Brownian motions and m_t is a weighted average defined by

$$m_t := \frac{\sum_{i=1}^N \theta_t^i \exp\left(-\alpha L(\theta_t^i)\right)}{\sum_{i=1}^N \exp\left(-\alpha L(\theta_t^i)\right)}.$$

One can alternatively consider other types of noise for (3) (see (Carrillo et al., 2021, 2022)). For instance, one may substitute the diffusion term in (3) with a geometric component-wise Brownian motion as in (Carrillo et al., 2021). This noise model improves the performance of the CBO algorithm in that its convergence rate toward the global optimizer of the loss function L in the mean-field limit becomes independent of the dimension d (Fornasier et al., 2022). Due to this, this noise model is suitable for machine learning applications in high dimensions. One can also modify the dynamics and introduce an anisotropic noise term by using a covariance matrix defined similarly to m_t as in (Carrillo et al., 2022) to give rise to a method called Consensus Based Sampling (CBS) in optimization mode. Here we will stick to the basic CBO method for simplicity and refer the interested reader to (Carrillo et al., 2021, 2022) for more details on other existing variants of CBO.

We notice that the term $\exp(-\alpha L(\theta))$ in the formula for m_t is the Gibbs distribution corresponding to the objective function $L(\theta)$ and temperature $\frac{1}{\alpha}$. The motivation for assigning weights in this way comes from the Laplace principle (Miller, 2006; Dembo, 2009), which states that for any probability measure $\rho \in \mathcal{P}(\mathbb{R}^d)$ compactly supported with $\theta^* \in \mathrm{supp}(\rho)$ we have

$$\lim_{\alpha\to+\infty}\left(-\frac{1}{\alpha}\log\left(\int_{\mathbb{R}^d}\exp\left(-\alpha L(\theta)\right)d\rho(\theta)\right)\right) = L(\theta^*).$$

Hence, if θ^* is the unique minimizer of L, then the measure $\exp(-\alpha L(\theta))\rho(\theta)$, normalized by a constant factor, will assign most of its mass to a small neighborhood of θ^*, and if α is large enough, this measure will approximate the Dirac delta distribution at θ^*. Consequently, the weighted-average m_t, which is the empirical first moment of a normalized version of the measure $\exp(-\alpha L(\theta))\rho(\theta)$, is a reasonable target for particles to follow, as induced by equation (3).

Although CBO can be easily adapted to the distributed setting when cluster identities are known (as one could simply run CBO on each cluster), this approach is not directly applicable to CFL if the goal is to propose dynamics that are oblivious to agents' identities. Our problem is also different from standard multi-objective optimization, for which CBO has already been adapted; see (Borghi et al., 2023a, 2022). Indeed, in standard multi-objective optimization the goal is to find a point that is Pareto optimal for K different target functions, whereas our goal is to find, simultaneously, K global minimizers for K different objective functions. On the other hand, the CBO approach is inherently gradient-free, so it is particularly suitable when the objective function is not smooth enough or its derivative is expensive to evaluate. However, if communication costs are expensive as in real FL applications, one may consider introducing additional local gradient terms in the dynamics of each user so that training may continue even when there is no communication among users.

1.1 Contributions and Related Works

1.1.1 Contributions

Motivated by the discussion above, we propose a new CBO-inspired interacting particle system (see (4) below and the discussion right after) that is suited for the clustered federated learning setting. In our system, the evolution of each individual particle is completely determined by its own loss function, its own identity, and the locations of the other particles *but no* knowledge of their identities. More precisely, our main contributions can be summarized as follows:

(1) We introduce a novel CBO-type framework that enables the minimization of K objective functions in a CFL setting without knowing the cluster identity of any of the particles. This is achieved by introducing a mechanism that secretly forces consensus among particles belonging to the same cluster. Moreover, we incorporate a local gradient term for each agent in the particle dynamics, which dramatically reduces the number of communication rounds required for the CBO algorithm to achieve good performance.

(2) We provide rigorous theoretical justification for the proposed framework. In particular, we first prove the well-posedness of the proposed finite particle system and of its corresponding mean-field limit system. Secondly, we study the consensus formation in the mean-field dynamics and explore the dynamics' ability to concentrate around global minimizers of each of the underlying loss functions. Thirdly, we study the approximation of the finite particle system to the mean-field system, and further establish a non-asymptotic concentration bound of the finite particle system around global minimizers.

(3) We discretize our continuous dynamics in a reasonable way and fit it into the conventional federated training protocol to obtain a new federated learning algorithm that we call FedCBO. We conduct extensive numerical experiments to verify the effectiveness and efficacy of our algorithm and demonstrate that it outperforms other state-of-the-art methods in the non-i.i.d. data setting.

1.1.2 Related work in clustered federated learning

In the setting of CFL, it is assumed that there is an underlying cluster structure among users, and the goal is to identify the clusters' identities and federate among each group. Both IFCA (Ghosh

et al., 2020) and HypCluster (Mansour et al., 2020) alternate between identifying cluster identities of users and updating models for the user clusters via local gradient descent. These methods identify cluster identities by finding the model with the lowest loss on each local dataset. FedSEM (Long et al., 2023) groups the users at each federated step by measuring the distance between users using model parameters and accuracy and then running a simple K-means algorithm. All these methods require prior knowledge or estimation of the number of underlying clusters, which may be difficult to have/do in practice. In contrast, as we discuss below, our method does not require any prior knowledge of the clustering structure, and consensus among clients in the same cluster will be automatically induced by our particle dynamics.

In (Sattler et al., 2020), clusters are found in a hierarchical way. In particular, clients are recursively divided into two sets based on the cosine similarity of the clients' model gradients or weight-updates. WeCFL (Ma et al., 2022) formalizes clustered federated learning problems into a unified bi-level optimization framework. Unlike the two framework mentioned above (Sattler et al., 2020; Ma et al., 2022), which conduct the convergence analysis under convexity assumptions, our paper considers target functions that are non-convex and provide an asymptotic convergence result in the mean-field regime.

1.1.3 Related work in consensus-based optimization

The idea of using interacting particle dynamics with consensus-inducing terms to solve global optimization problems was first introduced in (Pinnau et al., 2017). Since then, this approach has gained a lot of interest from both theoretical and applied perspectives.

On the theoretical side, (Carrillo et al., 2018) provided the first local convergence analysis of a mean-field CBO equation under relatively stringent assumptions on the initialization of the system. This is achieved by first proving consensus formation at the mean-field level in the infinite time horizon and then tuning the consensus point using the Laplace principle. Later, (Fornasier et al., 2024a) relaxed some of these stringent assumptions and proved that mean-field dynamics can reach an arbitrary level of concentration around a global minimizer within a finite time interval; however, this time horizon may be difficult to estimate a priori. By showing that the finite particle CBO system converges to the mean-field limit, (Huang and Qiu, 2022; Fornasier et al., 2024a) furthers the theoretical underpinnings of the CBO framework. In our paper, we use similar strategies as in (Fornasier et al., 2024a; Riedl, 2023) to study the behavior of our mean-field system (Section 4) and prove that our proposed finite particle dynamics form consensus around global minimizers of each underlying loss functions (Section 5). In our setting, we need to face new challenges due to the fact that particles may have different dynamics that depend on the loss functions that they try to optimize. For instance, we need to estimate the time horizon needed to achieve a given accuracy at the mean-field level differently from (Fornasier et al., 2024a; Riedl, 2023). Likewise, our convergence of the finite particle system toward a suitable mean-field limit involves additional technical difficulties arising from the fact that in our setting there are multiple types of particles interacting with each other. More recently, (Riedl et al., 2023) establishes the connection between consensus-based optimization (derivative-free method) and gradient-based method, and interprets CBO as a stochastic relaxation of gradient descent. In (Fornasier et al., 2024b), the authors incorporate truncated noise in the original CBO system, which enhances a better theoretical well-behavedness of the law of the dynamics.

On the algorithmic side, with motivations from a variety of applications, researchers have extended and adapted the original CBO model to include new settings such as global optimization on compact manifolds like the sphere (Fornasier et al., 2021), general constraints (Bae et al., 2022; Carrillo et al., 2023; Borghi et al., 2023b), high-dimensional machine learning problems (Carrillo

et al., 2021), global optimization of objective functions with multiple minimizers (Bungert et al., 2024; Fornasier and Sun, 2024), sampling from distributions (Carrillo et al., 2022; Bungert et al., 2024), saddle point optimization problems (Huang et al., 2024), and non-convex multi-player games (Chenchene et al., 2023). In (Riedl, 2023), the author introduces a gradient term in the CBO system which is shown to be beneficial numerically when applied the compressed sensing problems. In our paper, we also incorporate gradient information in our particle dynamics, but our motivation, different from the one in (Riedl, 2023), is to reduce communication costs among users, one of the important practical constraints in federated learning. For a more comprehensive review of the development of CBO-type methods we refer the interested reader to the recent survey (Totzeck, 2021).

1.2 Notation

We use $|\cdot|$ to denote the absolute value or ℓ_2-norm of vectors in Euclidean space and denote by $B_r(\theta)$ the open ball of radius r centered at $\theta \in \mathbb{R}^d$. We use $k \in [K]$ as a short notation for $k = 1, 2, \dots, K$. We denote by $\mathcal{C}(X, Y)$ the space of continuous functions $f : X \to Y$ between $X \subset \mathbb{R}^n$ and a given topological space Y. The space $\mathcal{C}(X, Y)$ is endowed with the sup-norm as is standard. When $Y = \mathbb{R}$, we simply use the notation $\mathcal{C}(X)$. We also use $\mathcal{C}_c^k(X)$ and $\mathcal{C}_b^k(X)$ to denote, respectively, the space of real-valued functions that are k-times continuously differentiable with compact support and the space of bounded functions that are k-times continuously differentiable. Let $\mathcal{P}(\mathbb{R}^d)$ be the space of all Borel probability measures over $\mathbb{R}^d$ equipped with the Levy-Prokhorov metric, which metrizes the topology of weak convergence. For a given $p \geq 1$, we let $\mathcal{P}_p(\mathbb{R}^d) \subseteq \mathcal{P}(\mathbb{R}^d)$ be the collection of probability measures $\rho \in \mathcal{P}(\mathbb{R}^d)$ with finite p-th moments, i.e., $\int_{\mathbb{R}^d} |\theta|^p d\rho(\theta) < \infty$. The space $\mathcal{P}_p(\mathbb{R}^d)$ is endowed with the p-Wasserstein distance W_p $(1 \leq p < \infty)$ defined according to

$$W_p(\rho, \hat{\rho}) := \left(\inf_{\pi \in \Gamma(\rho, \hat{\rho})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \left| \theta - \hat{\theta} \right|^p \pi(d\theta, d\hat{\theta}) \right)^{1/p}, \qquad \rho, \hat{\rho} \in \mathcal{P}_p(\mathbb{R}^d),$$

where $\Gamma(\rho, \hat{\rho})$ denotes the set of all joint probability measures over $\mathbb{R}^d \times \mathbb{R}^d$ with first and second marginals ρ and $\hat{\rho}$, respectively.

For $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$, we denote the law at time t as $\rho_t \in \mathcal{P}(\mathbb{R}^d)$. Given a continuous function $f \in \mathcal{C}(\mathbb{R}^d)$ and a fixed probability measure $\rho \in \mathcal{P}(\mathbb{R}^d)$, we denote by $\|f\|_{\mathbb{L}^1(\rho)} := \int_{\mathbb{R}^d} |f(\theta)| d\rho(\theta)$ the L^1-norm of f with respect to the measure ρ.

1.3 Organization of the Paper

The rest of the paper is organized as follows. In Section 2.1, we introduce the interacting particle system motivating our FedCBO algorithm. In Section 2.2, we state our main theoretical results, which include the well-posedness of both our proposed interacting particle dynamics and its associated mean-field limit system (Theorems 1 and 2), the behavior of the mean-field limit system in time and its ability to concentrate around global optimizers for each of the objective functions (Theorem 3), and finally, large time convergence property of proposed finite particle system (Theorem 4). Motivated by our particle system, in Section 2.3 we introduce our FedCBO algorithm. In Section 2.4, we present a series of numerical experiments to validate our proposed algorithm. Section 3 is devoted to the proof of Theorem 2, Section 4 to the proof of Theorem 3, and Section 5 to the proof of Theorem 4. We wrap up the paper in Section 6, where we summarize our contributions and discuss future research directions.

2. CBO for Clustered Federated Learning

2.1 Model Formulation

Finite particle system: In the rest of the paper we assume, without the loss of generality, that there are only two clusters in the CFL problem (2), i.e., $K = 2$. Indeed, it will become clear from our discussion below that extending the proposed model and its corresponding theoretical analysis to the case $K > 2$ is completely straightforward. We also assume that all agents in class 1 use a single loss function L_1 and all agents in class 2 use a single loss function L_2 (see the discussion on this assumption in Remark 5).[1] In order to optimize L_1 and L_2 simultaneously, we consider a collection of $N = N_1 + N_2 \in \mathbb{N}$ interacting particles with positions $\{\theta_t^{1,i_1}\}_{i_1=1}^{N_1} \in \mathbb{R}^d$ (class 1 particles) and $\{\theta_t^{2,i_2}\}_{i_2=1}^{N_2} \in \mathbb{R}^d$ (class 2 particles) described by the system of stochastic differential equations:

$$d\theta_t^{1,i_1} = -\lambda_1 \left(\theta_t^{1,i_1} - m_t^1\right) dt - \lambda_2 \nabla L_1(\theta_t^{1,i_1}) dt + \sigma_1 \left|\theta_t^{1,i_1} - m_t^1\right| dB_t^{1,i_1} + \sigma_2 \left|\nabla L_1(\theta_t^{1,i_1})\right| d\widetilde{B}_t^{1,i_1}, \quad (4a)$$

$$d\theta_t^{2,i_2} = -\lambda_1 \left(\theta_t^{2,i_2} - m_t^2\right) dt - \lambda_2 \nabla L_2(\theta_t^{2,i_2}) dt + \sigma_1 \left|\theta_t^{2,i_2} - m_t^2\right| dB_t^{2,i_2} + \sigma_2 \left|\nabla L_2(\theta_t^{2,i_2})\right| d\widetilde{B}_t^{2,i_2}, \quad (4b)$$

$$m_t^1 := \frac{\sum_{k=1,2}\sum_{i_k=1}^{N_k} \theta_t^{k,i_k} w_{\boldsymbol{L_1}}^{\alpha}(\theta_t^{k,i_k})}{\sum_{k=1,2}\sum_{i_k=1}^{N_k} w_{\boldsymbol{L_1}}^{\alpha}(\theta_t^{k,i_k})}, \qquad m_t^2 := \frac{\sum_{k=1,2}\sum_{i_k=1}^{N_k} \theta_t^{k,i_k} w_{\boldsymbol{L_2}}^{\alpha}(\theta_t^{k,i_k})}{\sum_{k=1,2}\sum_{i_k=1}^{N_k} w_{\boldsymbol{L_2}}^{\alpha}(\theta_t^{k,i_k})}, \quad (4c)$$

with $\lambda_1, \lambda_2, \sigma_1, \sigma_2 > 0$, $w_{L_k}^{\alpha}(\theta) := \exp(-\alpha L_k(\theta))$ for $k = 1, 2$, and $\alpha > 0$. In the sequel, we may use the terms particle and user interchangeably to refer to an agent.

Let us discuss the above system term by term. Equation (4a) describes the time evolution of the model parameters of agent i_1 in class 1, while equation (4b) does the same for agent i_2 in class 2. The term m_t^1 defined in (4c) is a weighted average of *all* particle positions $\{\theta_t^{1,i_1}\}_{i_1=1}^{N_1}, \{\theta_t^{2,i_2}\}_{i_2=1}^{N_2}$ with respect to the loss function L_1. In particular, note that an agent in class 1 *can* compute m_t^1 without knowing the class identities of any of the other agents, an essential feature for our purposes. If we imagine for a moment that class 1 particles concentrate around regions where L_1 is small, one should expect that $\{\theta_t^{1,i_1}\}_{i_1=1}^{N_1}$ have smaller L_1-loss than the class 2 particles $\{\theta_t^{2,i_2}\}_{i_2=1}^{N_2}$, which presumably should concentrate around regions where the loss function L_2 is small. Then, intuitively, in the expression for m_t^1 class 1 particles $\{\theta_t^{1,i_1}\}_{i_1=1}^{N_1}$ will receive higher weights than class 2 particles and hence m_t^1 should be close to the weighted average of the particles $\{\theta_t^{1,i_1}\}_{i_1=1}^{N_1}$ only. Thus, m_t^1 can be thought of as an evolving consensus point that corresponds to class 1 particles only. A similar intuition holds for m_t^2, which is an evolving consensus point for class 2 particles only. The first part of the drift terms in both (4a) and (4b) can then be thought of as a consensus-inducing term for each of the classes. The second part of the drift terms in (4a) and (4b), on the other hand, introduces local gradient information, which can be interpreted as a local model update through gradient descent. In federated learning, using local gradient information ensures that all models continue to update even when there is no communication between them; as discussed in (McMahan et al., 2017), communication in federated settings is in general costly. Finally, the diffusion terms in the dynamics guarantee that each particle continues to explore the optimization landscape of its loss function until it reaches a critical point and aligns with its class consensus point. The

1. In practice, each agent only has access to finite data samples and thus their empirical loss function may actually differ from that of other agents in the same cluster. We leave the modelling and study of this more realistic and more difficult setting to future work.

d-dimensional Brownian motions $\{B^{k,i_k}\}_{k,i_k}$, $\{\widetilde{B}^{k,i_k}\}_{k,i_k}$ for $k=1,2$ are assumed to be independent of each other.

Let us denote by $\boldsymbol{\theta}_t^{1,N} := \{\theta_t^{1,(i_1,N)}\}_{i_1=1}^{N_1}$, $\boldsymbol{\theta}_t^{2,N} := \{\theta_t^{2,(i_2,N)}\}_{i_2=1}^{N_2}$[2] with $N = N_1 + N_2 \in \mathbb{N}$ the solution of the particle system (4). Consider the empirical measures

$$\rho_t^{1,N} := \frac{1}{N_1}\sum_{i_1=1}^{N_1} \delta_{\theta_t^{1,(i_1,N)}}, \qquad \rho_t^{2,N} := \frac{1}{N_2}\sum_{i_2=1}^{N_2} \delta_{\theta_t^{2,(i_2,N)}}, \qquad \rho_t^N := \frac{N_1}{N}\rho_t^{1,N} + \frac{N_2}{N}\rho_t^{2,N}, \tag{5}$$

where we use δ_θ to represent a Dirac delta measure at $\theta \in \mathbb{R}^d$. Observe that m_t^1, m_t^2 can be rewritten in terms of ρ_t^N as follows:

$$m_t^1 = \frac{1}{\|w_{L_1}^\alpha\|_{\mathbb{L}^1(\rho_t^N)}}\int \theta w_{L_1}^\alpha d\rho_t^N := m_{L_1}^\alpha[\rho_t^N], \qquad m_t^2 = \frac{1}{\|w_{L_2}^\alpha\|_{\mathbb{L}^1(\rho_t^N)}}\int \theta w_{L_2}^\alpha d\rho_t^N := m_{L_2}^\alpha[\rho_t^N].$$

In turn, system (4) can be rewritten as

$$d\theta_t^{1,i_1} = -\lambda_1(\theta_t^{1,i_1} - m_{L_1}^\alpha[\rho_t^N])dt - \lambda_2\nabla L_1(\theta_t^{1,i_1})dt + \sigma_1|\theta_t^{1,i_1} - m_{L_1}^\alpha[\rho_t^N]|dB_t^{1,i_1} + \sigma_2|\nabla L_1(\theta_t^{1,i_1})|d\widetilde{B}_t^{1,i_1},$$
$$d\theta_t^{2,i_2} = -\lambda_1(\theta_t^{2,i_2} - m_{L_2}^\alpha[\rho_t^N])dt - \lambda_2\nabla L_2(\theta_t^{2,i_2})dt + \sigma_1|\theta_t^{2,i_2} - m_{L_2}^\alpha[\rho_t^N]|dB_t^{2,i_2} + \sigma_2|\nabla L_2(\theta_t^{2,i_2})|d\widetilde{B}_t^{2,i_2}.$$

Mean-field system: Based on the above rewriting of the finite particle system in terms of empirical measures, we can formally postulate a mean-field SDE system characterizing the time evolution of "typical" particles as $N_1, N_2 \to \infty$. Precisely, we consider the system of two SDEs:

$$d\overline{\theta}_t^1 = -\lambda_1\left(\overline{\theta}_t^1 - m_{L_1}^\alpha[\rho_t]\right)dt - \lambda_2\nabla L_1(\overline{\theta}_t^1)dt + \sigma_1\left|\overline{\theta}_t^1 - m_{L_1}^\alpha[\rho_t]\right|dB_t^1 + \sigma_2\left|\nabla L_1(\overline{\theta}_t^1)\right|d\widetilde{B}_t^1, \tag{7a}$$

$$d\overline{\theta}_t^2 = -\lambda_1\left(\overline{\theta}_t^2 - m_{L_2}^\alpha[\rho_t]\right)dt - \lambda_2\nabla L_2(\overline{\theta}_t^2)dt + \sigma_1\left|\overline{\theta}_t^2 - m_{L_2}^\alpha[\rho_t]\right|dB_t^2 + \sigma_2\left|\nabla L_2(\overline{\theta}_t^2)\right|d\widetilde{B}_t^2, \tag{7b}$$

where for $k=1,2$,

$$m_{L_k}^\alpha[\rho_t] = \int \theta d\eta_{k,t}^\alpha, \quad \eta_{k,t}^\alpha = w_{L_k}^\alpha\rho_t/\|w_{L_k}^\alpha\|_{\mathbb{L}^1(\rho_t)}, \quad \rho_t^k = Law(\overline{\theta}_t^k), \quad \rho_t = w_1\rho_t^1 + w_2\rho_t^2,$$

subject to the independent initial conditions $\overline{\theta}_0^1 \sim \rho_0^1$ and $\overline{\theta}_0^2 \sim \rho_0^2$; in the above, $B^1, \widetilde{B}^1, B^2, \widetilde{B}^2$ are independent d-dimensional Brownian motions. Equations (7a) and (7b) describe, respectively, the effective time evolutions of individual particles of class 1 and 2 in the regime of a large number of particles. The weight w_k represents the asymptotic proportion of particles of type k in the system. Finally, ρ_t^1 and ρ_t^2 represent, respectively, the distributions of particles of class 1 and 2 at time t. Notice that the system (7) is coupled through the distribution ρ_t of agents of both types. In Section 5, we discuss in precise mathematical terms the relationship between the finite system of interacting particles (4) and the mean-field limit system described in (7).

The system of Fokker-Planck equations corresponding to (7) reads:

$$\partial_t\rho_t^1 := \Delta(\kappa_t^1\rho_t^1) + \nabla\cdot(\mu_t^1\rho_t^1), \qquad \lim_{t\to 0}\rho_t^1 = \rho_0^1 \tag{8a}$$

$$\partial_t\rho_t^2 := \Delta(\kappa_t^2\rho_t^2) + \nabla\cdot(\mu_t^2\rho_t^2), \qquad \lim_{t\to 0}\rho_t^2 = \rho_0^2, \tag{8b}$$

2. Here $\{\theta_t^{1,(i_1,N)}\}_{i_1=1}^{N_1}$ and $\{\theta_t^{2,(i_2,N)}\}_{i_2=1}^{N_2}$ are defined exactly as $\{\theta_t^{1,i_1}\}_{i_1=1}^{N_1}$ and $\{\theta^{2,i_2}\}_{i_2=1}^{N_2}$ in (4). The super-index N in $\theta_t^{1,(i_1,N)}, \theta_t^{2,(i_2,N)}$ emphasizes that this is the system of interacting particles with N total particles. We will omit this super-index when the context is clear.

where

$$\mu_t^k := \lambda_1 \left(\theta - m_{L_k}^\alpha[\rho_t]\right) + \lambda_2 \nabla L_k(\theta), \quad \kappa_t^k := \frac{\sigma_1^2}{2} \left|\theta - m_{L_k}^\alpha[\rho_t]\right|^2 + \frac{\sigma_2^2}{2} \left|\nabla L_k(\theta)\right|^2, \quad \text{for } k = 1, 2.$$

This is a non-linear system of equations that describes the time evolution of the distributions of agents of each class in the mean-field limit.

In the sequel, we interpret the Fokker-Planck system (8) in the weak sense.

Definition 1 *For $k = 1, 2$, let $\rho_0^k \in \mathcal{P}(\mathbb{R}^d)$. Let $T > 0$ be a given time horizon. We say that $\rho^1, \rho^2 \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ satisfy, in the weak sense, the Fokker-Planck equation (8) for the time interval $[0, T]$ and with initial conditions (ρ_0^1, ρ_0^2) if $\forall \phi \in \mathcal{C}_c^\infty(\mathbb{R}^d)$, $\forall t \in (0, T)$, and $k = 1, 2$, we have*

$$\begin{aligned} \frac{d}{dt} \int \phi(\theta) d\rho_t^k(\theta) = &\int \left(-\lambda_1(\theta - m_{L_k}^\alpha[\rho_t]) - \lambda_2 \nabla L_k(\theta)\right) \cdot \nabla \phi(\theta) d\rho_t^k(\theta) \\ &+ \int \left(\frac{\sigma_1^2}{2} \left|\theta - m_{L_k}^\alpha[\rho_t]\right|^2 + \frac{\sigma_2^2}{2} \left|\nabla L_k(\theta)\right|^2\right) \Delta \phi(\theta) d\rho_t^k(\theta), \end{aligned} \tag{9}$$

and in addition $\lim_{t \to 0} \rho_t^k = \rho_0^k$ (in the sense of weak convergence of probability measures).

Our main theoretical results, presented in the next section, are split into four key theorems. First, we discuss the well-posedness of the proposed finite particle system (4) and of its corresponding mean-field SDE system (7) in Theorem 1 and Theorem 2, respectively. Second, we discuss in Theorem 3 the long-time behavior properties of the mean-field PDEs (8) and show that, under some mild assumptions on initialization and the correct tuning of parameters, each of the distributions ρ_t^1 and ρ_t^2 concentrates around the global minimizers of L_1 and L_2, respectively, within a certain time interval. Finally, we prove that by running the *finite* particle system (4) for long enough, the particles in each cluster will reach consensus around the global minimizer of their objective function. This result is presented in Theorem 4 and is established by combining the quantitative mean-field approximation result proved in Proposition 1 and the long-time behavior of the mean-field system discussed in Theorem 3. The practical implication of the combination of these theoretical results is the following: by considering the system (4) with sufficiently large N_1 and N_2, and assuming appropriate initialization, particles of class 1 will concentrate around the global minimizer of the loss function L_1, while particles of class 2 will do the same around the global minimizer of L_2. In Section 2.3, we use our mathematical model to motivate a new algorithm for clustered federated learning. In Section 2.4, we show through numerical experimentation that the proposed algorithm can indeed produce highly-performing learning models for groups of users with similar data sets.

2.2 Main Theoretical Results

In all our theoretical analysis, we make the following assumptions on the loss functions L_1, L_2.

Assumption 1 *For $k = 1, 2$, the loss function $L_k : \mathbb{R}^d \to \mathbb{R}$ is assumed to be bounded from below and we denote $\underline{L_k} := \inf L_k$ the largest lower bound. Moreover, we assume there exist constants*

$$M_{L_k}, C_{L_k}, M_{\nabla L_k}, C_{\nabla L_k}, M, c_{q_k} > 0$$

such that for all $\theta, \hat{\theta} \in \mathbb{R}^d$,

$$L_k(\theta) - \underline{L_k} \leq C_{L_k}\left(1 + |\theta|^2\right) \tag{10a}$$

$$\left|\nabla L_k(\theta) - \nabla L_k(\hat{\theta})\right| \leq M_{\nabla L_k}\left|\theta - \hat{\theta}\right| \tag{10b}$$

$$|\nabla L_k(\theta)| \leq C_{\nabla L_k} \tag{10c}$$

$$L_k(\theta) \leq \overline{L_k} := \sup L_k \quad \text{or} \quad L_k(\theta) - \underline{L_k} \geq c_{q_k}|\theta|^2 \ \text{ for all } |\theta| \geq M. \tag{10d}$$

In simple words, in (10a) we assume the loss functions L_k are bounded above by quadratic functions. We also assume the gradients of L_k to be Lipschitz and bounded in (10b) and (10c). In (10d), we consider loss functions L_k that either are (1) bounded from above, in particular, L_k has the upper bound $\overline{L_k} := \sup L_k$; or (2) quadratic growth at infinity, i.e., there exist constants $M > 0$ and $c_{q_k} > 0$ such that $L_k(\theta) - \underline{L_k} \geq c_{q_k}|\theta|^2$ for all $|\theta| \geq M$.

On our first main result, we study the existence of a unique process $\{\boldsymbol{\theta}_t^N \,|\, t \geq 0\}$ with

$$\boldsymbol{\theta}^N := \left(\theta^{1,(1,N)}, \cdots \theta^{1,(N_1,N)}, \theta^{2,(1,N)}, \cdots, \theta^{2,(N_2,N)}\right)^T \in \mathbb{R}^{Nd},$$

following the interacting particle system (4). In particular, by invoking standard existence results of strong solutions for systems of SDEs from (Durrett, 2018), we have the following theorem.

Theorem 1 (Well Posedness of the Microscopic Model) *For fixed* $N \in \mathbb{N}$*, the stochastic differential equation system* (4) *has a unique strong solution* $\{\boldsymbol{\theta}_t^N \,|\, t \geq 0\}$ *for any initial condition* $\boldsymbol{\theta}_0^N$ *satisfying* $\mathbb{E}|\boldsymbol{\theta}_0^N|^2 < \infty$*.*

The above theorem for fixed number of particles N doesn't directly generalize to the mean-field limit $N \to \infty$ (see Remark 10). In our second main result, we establish the well-posedness of the mean-field system of equations.

Theorem 2 (Well Posedness of mean-field equations) *Let* L_1, L_2 *satisfy Assumption 1 and be either bounded or have quadratic growth at infinity, and let* $\rho_0^1, \rho_0^2 \in \mathcal{P}_4(\mathbb{R}^d)$*. Then there exist unique nonlinear processes* $\overline{\theta}^1, \overline{\theta}^2 \in \mathcal{C}([0,T], \mathbb{R}^d), T > 0$*, satisfying*

$$\begin{aligned}
d\overline{\theta}_t^1 &= -\lambda_1\left(\overline{\theta}_t^1 - m_{L_1}^{\alpha}[\rho_t]\right)dt - \lambda_2 \nabla L_1(\overline{\theta}_t^1)dt + \sigma_1\left|\overline{\theta}_t^1 - m_{L_1}^{\alpha}[\rho_t]\right| dB_t^1 + \sigma_2\left|\nabla L_1(\overline{\theta}_t^1)\right| d\widetilde{B}_t^1, \\
d\overline{\theta}_t^2 &= -\lambda_1\left(\overline{\theta}_t^2 - m_{L_2}^{\alpha}[\rho_t]\right)dt - \lambda_2 \nabla L_2(\overline{\theta}_t^2)dt + \sigma_1\left|\overline{\theta}_t^2 - m_{L_2}^{\alpha}[\rho_t]\right| dB_t^2 + \sigma_2\left|\nabla L_2(\overline{\theta}_t^2)\right| d\widetilde{B}_t^2, \\
\rho_t &= w_1\rho_t^1 + w_2\rho_t^2, \qquad w_1 + w_2 = 1
\end{aligned}$$

in the strong sense, with $\rho_t^1 = Law(\overline{\theta}_t^1), \rho_t^2 = Law(\overline{\theta}_t^2), \rho_t \in \mathcal{C}([0,T], \mathcal{P}_4(\mathbb{R}^d))$ *satisfying the corresponding Fokker-Planck equations (8a) and (8b) in the weak sense, with* $\lim_{t\to 0}\rho_t^k = \rho_0^k \in \mathcal{P}_4(\mathbb{R}^d)$ *for* $k = 1, 2$*.*

Remark 1 *As illustrated in (Fornasier et al., 2024a, Theorem 8), the additional regularity of* $\rho^1, \rho^2 \in \mathcal{C}\left([0,T], \mathcal{P}_4(\mathbb{R}^d)\right)$ *stated in Theorem 2 is a consequence of the regularity of the initial conditions* $\rho_0^1, \rho_0^2 \in \mathcal{P}_4(\mathbb{R}^d)$*. Namely, if* $\rho^1, \rho^2 \in \mathcal{C}\left([0,T], \mathcal{P}_4(\mathbb{R}^d)\right)$ *solves* (8a) *and* (8b) *in the weak sense, identity* (9) *holds for all* $\phi \in \mathcal{C}_*^2(\mathbb{R}^d)$*, where for some constant* $C > 0$ *we define the test function space*

$$\mathcal{C}_*^2(\mathbb{R}^d) := \left\{\phi \in \mathcal{C}^2(\mathbb{R}^d) : \|\nabla\phi(v)\|_2 \leq C(1 + \|v\|_2) \ \text{ and } \sup_{v\in\mathbb{R}^d}|\Delta\phi(v)| < \infty\right\}.$$

Knowing that both the mean-field SDE system (7) and its corresponding Fokker-Planck system are well-posed, we can now state precisely the global convergence property of the mean-field system. For our next result, we impose additional assumptions on the loss functions L_1, L_2.

Assumption 2 *For $k = 1, 2$, we assume that the function $L_k \in \mathcal{C}(\mathbb{R}^d)$ satisfies*

(I) There exists $\theta_k^* \in \mathbb{R}^d$ such that $L_k(\theta_k^*) = \inf_{\theta\in\mathbb{R}^d} L_k(\theta) := \underline{L_k}$.

(II) There exist $L_\infty^k, R_0^k, \eta_k > 0$, and $\nu_k \in (0, \frac{1}{2}]$ such that

$$|\theta - \theta_k^*| \leq \frac{1}{\eta_k}\big(L_k(\theta) - \underline{L_k}\big)^{\nu_k} \qquad \text{for all } \theta \in B_{R_0^k}(\theta_k^*), \tag{12}$$

$$L_\infty^k < L_k(\theta) - \underline{L_k} \qquad \text{for all } \theta \in \big(B_{R_0^k}(\theta_k^*)\big)^c. \tag{13}$$

Assumption 2 is similar to assumptions used in (Fornasier et al., 2024a). Assumption (I) states that the minimum value $\underline{L_k}$ of the objective function is reached at some point θ_k^*. Assumption (II) specifies certain required properties of the objective functions' landscapes for our theory to hold. Specifically, the first part, inequality (12), imposes lower bounds on the local growth of L_k around the global minimizer θ_k^*. This condition is also known as the inverse continuity condition as discussed in (Fornasier et al., 2021), and it has been observed to hold globally for objectives useful in machine learning problems as in (Xu et al., 2017; Fornasier et al., 2021). The second part of (II), condition (13), rules out the possibility that $L_k(\theta) \approx \underline{L_k}$ for some θ outside a neighborhood of θ_k^*.

Theorem 3 (Concentration of mean-field around global minimizers) *For $k = 1, 2$, suppose $L_k \in \mathcal{C}(\mathbb{R}^d)$ satisfy Assumptions 1 and 2. Moreover, let $\rho_0^k \in \mathcal{P}_4(\mathbb{R}^d)$ be such that $\rho_0^k(B_r(\theta_k^*)) > 0$ for all $r > 0$. Define $\mathcal{V}(\rho_t^k) := \frac{1}{2}\int |\theta - \theta_k^*|^2 d\rho_t^k(\theta)$. For any $\varepsilon \in (0, \mathcal{V}(\rho_0^1) + \mathcal{V}(\rho_0^2))$, $\vartheta \in (0, 1)$, parameters $\lambda_1, \lambda_2, \sigma_1, \sigma_2 > 0$ satisfying $2\lambda_1 > 2\lambda_2 M + d\sigma_1^2 + d\sigma_2^2 M^2$, where $M := \max\{M_{\nabla L_1}, M_{\nabla L_2}\}$, and the time horizon*

$$T^* := \frac{1}{(1-\vartheta)(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2)} \log\left(\frac{\mathcal{V}(\rho_0^1) + \mathcal{V}(\rho_0^2)}{\varepsilon}\right), \tag{14}$$

there exists $\alpha_0 > 0$, which depends among other problem dependent quantities, on ε, ϑ and the distance between the global minimizers θ_1^* and θ_2^* (see (48) for a precise definition), such that for all $\alpha > \alpha_0$, if $\rho^1, \rho^2 \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ are the weak solutions to the Fokker-Planck equations (8a) and (8b), respectively, on the time interval $[0, T^*]$ with initial conditions ρ_0^1, ρ_0^2, we have $\min_{t\in[0,T^*]}\big(\mathcal{V}(\rho_t^1) + \mathcal{V}(\rho_t^2)\big) \leq \varepsilon$. Furthermore, up until $\mathcal{V}(\rho_t^1) + \mathcal{V}(\rho_t^2)$ reaches the prescribed accuracy ε for the first time, we have the exponential decay

$$\mathcal{V}(\rho_t^1) + \mathcal{V}(\rho_t^2) \leq \big(\mathcal{V}(\rho_0^1) + \mathcal{V}(\rho_0^2)\big) \exp\big(-(1-\vartheta)(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2)t\big). \tag{15}$$

Remark 2 *The parameter λ_1 in (4) determines the strength of the force driving particles toward their respective consensus points. Similarly, λ_2 and σ_1, σ_2 characterize the strength of the gradient and noise terms, respectively. In Theorem 3, we require the parameters $\lambda_1, \lambda_2, \sigma_1, \sigma_2 > 0$ to satisfy $2\lambda_1 > 2\lambda_2 M + d\sigma_1^2 + d\sigma_2^2 M^2$. This requirement ensures that the consensus inducing terms dominate the other terms in the dynamics, a crucial feature for the system to reach consensus around the global minimizers of the loss functions. This, however, is an assumption that we impose for theoretical purposes, as in fact a stronger drift toward consensus translates to more communication rounds between agents. Nevertheless, as we will see in our numerical experiments in section 2.4, our proposed FedCBO algorithm, introduced in the next section, continues to induce consensus among cluster members even when reasonable communication constraints are imposed.*

Based on Theorem 3, we also establish a convergence result for the *finite* particle system (4a) and (4b) toward the global minimizers θ_1^* and θ_2^*, respectively, by following the proof techniques developed in (Fornasier et al., 2021, 2022). By excluding a bad set with small probability we can further establish a polynomial sample complexity for the approximation, provided that particles are initialized by sampling distributions satisfying the assumptions in Theorem 3.

Theorem 4 (Finite particle system converges in probability) *Let $\varepsilon_{total} > 0$ and $\xi \in (0, \frac{1}{2})$ be fixed. Under the same assumptions as in Theorem 3, let $\{\theta^{1,i_1}\}_{i_1=1}^{N_1}, \{\theta^{2,i_2}\}_{i_2=1}^{N_2}$ be generated by running the finite particle system (4) up to time T^* (defined in (14)) initialized by sampling the measures ρ_0^1 and ρ_0^2, respectively. Then there is $T \leq T^*$ such that*

$$\left| \frac{1}{N_1} \sum_{i_1=1}^{N_1} \theta_T^{1,i_1} - \theta_1^* \right|^2 + \left| \frac{1}{N_2} \sum_{i_2=1}^{N_2} \theta_T^{2,i_2} - \theta_2^* \right|^2 \leq \varepsilon_{total} \tag{16}$$

with probability greater than $1 - \left(\xi + \varepsilon_{total}^{-1}(2C_{MFA}(N_1^{-1} + N_2^{-1}) + 4\varepsilon)\right)$. Here, ε is the error from Theorem 3. $C_{MFA} > 0$ depends on the particle system parameters $\alpha, \lambda_1, \lambda_2, \sigma_1$ and σ_2, on the time horizon T^*, and on ξ^{-1}.

Remark 3 *Notice that the probability of the event where (16) holds can be made to be larger than $1 - \delta$ by choosing ε to be small enough so that $\frac{4\varepsilon}{\varepsilon_{total}} \leq \delta/3$, then taking ξ smaller than $\delta/3$, and, finally, N_1 and N_2 large enough so that $\frac{2C_{MFA}}{\varepsilon_{total}}(N_1^{-1} + N_2^{-1})$ is less than $\delta/3$.*

Remark 4 *The error analysis of the numerical scheme (20) approximating the continuous dynamics (4) is out of scope for this paper. Therefore, in Theorem 4 the error analysis is for the continuous time scheme only and we leave the analysis of the error introduced by time discretization for future work.*

2.3 The FedCBO Algorithm

To make the system (4) into a practical algorithm for federated learning, we need to make a series of adjustments. Firstly, we discretize the proposed continuous-time system, which can be done using a standard Euler-Maruyama discretization of (4). Secondly, the resulting discretized scheme must be adapted to fit the conventional federated training protocol where the number of communication rounds among users is restricted. The resulting algorithm, which we name FedCBO, is the combination of these adjustments. We provide more details next.

Let $\{\theta_0^{1,i_1}\}_{i_1=1}^{N_1}, \{\theta_0^{2,i_2}\}_{i_2=1}^{N_2}$ be sampled from fixed distributions $\rho_0^1, \rho_0^2 \in \mathcal{P}(\mathbb{R}^d)$, respectively. Since class memberships are not given, it is reasonable to assume that ρ_0^1 and ρ_0^2 are the same, but this assumption is not required. Consider the iterates

$$\theta_{n+1}^{1,i_1} \leftarrow \theta_n^{1,i_1} - \lambda_1\gamma\left(\theta_n^{1,i_1} - m_n^1\right) - \lambda_2\gamma\nabla L_1(\theta_n^{1,i_1}) + \sigma_1\sqrt{\gamma}\left|\theta_n^{1,i_1} - m_n^1\right| z_n^{1,i_1} + \sigma_2\sqrt{\gamma}\left|\nabla L_1(\theta_n^{1,i_1})\right| \widetilde{z}_n^{1,i_1}, \tag{17a}$$

$$\theta_{n+1}^{2,i_2} \leftarrow \theta_n^{2,i_2} - \lambda_1\gamma\left(\theta_n^{2,i_2} - m_n^2\right) - \lambda_2\gamma\nabla L_2(\theta_n^{2,i_2}) + \sigma_1\sqrt{\gamma}\left|\theta_n^{2,i_2} - m_n^2\right| z_n^{2,i_2} + \sigma_2\sqrt{\gamma}\left|\nabla L_2(\theta_n^{2,i_2})\right| \widetilde{z}_n^{2,i_2}, \tag{17b}$$

for $n = 0, 1, 2, \ldots$. Here, γ is the discretization step size; $z_n^{k,i_k}, \widetilde{z}_n^{k,i_k}$ for $k = 1, 2$ are independent normal random vectors $N(0, I_{d\times d})$; $m_n^k, k = 1, 2$ are the weighted averages of $\{\theta_n^{1,i_1}\}_{i_1=1}^{N_1}, \{\theta_n^{2,i_2}\}_{i_2=1}^{N_2}$

defined by

$$m_n^1\big[\{\theta_n^{1,i_1}\},\{\theta_n^{2,i_2}\}\big]=\frac{\sum_{k=1,2}\sum_{i_k=1}^{N_k}\theta_n^{k,i_k}w_{\boldsymbol{L_1}}^{\alpha}(\theta_n^{k,i_k})}{\sum_{k=1,2}\sum_{i_k=1}^{N_k}w_{\boldsymbol{L_1}}^{\alpha}(\theta_n^{k,i_k})}, \tag{18a}$$

$$m_n^2\big[\{\theta_n^{1,i_1}\},\{\theta_n^{2,i_2}\}\big]=\frac{\sum_{k=1,2}\sum_{i_k=1}^{N_k}\theta_n^{k,i_k}w_{\boldsymbol{L_2}}^{\alpha}(\theta_n^{k,i_k})}{\sum_{k=1,2}\sum_{i_k=1}^{N_k}w_{\boldsymbol{L_2}}^{\alpha}(\theta_n^{k,i_k})},, \tag{18b}$$

where $w_{L_k}^{\alpha}(\theta)=\exp(-\alpha L_k(\theta))$ for $k=1,2$. Given a fixed integer $\tau>0$, by summing over (17a) and (17b) τ times, we can rewrite (17a) and (17b) (omitting noise terms for simplicity. See the discussion in Appendix E) as

$$\theta_{(n+1)\tau}^{1,i_1}\leftarrow\theta_{n\tau}^{1,i_1}-\lambda_1\gamma\sum_{q=0}^{\tau-1}\Big(\theta_{n\tau+q}^{1,i_1}-m_{n\tau+q}^1\Big)-\lambda_2\gamma\sum_{q=0}^{\tau-1}\nabla L_1(\theta_{n\tau+q}^{1,i_1}), \tag{19a}$$

$$\theta_{(n+1)\tau}^{2,i_2}\leftarrow\theta_{n\tau}^{2,i_2}-\lambda_1\gamma\sum_{q=0}^{\tau-1}\Big(\theta_{n\tau+q}^{2,i_2}-m_{n\tau+q}^2\Big)-\lambda_2\gamma\sum_{q=0}^{\tau-1}\nabla L_2(\theta_{n\tau+q}^{2,i_2}), \tag{19b}$$

However, the above update rule would require the computation of the consensus points $m_{n\tau+q}^1$ and $m_{n\tau+q}^2$ at each iterate. This would result in an excessive amount of communication among users and server, a situation that must be avoided in practical settings. Indeed, note that an agent would need to download the parameters of all other users participating in the training to compute its corresponding $m_{n\tau+q}^k$. If this communication is done too often, it could quickly become prohibitively expensive. To accommodate our algorithm to this practical constraint, we consider a splitting scheme that *approximates* the update formula (19a) and (19b) in the following way: for $n=0,1,2,\cdots,$

$$\widehat{\theta}_{n\tau}^{1,i_1}\leftarrow\theta_{n\tau}^{1,i_1},\quad \widehat{\theta}_{n\tau}^{2,i_2}\leftarrow\theta_{n\tau}^{2,i_2}, \tag{20a}$$

$$\widehat{\theta}_{n\tau+q+1}^{1,i_1}\leftarrow\widehat{\theta}_{n\tau+q}^{1,i_1}-\lambda_2\gamma\nabla L_1(\widehat{\theta}_{n\tau+q}^{1,i_1}),\quad \widehat{\theta}_{n\tau+q+1}^{2,i_2}\leftarrow\widehat{\theta}_{n\tau+q}^{2,i_2}-\lambda_2\gamma\nabla L_2(\widehat{\theta}_{n\tau+q}^{2,i_2})\qquad \text{for } q=0,\dots,\tau-1, \tag{20b}$$

$$\theta_{(n+1)\tau}^{1,i_1}\leftarrow\widehat{\theta}_{(n+1)\tau}^{1,i_1}-\lambda_1\gamma\left(\widehat{\theta}_{(n+1)\tau}^{1,i_1}-m_{(n+1)\tau}^1\right),\quad \theta_{(n+1)\tau}^{2,i_2}\leftarrow\widehat{\theta}_{(n+1)\tau}^{2,i_2}-\lambda_1\gamma\left(\widehat{\theta}_{(n+1)\tau}^{2,i_2}-m_{(n+1)\tau}^2\right), \tag{20c}$$

where the consensus points $m_{(n+1)\tau}^k:=m_{(n+1)\tau}^k\big[\{\widehat{\theta}_{(n+1)\tau}^{1,i_1}\},\{\widehat{\theta}_{(n+1)\tau}^{2,i_2}\}\big]$ for $k=1,2$ are the weighted average of $\{\widehat{\theta}_{(n+1)\tau}^{1,i_1}\}_{i_1=1}^{N_1},\{\widehat{\theta}_{(n+1)\tau}^{2,i_2}\}_{i_2=1}^{N_2}$ defined as in (18a) and (18b). In simple terms, at each communication round we first update models through gradient descent τ times (20b) and then compute the consensus points once (20c). For the above scheme to resemble (19a) as much as possible, we set a larger value for λ_1 than for λ_2. In the standard terminology in federated training, (20b) can be interpreted as a local update of each user's model parameters through τ epochs of local gradient descent, while (20c) can be viewed as the aggregation step. One interesting feature of our update rules is that the model aggregation does not occur at the global server. Instead, agents may download other users' models and aggregate them through (20c) locally. Thus, the server can be assumed to be completely oblivious to not only class memberships but also to the actual values of all user parameters. This feature makes our FedCBO approach a rather decentralized approach to federated learning.

We are ready to present the FedCBO algorithm (Algorithm 1) in precise terms. At the n-th iteration of FedCBO, the central server selects a subset of participating agents $G_n\subseteq[N]$; in

Algorithm 1 FedCBO

Input: Initialized model $\theta_0^j \in \mathbb{R}^d, j \in [N]$; Number of iterations T; Number of local gradient steps τ; Number of models downloaded M; CBO system hyperparameters $\lambda_1, \lambda_2, \alpha$; Discretization step size γ; Initialized sampling likelihood $P_0 \in \mathbb{R}^{N\times(N-1)}$;

for $n = 0, \cdots, T-1$ **do**
  $G_n \leftarrow$ random subset of agents (participating devices);
  LocalUpdate$(\theta_n^j, \tau, \lambda_2, \gamma)$ for $j \in G_n$;
  LocalAggregation(agent j) for $j \in G_n$;
end for

Output: θ_T^j for $j \in [N]$.

LocalUpdate$(\widehat{\theta}_0, \tau, \lambda_2, \gamma)$ at j-th agent

for $q = 0, \cdots, \tau-1$ **do**
  (stochastic) gradient descent $\widehat{\theta}_{q+1} \leftarrow \widehat{\theta}_q - \lambda_2\gamma\nabla\widetilde{L}_j(\widehat{\theta}_q)$;
end for
return $\widehat{\theta}_\tau$;

practice, the server can select this group among the agents that are currently online and available. Each selected agent $j \in G_n$ performs local SGD updates on its model θ_n^j using its personal data set (denoted the loss function as $\widetilde{L}_j$). After the local update, each participating agent $j \in G_n$ begins local aggregation (Algorithm 2). In particular, agent j first selects a subset of agents $A_n \subseteq G_n$ using, for example, a ε-greedy sampling strategy (Zhang et al., 2021) (see Remark 6 for details) and then downloads their models (see Remark 7 for a discussion on data privacy vulnerabilities of this and other federated learning schemes). Agent j then evaluates all downloaded models $\theta_j^i, i \in A_n$, on its local dataset and obtains their corresponding losses $\widetilde{L}_j^i$. Using the losses $\widetilde{L}_j^i$, agent j calculates the consensus point m_j following (21) and updates its own model θ_n^j following equation (22). Finally, agent j updates its sampling likelihood vector P_n^j according to (23) for future communication rounds. As had already been suggested above, the model aggregation step in FedCBO is different from the one in most conventional federated learning algorithms. In FedCBO, models are aggregated locally on each device, whereas conventional federated learning algorithms average models at the global server.

Remark 5 *In practice, each agent only has access to finite data samples and thus their empirical loss functions may actually differ from those of the other agents in the same cluster. This is a different setting from our theoretical assumption that agents in the same cluster have the same objective function. On the experimental side, our numerical results show that when agents in the same cluster have similar but still different objective functions, our FedCBO algorithm still works well. On the theoretical side, to generalize our theoretical results to more practical settings we would need to study the case where agents in the same cluster have different objective functions that are nonetheless slight perturbations of a common underlying "true" loss function. This is closely related to analyzing the sensitivity of our proposed interacting particle system to perturbations of the objective functions. We leave the detailed modeling and study of this more realistic and difficult setting to future work.*

Algorithm 2 LocalAggregation(agent j)

Input: Agent j's model $\theta_n^j \in \mathbb{R}^d$; Participating devices at n iteration G_n; Sampling likelihood $P_n^j \in \mathbb{R}^{N-1}$; CBO system hyperparameters λ_1, α; Discretization step size γ; Random sample proportion $\varepsilon \in (0,1)$; Number of models downloaded M;

$A_n \leftarrow \boldsymbol{\varepsilon}$**-greedySampling**$(P_n^j, G_n, M)$;

Agent j downloads models θ_n^i for $i \in A_n$;

Evaluate models θ_n^i on agent j's data set respectively and denote the corresponding loss as $\widetilde{L}_j^i$;

Calculate consensus point m_j by

$$m_j \leftarrow \frac{1}{\sum_{i\in A_n} \mu_j^i} \sum_{i\in A_n} \theta_n^i \mu_j^i, \qquad \text{with } \mu_j^i = \exp(-\alpha \widetilde{L}_j^i) \tag{21}$$

Update agent j's model by

$$\theta_{n+1}^j \leftarrow \theta_n^j - \lambda_1 \gamma \left(\theta_n^j - m_j\right), \tag{22}$$

Update sampling likelihood P_n^j by

$$P_{n+1}^{j,i} \leftarrow P_n^{j,i} + \left(\widetilde{L}_j^j - \widetilde{L}_j^i\right), \qquad \text{for } i \in A_n \tag{23}$$

Output: $\theta_{n+1}^j, P_{n+1}^j$

$\underline{\boldsymbol{\varepsilon}\textbf{-greedySampling}}(P_n^j, G_n, M)$

Randomly sample $\varepsilon * M$ number of agents from G_n, denoted as A_n^1;

Select $(1-\varepsilon) * M$ numbers of agents in $G_n \backslash A_n^1$ with top value $P_j^{j,i}, i \in G_n \backslash A_n^1$, denoted as A_n^2;

return $A_n = A_n^1 \cup A_n^2$

Remark 6 (ε-greedy sampling strategy) *For each agent j, we use ε-greedy sampling scheme as in (Zhang et al., 2021) to select which models to download from other agents. In particular, we maintain a matrix P consisting of row vectors $p^j = (p^{j,1}, \ldots, p^{j,N})$, where $p^{j,i}$ measures the likelihood of agent j downloading model θ^i. Initially, we set P to be the zero matrix, i.e., each model has an equal chance of being selected by any other agent. During each federated iteration, we update P by (23). Since the number of allowed downloaded models M is much smaller than the total number of agents N, we may benefit from extra exploration by randomly selecting ε proportion of agents and then selecting the remaining proportion of agents based on the top sampling likelihoods according to P. After a few iterations, the likelihood matrix P should become more accurate in identifying similar agents. Therefore, we gradually decrease the value of ε to control the random exploration rate.*

Remark 7 *Given the data privacy constraints motivating federated learning methods, individual agents must not share their local data with the global server or other agents. In standard federated training protocols, agents typically exchange either the gradients or the parameters of the models that are being trained on their respective local data sets. It should be noted that the sharing of*

gradients or model parameters is not entirely secure, as it is possible for the private training data to be retrieved from publicly shared gradients (Zhu and Han, 2020; Geiping et al., 2020; Zhao et al., 2020; Yin et al., 2021; Huang et al., 2021; Li et al., 2022) or shared parameters (Haim et al., 2022; Buzaglo et al., 2023). While the recovery of data from parameters is still quite difficult, our FedCBO framework may benefit from the incorporation of other privacy protection techniques like DP-SGD in (Abadi et al., 2016) and secure aggregation in (Bonawitz et al., 2017). A detailed implementation of these mechanisms is out of the scope of this paper.

2.4 Experiments

In this section, we present an empirical study of our proposed FedCBO algorithm and assess its performance in relation to other state-of-the-art methodologies designed for the clustered federated learning setting[3].

Dataset Setup: We follow the approach used in (Ghosh et al., 2020) to create a clustered FL setting that is based on the standard MNIST dataset (Lecun et al., 1998). Precisely, we begin with the original MNIST dataset containing $60,000$ training images and $10,000$ test images. We augment this dataset by applying 0, 90, 180, and 270 degrees of rotation to each image, producing in this way $k = 4$ clusters, each of them corresponding to one of the 4 rotation angles. For training, we randomly partition the total number of training images $60000k$ into N agent machines so that each agent holds $n = \frac{60000k}{N}$ images, all coming *from the same rotation angle*. For inference, we do not split the test data. Therefore, the model from each local agent will be evaluated on 10000 rotated test images according to the cluster to which the agent belongs to. A few examples of the rotated MNIST dataset are shown in Fig. 2.

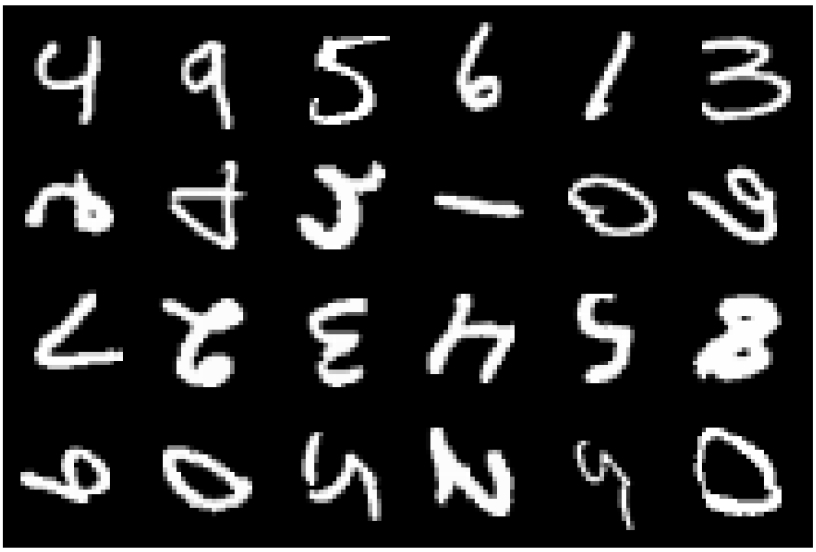

Figure 2: Examples of rotated MNIST dataset. Each row contains a collection of samples from one particular rotation.

Baselines & Implementations: We compare our FedCBO algorithm with three baseline algorithms: IFCA (Ghosh et al., 2020), FedAvg (McMahan et al., 2017), and a *local model training* scheme. We use fully connected neural networks with a single hidden layer of size 200 and ReLU

3. Implementation of our experiments is open sourced at `https://github.com/SixuLi/FedCBO`.

activation as base model. We set the total number of agents $N = 1200$ and the number of communication rounds $T = 100$. In each communication round, all agents participate in the training, i.e., $|G_n| = N$ for all n. When an agent trains its own model on its local dataset (local update step in each round), we run $\tau = 10$ epochs of stochastic gradient descent (SGD) with a learning rate of $\gamma = 0.1$ and momentum of 0.9. In what follows, we provide some implementation details for each baseline algorithm:

- **FedCBO:** We set the model download budget $M = 200$. We choose the hyperparameters $\lambda_1 = 10$, $\lambda_2 = 1$ and $\alpha = 10$. For the ε-greedy sampling, we use the decay scheme $\varepsilon(n) = \max\{0.5 - 0.01n, 0.1\}$, i.e., the initial random sample proportion $\varepsilon = 0.5$. This parameter is decreased by 0.01 at each communication round until it reaches the threshold 0.1.

- **IFCA (Ghosh et al., 2020):** We set the number of models initialized at the global server to equal the number of underlying clusters ($k = 4$) as suggested in (Ghosh et al., 2020) for a fair comparison. This should provide the best result for the IFCA algorithm.

- **FedAvg (McMahan et al., 2017):** The algorithm tries to train a single global model that works for all the local distributions. Hence, in the model aggregation step, the local models trained by the agents are averaged to obtain the updated global model.

- **Local model training:** Agents train their own model using only their local data and with no communication with the global server or to any of the other agents. To ensure a fair comparison, each agent trains its model for a total of $T * \tau = 1000$ epochs.

Remark 8 *Notice that in FedCBO we do not have to input the number of underlying clusters k, in contrast to IFCA where we need to input this value or an estimate thereof.*

For FedCBO and the local model training scheme we perform inference by testing the trained model on the test data with the same distribution as their training data (i.e., data points with the same rotation). For IFCA and FedAvg, following (Ghosh et al., 2020), we run inference on all learned models (k models for IFCA and one model for FedAvg) on each data distribution and calculate the accuracy of the model that produces the smallest loss value. We conduct experiments with 5 different random seeds for all the algorithms and report the average accuracy and standard deviation.

Experimental Results: The test results are summarized in Table 1. We observe that our FedCBO algorithm outperforms the three baseline methods. Although the IFCA algorithm can gradually estimate the cluster identities of users correctly and average over users' models that are estimated to belong to the same clusters, it gives models the same weights during the model aggregation step, thus preventing the algorithm from further utilizing the relative similarities between different models. In contrast, as we run the FedCBO algorithm we observe that, during the (local) model aggregation steps, agents successfully select models from other users having the same data distribution (as discussed in Remark 9) and assign different importance (weights) to the downloaded models using (21) during the aggregation steps. In this way, each user can better utilize the most beneficial models from others. As pointed out in (Ghosh et al., 2020), the FedAvg baseline performs worse than FedCBO and IFCA as it tries to fit heterogeneous data using a single model and thus cannot provide cluster-wise predictions. Since each agent only stores a small amount of data, the local model training scheme can easily overfit to the local dataset. This explains why it produces the worst performance among all other methodologies.

Table 1: Test accuracy ± standard deviation % on rotated MNIST.

FedCBO	IFCA	FedAvg	Local
$\mathbf{96.51 \pm 0.04}$	94.44 ± 0.01	85.50 ± 0.19	81.27 ± 0.02

Remark 9 *To verify the correctness of the sampling scheme in FedCBO, we define the successful selection rate (SR) for agent j at iteration n as follows:*

$$SR_n^j := \frac{\textit{Number of selected agents in the same cluster as agent } j}{\textit{Total number of selected agents}}, \tag{24}$$

where the total number of selected agents equals the model download budget M. During the FedCBO algorithm, we calculate the average successful selection rate $SR_n := \frac{1}{N}\sum_{j=1}^{N} SR_n^j$ at each communication round n, which corresponds to the blue curve in Fig. 3. Meanwhile, when implementing the ε-greedy sampling, we set the random exploration proportion ε to 0.5 at $n = 0$ and use a decay scheme of $\varepsilon(n) = \max\{0.5 - 0.01n, 0.1\}$. Hence we can calculate the oracle expected successful selection rate at each round: this is shown as the orange curve in Fig. 3. We note that the empirical average successful SR (blue curve) is very close to the best expected successful SR (orange curve). This indicates that our FedCBO algorithm can successfully identify the agents with the same data distributions. We leave the task of designing better sampling strategies to close the gap between empirical successful SR and oracle successful SR to future work.

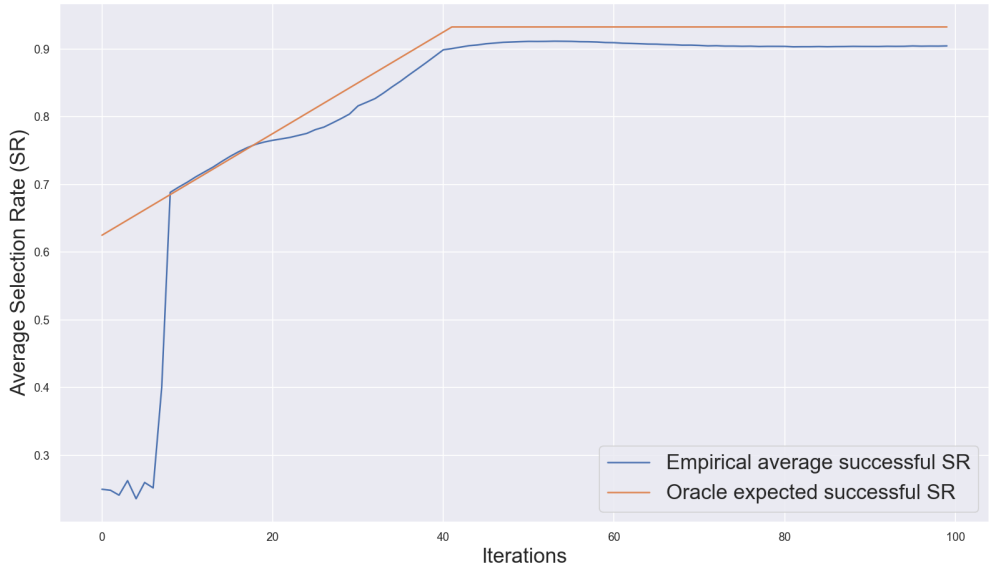

Figure 3: Average successful selection rate (SR) at each communication round.

3. Well-posedness of Finite Particle System and Mean-field System

3.1 Well-posedness of the Microscopic Model

In this section, we prove Theorem 1. Before that, we first introduce some useful notation and preliminary lemmas. In this section, to make the notation simpler we write the solution of the

interacting particle system (4) as $\{\boldsymbol{\theta}^N \,|\, t \geq 0\}$ with $\boldsymbol{\theta}_t^N := \left(\theta^{(1,N)}, \cdots, \theta^{(N,N)}\right)^T$ without distinguishing between agents of cluster 1 or 2. Moreover, we assume the objective functions L_1 and L_2 to be the same, and we denote them as L.[4] For an arbitrary but fixed $N \in \mathbb{N}$, we rewrite the system (4) as

$$d\boldsymbol{\theta}_t^N = -\boldsymbol{F}_N(\boldsymbol{\theta}_t^N)dt + \boldsymbol{M}_N(\boldsymbol{\theta}_t^N)d\boldsymbol{B}_t^N, \tag{25}$$

where $\boldsymbol{B} := \left(B^{(1,N)}, \cdots, B^{(N,N)}\right)^T$ is the standard Brownian motion in $\mathbb{R}^{Nd}$, and

$$\begin{aligned}
\boldsymbol{F}_N(\boldsymbol{\theta}) &:= \left(F_N^1(\boldsymbol{\theta}), \cdots, F_N^N(\boldsymbol{\theta})\right)^T \in \mathbb{R}^{Nd}, \quad \text{with } F_N^i(\boldsymbol{\theta}) := \lambda_1 G_N^i(\boldsymbol{\theta}) + \lambda_2 \nabla L(\theta^i), \\
\boldsymbol{M}_N(\boldsymbol{\theta}) &:= \operatorname{diag}\left(|M_N^1(\boldsymbol{\theta})|\mathbb{I}_d, \cdots, |M_N^N(\boldsymbol{\theta})|\mathbb{I}_d\right) \in \mathbb{R}^{Nd\times Nd}, \\
\text{with } M_N^i(\boldsymbol{\theta}) &:= \sigma_1|G_N^i(\boldsymbol{\theta})| + \sigma_2|\nabla L(\theta^i)|, \quad \text{and} \quad G_N^i(\boldsymbol{\theta}) := \frac{\sum_{j\neq i}(\theta^i - \theta^j)\omega_L^\alpha(\theta^j)}{\sum_j \omega_L^\alpha(\theta^j)}.
\end{aligned}$$

As in Assumption 1, we assume that the gradient of objective function L satisfies the following conditions for all $\theta, \hat{\theta} \in \mathbb{R}^d$:

$$|\nabla L(\theta) - \nabla L(\hat{\theta})| \leq M_{\nabla L}|\theta - \hat{\theta}| \qquad \text{and} \qquad |\nabla L(\theta)| \leq C_{\nabla L}, \tag{26}$$

with constants $M_{\nabla L}, C_{\nabla L} > 0$. Under these conditions on ∇L, we can show that G_N^i, for $i \in [N]$, is locally Lipschitz continuous and has linear growth. Consequently, $\boldsymbol{F}_N$ and $\boldsymbol{M}_N$ are locally Lipschitz continuous and have linear growth. We summarize the previous observations in the following lemma.

Lemma 1 ((Carrillo et al., 2018, Lemma 2.1)) *Let $N \in \mathbb{N}, \alpha, r > 0$ be arbitrary. Then, for any $\boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \in \mathbb{R}^{Nd}$ with $|\boldsymbol{\theta}|, |\hat{\boldsymbol{\theta}}| \leq r$ and all $i \in [N]$, it holds*

$$\begin{aligned}
|G_N^i(\boldsymbol{\theta}) - G_N^i(\hat{\boldsymbol{\theta}})| &\leq |\theta^i - \hat{\theta}^i| + \left(1 + \frac{2c_r}{N}\sqrt{N|\hat{\theta}^i|^2 + |\hat{\boldsymbol{\theta}}|^2}\right)|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|, \\
|G_N^i(\boldsymbol{\theta})| &\leq |\theta^i| + |\boldsymbol{\theta}|,
\end{aligned}$$

where $c_r := \alpha C_{\nabla L}\exp(\alpha\|L - \underline{L}\|_{L^\infty(B_r)})$ and $B_r := \{\theta \in \mathbb{R}^d \,|\, |\theta| \leq r\}$.

Combining the above lemma with the assumptions (26) we deduce the following corollary.

Corollary 1 *Under the same assumptions as in Lemma 1, it holds*

$$\begin{aligned}
|F_N^i(\boldsymbol{\theta}) - F_N^i(\hat{\boldsymbol{\theta}})| &\leq (\lambda_1 + \lambda_2 M_{\nabla L})\,|\theta^i - \hat{\theta}^i| + \lambda_1\left(1 + \frac{2c_r}{N}\sqrt{N|\hat{\theta}^i|^2 + |\hat{\boldsymbol{\theta}}|^2}\right)|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|, \\
|M_N^i(\boldsymbol{\theta}) - M_N^i(\hat{\boldsymbol{\theta}})| &\leq (\sigma_1 + \sigma_2 M_{\nabla L})\,|\theta^i - \hat{\theta}^i| + \lambda_1\left(1 + \frac{2c_r}{N}\sqrt{N|\hat{\theta}^i|^2 + |\hat{\boldsymbol{\theta}}|^2}\right)|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|, \\
|M_N^i(\boldsymbol{\theta})| &\leq \sigma_1(|\theta^i| + |\boldsymbol{\theta}|) + \sigma_2 C_{\nabla L},
\end{aligned} \tag{27}$$

where $c_r := \alpha C_{\nabla L}\exp(\alpha\|L - \underline{L}\|_{L^\infty(B_r)})$ and $B_r := \{\theta \in \mathbb{R}^d \,|\, |\theta| \leq r\}$.

Proof Based on Lemma 1 and the assumptions on ∇L, we can compute

$$\begin{aligned}
|F_N^i(\boldsymbol{\theta}) - F_N^i(\hat{\boldsymbol{\theta}})| &= |\lambda_1(G_N^i(\boldsymbol{\theta}) - G_N^i(\hat{\boldsymbol{\theta}})) + \lambda_2(\nabla L(\theta^i) - \nabla L(\hat{\theta}^i))| \\
&\leq \lambda_1|G_N^i(\boldsymbol{\theta}) - G_N^i(\hat{\boldsymbol{\theta}})| + \lambda_2|\nabla L(\theta^i) - \nabla L(\hat{\theta}^i)| \\
&\leq (\lambda_1 + \lambda_2 M_{\nabla L})\,|\theta^i - \hat{\theta}^i| + \lambda_1\left(1 + \frac{2c_r}{N}\sqrt{N|\hat{\theta}^i|^2 + |\hat{\boldsymbol{\theta}}|^2}\right)|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|.
\end{aligned}$$

4. For the case that objective functions L_1 and L_2 are different and two classes of agents are distinguishable, the proof of the well-posedness of the microscopic model can be easily adapted.

$$
\begin{aligned}
|M_N^i(\boldsymbol{\theta}) - M_N^i(\hat{\boldsymbol{\theta}})| &= \sigma_1 \left(|G_N^i(\boldsymbol{\theta})| - |G_N^i(\hat{\boldsymbol{\theta}})| \right) + \sigma_2 \left(|\nabla L(\theta^i)| - |\nabla L(\hat{\theta}^i)| \right) \\
&\leq \sigma_1 |G_N^i(\boldsymbol{\theta}) - G_N^i(\hat{\boldsymbol{\theta}})| + \sigma_2 |\nabla L(\theta^i) - \nabla L(\hat{\theta}^i)| \\
&\leq (\sigma_1 + \sigma_2 M_{\nabla L}) \, |\theta^i - \hat{\theta}^i| + \lambda_1 \left(1 + \frac{2c_r}{N} \sqrt{N|\hat{\theta}^i|^2 + |\hat{\boldsymbol{\theta}}|^2} \right) |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|. \\
|M_N^i(\boldsymbol{\theta})| &= \sigma_1 |G_N^i(\boldsymbol{\theta})| + \sigma_2 |\nabla L(\theta^i)| \leq \sigma_1 (|\theta^i| + |\boldsymbol{\theta}|) + \sigma_2 C_{\nabla L}.
\end{aligned}
$$

■

Now we may invoke standard results on the existence of strong solutions of SDEs from (Durrett, 2018) to prove Theorem 1.

Proof [Proof of Theorem 1] We first show that there exists a constant $b_N > 0$ such that

$$
-2\boldsymbol{\theta} \cdot \boldsymbol{F}_N(\boldsymbol{\theta}) + \operatorname{trace}(\boldsymbol{M}_N \boldsymbol{M}_N^T)(\boldsymbol{\theta}) \leq b_N (1 + |\boldsymbol{\theta}|^2). \tag{28}
$$

By Corollary 1, the following holds:

$$
\begin{aligned}
-\theta^i \cdot F_N^i(\boldsymbol{\theta}) &= -\theta^i \cdot \left(\lambda_1 \frac{\sum_{j \neq i} (\theta^i - \theta^j) \omega_L^\alpha(\theta^j)}{\sum_j \omega_L^\alpha(\theta^j)} + \lambda_2 \nabla L(\theta^i) \right) \leq -\lambda_1 |\theta^i|^2 + (\lambda_1 |\boldsymbol{\theta}| + \lambda_2 C_{\nabla L}) |\theta^i| \\
|M_N^i(\boldsymbol{\theta})|^2 &\leq \left| \sigma_1 (|\theta^i| + |\boldsymbol{\theta}|) + \sigma_2 C_{\nabla L} \right|^2 \leq 4\sigma_1^2 (|\theta^i|^2 + |\boldsymbol{\theta}|^2) + 2\sigma_2^2 C_{\nabla L}^2.
\end{aligned}
$$

Then one can obtain

$$
\begin{aligned}
-2\boldsymbol{\theta} \cdot \boldsymbol{F}_N(\boldsymbol{\theta}) + \operatorname{trace}(\boldsymbol{M}_N \boldsymbol{M}_N^T)(\boldsymbol{\theta}) &= \sum_i \left(-2\theta^i \cdot F_N^i(\theta) + d |M_N^i(\theta)|^2 \right) \\
&\leq \sum_i \big(-2\lambda_1 |\theta^i|^2 + 2 \left(\lambda_1 |\boldsymbol{\theta}| + \lambda_2 C_{\nabla L} \right) |\theta^i| \\
&\qquad + 4d\sigma_1^2 \left(|\theta^i|^2 + |\boldsymbol{\theta}|^2 \right) + 2d\sigma_2^2 C_{\nabla L}^2 \big) \\
&\leq -2\lambda_1 |\boldsymbol{\theta}|^2 + \lambda_1 N |\boldsymbol{\theta}|^2 + \lambda_1 |\boldsymbol{\theta}|^2 + \lambda_2 C_{\nabla L} N + \lambda_2 C_{\nabla L} |\boldsymbol{\theta}|^2 \\
&\quad + 4d\sigma_1^2 |\boldsymbol{\theta}|^2 + 4d\sigma_1^2 N |\boldsymbol{\theta}|^2 + 2d\sigma_2^2 C_{\nabla L}^2 N \\
&= \left(\lambda_1 (N-1) + \lambda_2 + 4d\sigma_1^2 (N+1) \right) |\boldsymbol{\theta}|^2 \\
&\quad + \left(\lambda_2 C_{\nabla L} + 2d\sigma_2^2 C_{\nabla L}^2 \right) N \\
&\leq b_N \left(1 + |\boldsymbol{\theta}|^2 \right).
\end{aligned}
$$

Together with the local Lipschitz continuity and linear growth of $\boldsymbol{F}_N$ and $\boldsymbol{M}_N$, we deduce the desired result by applying Theorem 3.1 in (Durrett, 2018). ■

Remark 10 *From the estimate* (28)*, we can obtain a uniform bound on the second moment of* $\boldsymbol{\theta}_t^N$*. In particular, by the Itô formula, we have*

$$
\frac{d}{dt} \mathbb{E} |\boldsymbol{\theta}_t^N|^2 = -\mathbb{E}[\boldsymbol{\theta}_t^N \cdot \boldsymbol{F}_N(\boldsymbol{\theta}_t^N)] + \mathbb{E}[trace(\boldsymbol{M}_N \boldsymbol{M}_N^T)(\boldsymbol{\theta}_t^N)] \leq b_N \left(1 + \mathbb{E} |\boldsymbol{\theta}_t^N|^2 \right).
$$

Therefore, Grönwall inequality yields

$$
\mathbb{E} |\boldsymbol{\theta}_t^N|^2 \leq \exp(b_N t) \mathbb{E} |\boldsymbol{\theta}_0^N|^2 + b_N \int_0^t \exp(b_N (t-s)) ds \qquad \textit{for all } t \geq 0,
$$

i.e., the solution exists globally in time for each fixed $N \in \mathbb{N}$*.*

3.2 Well-posedness of Mean-field System

In this section we prove Theorem 2. We present the details in the case in which the loss functions are assumed to be bounded. The proof for the quadratic growth case is similar, and we refer the reader to Appendix B for more details.

Before we prove Theorem 2, we first need a "Lipschitz continuity" property for the operators $m^\alpha_{L_k}$ for $k = 1, 2$. The proof of the next lemma can be found in Appendix B.

Lemma 2 *Let objective functions L_1, L_2 satisfy Assumption 1 and let $\nu^1, \nu^2 \in \mathcal{C}([0,T], \mathcal{P}_2(\mathbb{R}^d))$ be such that $\sup_{t\in[0,T]} \int |\theta|^4 d\nu_t^1 \leq K$, $\sup_{t\in[0,T]} \int |\theta|^4 d\nu_t^2 \leq K$. Let us denote by $\overline{L}_1, \overline{L}_2$ the supremum of each of the loss functions. Let $w_1, w_2 > 0$ be such that $w_1 + w_2 = 1$, and let $\nu := w_1\nu^1 + w_2\nu^2$. Then, for all $s, t \in (0,T)$, the following stability estimates hold*

$$\left|m^\alpha_{L_k}[\nu_t] - m^\alpha_{L_k}[\nu_s]\right| \leq C\left(\sqrt{w_1}W_2\left(\nu_t^1, \nu_s^1\right) + \sqrt{w_2}W_2\left(\nu_t^2, \nu_s^2\right)\right), \tag{29}$$

for $k = 1, 2$ and for a constant C that depends only on $\alpha, C_{\nabla L_k}$ and K.

With the above lemma in hand, the proof of Theorem 2 reduces to a careful application of the Leray-Schauder fixed point theorem which follows similar steps as in (Carrillo et al., 2018, Theorem 3.1).

Proof [Proof of Theorem 2] **Step 1:** For a given pair $u^1, u^2 \in \mathcal{C}([0,T], \mathbb{R}^d)$, we may apply standard theory of SDEs (see, e.g., Chapter 6.2 of (Arnold, 1974)) to conclude that there exists a unique strong solution to the SDE

$$dY_t^1 = -\lambda_1\left(Y_t^1 - u_t^1\right)dt - \lambda_2\nabla L_1(Y_t^1)dt + \sigma_1\left|Y_t^1 - u_t^1\right|dB_t^1 + \sigma_2\left|\nabla L_1(Y_t^1)\right|d\widetilde{B}_t^1 \tag{30a}$$

$$dY_t^2 = -\lambda_1\left(Y_t^2 - u_t^2\right)dt - \lambda_2\nabla L_2(Y_t^2)dt + \sigma_1\left|Y_t^2 - u_t^2\right|dB_t^2 + \sigma_2\left|\nabla L_2(Y_t^2)\right|d\widetilde{B}_t^2, \tag{30b}$$

for initial conditions $Y_0^1 \sim \rho_0^1$ and $Y_0^2 \sim \rho_2$ (independent of each other), where $\rho_0^1, \rho_0^2 \in \mathcal{P}_4(\mathbb{R}^d)$. Let $\nu_t^1 = \mathrm{Law}(Y_t^1)$ and $\nu_t^2 = \mathrm{Law}(Y_t^2)$ be the laws of Y_t^1 and Y_t^2, respectively. Since the variables Y^1, Y^2 take values in $\mathcal{C}([0,T], \mathbb{R}^d)$, it follows that $\nu^1, \nu^2 \in \mathcal{C}([0,T], \mathcal{P}(\mathbb{R}^d))$. Moreover, ν_t^1, ν_t^2 satisfy the following system of Fokker-Planck equations (in weak form):

$$\frac{d}{dt}\int \varphi d\nu_t^1 = \int\left[\left(-\lambda_1(\theta - u_t^1) - \lambda_2\nabla L_1(\theta)\right)\cdot\nabla\varphi + \left(\frac{\sigma_1^2}{2}\left|\theta - u_t^1\right|^2 + \frac{\sigma_2^2}{2}\left|\nabla L_1(\theta)\right|^2\right)\Delta\varphi\right]d\nu_t^1, \tag{31a}$$

$$\frac{d}{dt}\int \varphi d\nu_t^2 = \int\left[\left(-\lambda_1(\theta - u_t^2) - \lambda_2\nabla L_2(\theta)\right)\cdot\nabla\varphi + \left(\frac{\sigma_1^2}{2}\left|\theta - u_t^2\right|^2 + \frac{\sigma_2^2}{2}\left|\nabla L_2(\theta)\right|^2\right)\Delta\varphi\right]d\nu_t^2, \tag{31b}$$

for all $\varphi \in \mathcal{C}_c^2(\mathbb{R}^d)$. Let us consider the product space $\mathcal{C}([0,T], \mathbb{R}^d) \times \mathcal{C}([0,T], \mathbb{R}^d)$ endowed with the norm $\|\cdot\|$ defined as

$$\|(f,g)\| := \|f\|_\infty + \|g\|_\infty,$$

where $f, g \in \mathcal{C}([0,T], \mathbb{R}^d)$ and $\|f\|_\infty := \sup_{t\in[0,T]}|f_t|$. Notice that $m^\alpha_{L_k}[\nu] \in \mathcal{C}([0,T], \mathbb{R}^d)$ for $k = 1, 2$ and thus we can define the map

$$\begin{aligned}\mathcal{T} : \mathcal{C}([0,T], \mathbb{R}^d) \times \mathcal{C}([0,T], \mathbb{R}^d) &\longrightarrow \mathcal{C}([0,T], \mathbb{R}^d) \times \mathcal{C}([0,T], \mathbb{R}^d),\\ (u^1, u^2) &\longrightarrow \mathcal{T}(u^1, u^2) = (m^\alpha_{L_1}[\nu], m^\alpha_{L_2}[\nu]),\end{aligned}$$

where $\nu = w_1\nu^1 + w_2\nu^2$. Next, we show that $\mathcal{T}$ has a unique fixed point.

Step 2: First, we show that $\mathcal{T}$ is compact, i.e., any bounded sequence $\{(f_n, g_n)\}_{n\in\mathbb{N}}$ in $\mathcal{C}([0,T],\mathbb{R}^d)\times\mathcal{C}([0,T],\mathbb{R}^d)$ is precompact. Let $(\varphi_n,\psi_n) := \mathcal{T}(f_n,g_n)$. It is sufficient to show that each of the sequences $\{\varphi_n\}_{n\in\mathbb{N}}$ and $\{\psi_n\}_{n\in\mathbb{N}}$ is precompact. We show the details for the sequence $\{\varphi_n\}_{n\in\mathbb{N}}$. Since $\rho_0^1,\rho_0^2\in\mathcal{P}_4(\mathbb{R}^d)$, standard theory of SDEs (see, e.g., Chapter 7 of (Arnold, 1974)) provides a fourth-order moment estimate for solutions to (30a) and (30b) of the form

$$\mathbb{E}|Y_t^1|^4 \le \big(1+\mathbb{E}|Y_0^1|^4\big)\exp(ct) \qquad \text{and} \qquad \mathbb{E}|Y_t^2|^4 \le \big(1+\mathbb{E}|Y_0^2|^4\big)\exp(ct),$$

for some constant $c>0$ only depending on $\|u^1\|_\infty$ and $\|u^2\|_\infty$. In particular,

$$\sup_{t\in[0,T]}\int|\theta|^4 d\nu_t^1, \sup_{t\in[0,T]}\int|\theta|^4 d\nu_t^2 \le K$$

for some $K<\infty$. On the other hand, for any $t>s$ in $(0,T)$, the Itô isometry and Cauchy-Schwarz inequality yield

$$\begin{aligned}
\mathbb{E}\left|Y_t^1-Y_s^1\right|^2 &= \mathbb{E}\bigg|\int_s^t\big(-\lambda_1\left(Y_\tau^1-u_\tau^1\right)-\lambda_2\nabla L_1(Y_\tau^1)\big)d\tau \\
&\qquad +\int_s^t\sigma_1\left|Y_\tau^1-u_\tau^1\right|dB_\tau^1+\int_s^t\sigma_2\left|\nabla L_1(Y_\tau^1)\right|d\widetilde{B}_\tau^1\bigg|^2 \\
&\le 2\mathbb{E}\bigg|\int_s^t\big(-\lambda_1\left(Y_\tau^1-u_\tau^1\right)-\lambda_2\nabla L_1(Y_\tau^1)\big)d\tau\bigg|^2 \\
&\qquad +2\mathbb{E}\bigg[\int_s^t\big(\sigma_1^2\left|Y_\tau^1-u_\tau^1\right|^2+\sigma_2^2\left|\nabla L_1(Y_\tau^1)\right|^2\big)d\tau\bigg] \\
&\le 4\lambda_1^2|t-s|\,\mathbb{E}\left[\int_s^t\left|Y_\tau^1-u_\tau^1\right|^2 d\tau\right]+4\lambda_2^2|t-s|\,\mathbb{E}\left[\int_s^t\left|\nabla L_1(Y_\tau^1)\right|^2 d\tau\right] \\
&\qquad +2\mathbb{E}\bigg[\int_s^t\big(\sigma_1^2|Y_\tau^1-u_\tau^1|^2+\sigma_2^2\left|\nabla L_1(Y_\tau^1)\right|^2\big)d\tau\bigg] \\
&\le 4\lambda_1^2\left(K+\|u^1\|_\infty^2\right)T|t-s|+4\lambda_2^2C_{\nabla L_1}^2T|t-s| \\
&\qquad +2\sigma_1^2\left(K+\|u^1\|_\infty^2\right)|t-s|+2\sigma_2^2C_{\nabla L_1}^2|t-s| \\
&:= C_1|t-s|.
\end{aligned}$$

Similarly, $\mathbb{E}\left|Y_t^2-Y_s^2\right|^2 \le C_2|t-s|$ for a constant $C_2>0$. Therefore, $W_2\left(\nu_t^1,\nu_s^1\right)\le c_1|t-s|^{\frac{1}{2}}$ and $W_2\left(\nu_t^2,\nu_s^2\right)\le c_2|t-s|^{\frac{1}{2}}$, for some constants $c_1,c_2>0$ only depending on $\|u^1\|_\infty$ and $\|u^2\|_\infty$. Applying Lemma 2, we obtain

$$\left|m_{L_1}^\alpha[\nu_t]-m_{L_1}^\alpha[\nu_s]\right| \le C(\sqrt{w_1}c_1+\sqrt{w_2}c_2)\,|t-s|^{\frac{1}{2}},$$

which proves that $t\to m_{L_1}^\alpha[\nu_t]$ is Hölder continuous with exponent $\frac{1}{2}$. From this we can conclude that $\{\varphi_n\}_{n\in\mathbb{N}}$ is precompact due to the compact embedding $\mathcal{C}^{0,1/2}([0,T],\mathbb{R}^d)\hookrightarrow\mathcal{C}([0,T],\mathbb{R}^d)$, where $\mathcal{C}^{0,1/2}([0,T],\mathbb{R}^d)$ is the space of $\frac{1}{2}$-Hölder continuous functions from $[0,T]$ into $\mathbb{R}^d$.

Step 3: Next, we verify the conditions in the Leray-Schauder fixed point theorem. For that purpose, suppose the pair $(u^1,u^2)\in\mathcal{C}([0,T],\mathbb{R}^d)\times\mathcal{C}([0,T],\mathbb{R}^d)$ satisfies $(u^1,u^2)=\tau\mathcal{T}(u^1,u^2)$ for some $\tau\in[0,1]$. In particular, there exist $\nu_1,\nu_2\in\mathcal{C}([0,T],\mathcal{P}_2(\mathbb{R}^d))$ satisfying (31a) and (31b), respectively, such that $(u^1,u^2)=\tau(m_{L_1}^\alpha[\nu],m_{L_2}^\alpha[\nu])$, where $\nu=w_1\nu_1+w_2\nu_2$. Due to the boundedness

assumption on L_1[5], we have for all $t \in (0,T)$

$$\begin{aligned}\left|u_t^1\right|^2 &= \tau^2 \left|m_{L_1}[\nu]\right|^2 \leq \tau^2 \exp\left(\alpha(\overline{L_1} - \underline{L_1})\right) \int |x|^2 d\nu_t \\ &= \tau^2 \exp(\alpha(\overline{L_1} - \underline{L_1}))\bigg(w_1 \int |x|^2 d\nu_t^1 + w_2 \int |x|^2 d\nu_t^2\bigg).\end{aligned} \tag{32}$$

A computation of the second moment of ν_t^1 gives

$$\begin{aligned}\frac{d}{dt}\int |\theta|^2 d\nu_t^1 &= \int \bigg[\big(-\lambda_1\left(\theta - u_t^1\right) - \lambda_2 \nabla L_1(\theta)\big)\cdot 2\theta + 2d\left(\frac{\sigma_1^2}{2}\left|\theta - u_t^1\right|^2 + \frac{\sigma_2^2}{2}\left|\nabla L_1(\theta)\right|^2\right)\bigg] d\nu_t^1 \\ &= \int \bigg[-2\lambda_1|\theta|^2 + 2\lambda_1 \theta \cdot u_t^1 - 2\lambda_2 \theta \cdot \nabla L_1(\theta) \\ &\qquad + d\sigma_1^2\left(|\theta|^2 - 2\theta\cdot u_t^1 + |u_t^1|^2\right) + d\sigma_2^2\left|\nabla L_1(\theta)\right|^2\bigg] d\nu_t^1 \\ &\leq \int \left[(d\sigma_1^2 - 2\lambda_1 + |\chi| + \lambda_2)|\theta|^2 + (d\sigma_1^2 + |\chi|)|u_t^1|^2 + (\lambda_2 + d\sigma_2^2)|\nabla L_1(\theta)|^2\right] d\nu_t^1 \\ &\leq (d\sigma_1^2 - 2\lambda_1 + |\chi| + \lambda_2)\int |\theta|^2 d\nu_t^1 + (d\sigma_1^2 + |\chi|)|u_t^1|^2 + (\lambda_2 + d\sigma_2^2)C_{\nabla L_1}^2 \\ &\leq (d\sigma_1^2 + |\chi| + \lambda_2)\big(1 + \exp(\alpha(\overline{L_1} - \underline{L_1}))\big)\bigg(\int |\theta|^2 d\nu_t^1 + \int |\theta|^2 d\nu_t^2\bigg) + (\lambda_2 + d\sigma_2^2)C_{\nabla L_1}^2,\end{aligned}$$

where $\chi := \lambda_1 - d\sigma_1^2$. Similarly,

$$\frac{d}{dt}\int |\theta|^2 d\nu_t^2 \leq (d\sigma_1^2 + |\chi| + \lambda_2)\big(1 + \exp(\alpha(\overline{L_2} - \underline{L_2}))\big)\bigg(\int |\theta|^2 d\nu_t^1 + \int |\theta|^2 d\nu_t^2\bigg) + (\lambda_2 + d\sigma_2^2)C_{\nabla L_2}^2.$$

Adding the above two inequalities we conclude that

$$\frac{d}{dt}\bigg(\int |\theta|^2 d\nu_t^1 + \int |\theta|^2 d\nu_t^2\bigg) \leq C_1\bigg(\int |\theta|^2 d\nu_t^1 + \int |\theta|^2 d\nu_t^2\bigg) + C_2,$$

for some constants $C_1, C_2 > 0$. Using Grönwall's inequality we obtain

$$\int |\theta|^2 d\nu_t^1 + \int |\theta|^2 d\nu_t^2 \leq \bigg(\int |\theta|^2 d\nu_0^1 + \int |\theta|^2 d\nu_0^2\bigg)\exp(C_1 t) + \frac{C_2}{C_1}\big(\exp(C_1 t) - 1\big).$$

Then, from (32), we conclude that there is a constant $q_1 > 0$ such that $\|u^1\|_\infty < q_1$. A similar bound holds for $\|u^2\|_\infty$, i.e., there is a constant $q_2 > 0$ such that $\|u^2\|_\infty < q_2$. Hence, $\|(u^1, u^2)\| = \|u^1\|_\infty + \|u^2\|_\infty < q_1 + q_2$. We may now invoke the Leray-Schauder fixed point theorem (Section 9.2 in (Evans, 2010)) to conclude that there exists a fixed point (u^1, u^2) for the mapping $\mathcal{T}$ and thereby a solution of (30a) and (30b).

Step 4: As for uniqueness, we first note that a fixed point (u^1, u^2) of $\mathcal{T}$ must satisfy $\|(u^1, u^2)\| < q$. Hence, the fourth-order moment estimates provided in **Step 2** hold and $\sup_{t\in[0,T]} \int |x|^4 d\nu_t^k \leq K < \infty$ for $k = 1, 2$. Now suppose we have two fixed points (u^1, u^2) and $(\hat{u}^1, \hat{u}^2)$ with

$$\left\|\left(u^1, u^2\right)\right\|, \left\|\left(\hat{u}^1, \hat{u}^2\right)\right\| < q, \qquad \sup_{t\in[0,T]} \int |\theta|^4 d\nu_t^k, \sup_{t\in[0,T]} \int |\theta|^4 d\hat{\nu}_t^k \leq K \quad \text{for } k = 1, 2,$$

5. The proof for the case in which L_1 has quadratic growth at infinity is provided in Appendix B.2.

and consider their corresponding processes $(Y_t^1, Y_t^2), (\hat{Y}_t^1, \hat{Y}_t^2)$, which satisfy (30a) and (30b) with the same Brownian motions. Taking the differences $z_t^k := Y_t^k - \hat{Y}_t^k$ for $k = 1, 2$, we obtain

$$\begin{aligned} z_t^k = z_0^k &+ \int_0^t \left(-\lambda_1 z_s^k + \lambda_1 \left(u_s^k - \hat{u}_s^k\right) - \lambda_2 \left(\nabla L_k(Y_s^k) - \nabla L_k(\hat{Y}_s^k)\right)\right) ds \\ &+ \sigma_1 \int_0^t \left(\left|Y_s^k - u_s^k\right| - \left|\hat{Y}_s^k - \hat{u}_s^k\right|\right) dB_s^k + \sigma_2 \int_0^t \left(\left|\nabla L_k(Y_s^k)\right| - \left|\nabla L_k(\hat{Y}_s^k)\right|\right) d\widetilde{B}_s^k. \end{aligned}$$

Squaring both sides, taking expectations, and using Itô isometry we obtain

$$\begin{aligned} \mathbb{E}|z_t^k|^2 &= \mathbb{E}\Bigg[z_0^k + \int_0^t \left(-\lambda_1 z_s^k + \lambda_1 \left(u_s^k - \hat{u}_s^k\right) - \lambda_2 \left(\nabla L_k(Y_s^k) - \nabla L_k(\hat{Y}_s^k)\right)\right) ds \\ &\quad + \sigma_1 \int_0^t \left(\left|Y_s^k - u_s^k\right| - \left|\hat{Y}_s^k - \hat{u}_s^k\right|\right) dB_s^k + \sigma_2 \int_0^t \left(\left|\nabla L_k(Y_s^k)\right| - \left|\nabla L_k(\hat{Y}_s^k)\right|\right) d\widetilde{B}_s^k\Bigg]^2 \\ &\le 2\mathbb{E}|z_0^k|^2 + 2t\mathbb{E}\Bigg[\int_0^t \left(-\lambda_1 z_s^k + \lambda_1(u_s^k - \hat{u}_s^k) - \lambda_2 \left(\nabla L_k(Y_s^k) - \nabla L_k(\hat{Y}_s^k)\right)\right)^2 ds\Bigg] \\ &\quad + 2\sigma_1^2 \mathbb{E}\Bigg[\int_0^t \left(\left|Y_s^k - u_s^k\right| - \left|\hat{Y}_s^k - \hat{u}_s^k\right|\right)^2 ds\Bigg] + 2\sigma_2^2 \mathbb{E}\Bigg[\int_0^t \left(\left|\nabla L_k(Y_s^k)\right| - \left|\nabla L_k(\hat{Y}_s^k)\right|\right)^2 ds\Bigg] \\ &\le 2\mathbb{E}|z_0^k|^2 + 6t\lambda_1^2 \int_0^t \mathbb{E}|z_s^k|^2 ds + 6t\lambda_1^2 \int_0^t \left|u_s^k - \hat{u}_s^k\right|^2 ds + 6t\lambda_2^2 \int_0^t \mathbb{E}\left[\nabla L_k(Y_s^k) - \nabla L_k(\hat{Y}_s^k)\right]^2 ds \\ &\quad + 2\sigma_1^2 \mathbb{E}\Bigg[\int_0^t \left|\left(Y_s^k - \hat{Y}_s^k\right) - \left(u_s^k - \hat{u}_s^k\right)\right|^2 ds\Bigg] + 2\sigma_2^2 \mathbb{E}\left[\int_0^t \left|\nabla L_k(Y_s^k) - \nabla L_k(\hat{Y}_s^k)\right|^2 ds\right] \\ &\le 2\mathbb{E}|z_0^k|^2 + \left(6\lambda_1^2 t + 4\sigma_1^2\right) \int_0^t \mathbb{E}|z_s^k|^2 ds + \left(6\lambda_1^2 t + 4\sigma_1^2\right) \int_0^t \left|u_s^k - \hat{u}_s^k\right|^2 ds \\ &\quad + 6\lambda_2^2 M_{\nabla L_1}^2 \int_0^t \mathbb{E}|z_s^k|^2 ds + 2\sigma_2^2 M_{\nabla L_1}^2 \int_0^t \mathbb{E}|z_s^k|^2 ds \\ &= 2\mathbb{E}|z_0^k|^2 + \left[6\lambda_1^2 t + 4\sigma_1^2 + M_{\nabla L_1}^2 \left(6\lambda_2^2 + 2\sigma_2^2\right)\right] \int_0^t \mathbb{E}|z_s^k|^2 ds + \left(6\lambda_1^2 t + 4\sigma_1^2\right) \int_0^t \left|u_s^k - \hat{u}_s^k\right|^2 ds. \end{aligned}$$

By Lemma 2, for $k = 1, 2$, we get

$$\begin{aligned} |u_s^k - \hat{u}_s^k|^2 = |m_{L_k}^\alpha[\nu_s] - m_{L_k}^\alpha[\hat{\nu}_s]|^2 &\le C^2 \left(\sqrt{w_1} W_2(\nu_s^1, \hat{\nu}_s^1) + \sqrt{w_2} W_2(\nu_s^2, \hat{\nu}_s^2)\right)^2 \\ &\le C^2 \left(\sqrt{w_1}\sqrt{\mathbb{E}|z_s^1|^2} + \sqrt{w_2}\sqrt{\mathbb{E}|z_s^2|^2}\right)^2 \\ &\le 2C^2 \left(\mathbb{E}|z_s^1|^2 + \mathbb{E}|z_s^2|^2\right). \end{aligned} \tag{33}$$

We further obtain

$$\begin{aligned} \mathbb{E}|z_t^1|^2 &\le 2\mathbb{E}|z_0^1|^2 + \widetilde{C}_1 \int_0^t \mathbb{E}|z_s^1|^2 ds + \widetilde{C}_2 \int \mathbb{E}|z_s^2|^2 ds \\ \mathbb{E}|z_t^2|^2 &\le 2\mathbb{E}|z_0^2|^2 + \widetilde{C}_1 \int_0^t \mathbb{E}|z_s^2|^2 ds + \widetilde{C}_2 \int \mathbb{E}|z_s^1|^2 ds, \end{aligned}$$

where $\widetilde{C}_1 = (1 + 2C^2)(6\lambda_1^2 t + 4\sigma_1^2) + (6\lambda_2^2 + 2\sigma_2^2) \max\{M_{\nabla L_1}^2, M_{\nabla L_2}^2\}$ and $\widetilde{C}_2 = 2C^2(6\lambda_1^2 t + 4\sigma_1^2)$. Combining the above two inequalities we deduce

$$\mathbb{E}|z_t^1|^2 + \mathbb{E}|z_t^2|^2 \le 2\big(\mathbb{E}|z_0^1|^2 + \mathbb{E}|z_0^2|^2\big) + (\widetilde{C}_1 + \widetilde{C}_2) \int_0^t \left(\mathbb{E}|z_s^1|^2 + \mathbb{E}|z_s^2|^2\right) ds.$$

Then, by Grönwall's inequality and the fact that $\mathbb{E}|z_0^1|^2 = \mathbb{E}|z_0^2|^2 = 0$, we infer that $\mathbb{E}|z_t^1|^2 + \mathbb{E}|z_t^2|^2 = 0$ for all $t \in [0,T]$. From inequality (33), we obtain $\|u^1 - \hat{u}^1\|_\infty = \|u^2 - \hat{u}^2\|_\infty = 0$, i.e., $(u^1, u^2) \equiv (\hat{u}^1, \hat{u}^2)$, proving in this way the uniqueness. ■

Remark 11 *Notice that the stochastic processes Y^1 and Y^2 in* (30a) *and* (30b) *are independent from each other for any input functions (u^1, u^2). In turn, since* (7a) *and* (7b) *are realized as (Y^1, Y^2) for a specific choice of (u^1, u^2) (i.e., for a fixed point of $\mathcal{T}$), we conclude that the processes $\overline{\theta}^1, \overline{\theta}^2$ from* (7a) *and* (7b) *are independent as stochastic processes. Notice, however, that both SDEs share parameters, e.g., the distribution ρ appearing in both the drift and diffusion terms of the equations.*

4. Large time behavior of mean-field equation and consensus formation

In this section we prove the global convergence of the mean-field system as stated in Theorem 3 by following the strategies that were first used in (Fornasier et al., 2024a) and further refined in (Riedl, 2023). In Section 4.1, we outline the main steps in the proof and state some important preliminary lemmas. In Section 4.2 we complete the proof of Theorem 3.

4.1 Proof Sketch

We consider the evolution of the energy functional $\mathcal{V}(\rho_t^k)$ defined by

$$\mathcal{V}(\rho_t^k) = \frac{1}{2}\int |\theta - \theta_k^*|^2 d\rho_t^k(\theta)$$

for $k = 1, 2$. Our primary objective is to demonstrate that the combined energy functional $\mathcal{V}(\rho_t^1) + \mathcal{V}(\rho_t^2)$ decreases over time according to the differential inequality

$$\frac{d}{dt}\left(\mathcal{V}(\rho_t^1) + \mathcal{V}(\rho_t^2)\right) \leq -\left(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2\right)\left(\mathcal{V}(\rho_t^1) + \mathcal{V}(\rho_t^2)\right) \tag{34}$$

until a time $T \leq T^*$ at which $\mathcal{V}(\rho_T^1) + \mathcal{V}(\rho_T^2) \leq \varepsilon$. In the case $T = T^*$, one can easily check that $\mathcal{V}(\rho_{T^*}^1) + \mathcal{V}(\rho_{T^*}^2) \leq \varepsilon$ by the definition of T^* in Theorem 3.

To accomplish this, we first derive a differential inequality for the evolution of $\mathcal{V}(\rho_t^1) + \mathcal{V}(\rho_t^2)$ by using the dynamics of ρ^1 and ρ^2. In particular, by considering the test functions $\phi_1(\theta) := \frac{1}{2}|\theta - \theta_1^*|^2$ and $\phi_2(\theta) := \frac{1}{2}|\theta - \theta_2^*|^2$ on the PDE (9), respectively, we derive in Lemma 3 an initial form for the sought differential inequality:

$$\begin{aligned}\frac{d}{dt}\big(\mathcal{V}(\rho_t^1) + \mathcal{V}(\rho_t^2)\big) \leq &-(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2)\big(\mathcal{V}(\rho_t^1) + \mathcal{V}(\rho_t^2)\big) \\ &+ \sqrt{2}(\lambda_1 + d\sigma_1^2)\big(|m_{L_1}^\alpha[\rho_t] - \theta_1^*| + |m_{L_2}^\alpha[\rho_t] - \theta_2^*|\big)\sqrt{\mathcal{V}(\rho_t^1) + \mathcal{V}(\rho_t^2)} \\ &+ \frac{d\sigma_1^2}{2}\big(|m_{L_1}^\alpha[\rho_t] - \theta_1^*|^2 + |m_{L_2}^\alpha[\rho_t] - \theta_2^*|^2\big).\end{aligned}$$

To find suitable bounds on the second and third terms in the above inequality, we need to control the quantity $|m_{L_1}^\alpha[\rho_t] - \theta_1^*| + |m_{L_2}^\alpha[\rho_t] - \theta_2^*|$. This is done using the quantitative Laplace principle in Lemma 4. To be more specific, under the inverse continuity property (12), we show that

$$|m_{L_1}^\alpha[\rho_t] - \theta_1^*| + |m_{L_2}^\alpha[\rho_t] - \theta_2^*| \lesssim l(r) + \exp(-\alpha r)\left(\frac{1}{\rho_t(B_r(\theta_1^*))} + \frac{1}{\rho_t(B_r(\theta_2^*))}\right),$$

with $r > 0$ small enough and l a strictly positive but monotonically decreasing function with $l(r) \to 0$ as $r \to 0$. As long as $\rho_t(B_r(\theta_1^*)), \rho_t(B_r(\theta_2^*)) > 0$, one can choose

$$\alpha > \frac{\log\left(\frac{1}{\rho_t(B_r(\theta_1^*))} + \frac{1}{\rho_t(B_r(\theta_2^*))}\right) - \log(l(r))}{r}$$

to guarantee $|m^\alpha_{L_1}[\rho_t] - \theta_1^*| + |m^\alpha_{L_2}[\rho_t] - \theta_2^*| \lesssim l(r)$, which can be made arbitrarily small by suitable choices of $r \ll 1$ and $\alpha \gg 1$.

To conclude the proof, we need to show that $\rho_t(B_r(\theta_1^*)), \rho_t(B_r(\theta_2^*)) > 0$ for all $r > 0$. We prove this in Lemma 5 by showing that the initial masses $\rho_0(B_r(\theta_1^*)), \rho_0(B_r(\theta_2^*)) > 0$ can decay at most exponentially fast for any $r > 0$, and therefore remain positive in any finite time interval $[0, T]$.

Lemma 3 (Evolution of energy functional $\mathcal{V}$) *For $k = 1, 2$, let objective functions $L_k : \mathbb{R}^d \to \mathbb{R}$ and fix $\alpha, \lambda_1, \lambda_2, \sigma_1, \sigma_2 > 0$. Moreover, let $T > 0$ and $\rho^1, \rho^2 \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}))$ form the weak solution to the Fokker-Planck equations* (8a) *and* (8b)*. Then the functional $\mathcal{V}(\rho_t^k)$ satisfies*

$$\begin{aligned}\frac{d}{dt}\mathcal{V}(\rho_t^k) \leq &-\big(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2\big)\mathcal{V}(\rho_t^k) \\ &+ \sqrt{2}(\lambda_1 + d\sigma_1^2)|m^\alpha_{L_k}[\rho_t] - \theta_k^*|\sqrt{\mathcal{V}(\rho_t^k)} + \frac{d\sigma_1^2}{2}|m^\alpha_{L_k}[\rho_t^k] - \theta_k^*|^2,\end{aligned}$$

with $M := \max\{M_{\nabla L_1}, M_{\nabla L_2}\}$.

Lemma 4 (Quantitative Laplace principle) *For $k = 1, 2$, denote $\underline{L_k} := \inf_{\theta \in \mathbb{R}^d} L_k(\theta)$. Let $\rho \in \mathcal{P}(\mathbb{R}^d)$ and fix $\alpha > 0$. For any $r > 0$, we define $L_r^k := \sup_{\theta \in B_r(\theta_k^*)} L_k(\theta)$. Then, under Assumption 2, for any $r \in (0, \min\{R_0^1, R_0^2\}]$ and $q_k > 0$ such that $q_k + L_r^k - \underline{L_k} \leq L_\infty^k$, we have*

$$|m^\alpha_{L_k}[\rho] - \theta_k^*| \leq \frac{(q_k + L_r^k - \underline{L_k})^{\nu_k}}{\eta_k} + \frac{\exp(-\alpha q_k)}{\rho(B_r(\theta_k^*))}\int |\theta - \theta_k^*| d\rho(\theta).$$

Lemma 5 *For $k = 1, 2$, let $T > 0, r > 0$, and fix parameters $\alpha, \lambda_1, \lambda_2, \sigma_1, \sigma_2 > 0$. Assume $\rho^1, \rho^2 \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ weakly solve the Fokker-Planck equations* (8a) *and* (8b) *respectively with initial conditions $\rho_0^1, \rho_0^2 \in \mathcal{P}(\mathbb{R}^d)$. Furthermore, denote $B_k := \sup_{t \in [0,T]} |m^\alpha_{L_k}[\rho_t] - \theta_k^*|$. Then for all $t \in [0, T]$ we have*

$$\rho_t^k(B_r(\theta_k^*)) \geq \Big(\int \phi_r^k(\theta) d\rho_0^k(\theta)\Big) \exp\big(-(q_k^l + q_k^g)t\big),$$

with the functions ϕ_r^1 and ϕ_r^2 as in (64) *and*

$$\begin{aligned}q_k^l &:= \max\left\{\frac{2\lambda_1(\sqrt{c}r + B_k)\sqrt{c}}{(1-c)^2 r} + \frac{2\sigma_1^2(cr^2 + B_k^2)(2c+d)}{(1-c)^4 r^2}, \frac{4\lambda_1^2}{(2c-1)\sigma_1^2}\right\}, \\ q_k^g &:= \max\left\{\frac{2\lambda_2 c M_{\nabla L_k}}{(1-c)^2} + \frac{\sigma_2^2 M_{\nabla L_k}^2 c(2c+d)}{(1-c)^4}, \frac{4\lambda_2^2}{(2c-1)\sigma_2^2}\right\},\end{aligned}$$

where $c \in (\frac{1}{2}, 1)$ can be any constant that satisfies the inequality

$$(2c-1)c \geq d(1-c)^2. \tag{35}$$

The proofs of Lemmas 3,4, and 5 are presented in Appendix C.

4.2 Proof of Theorem 3

We are now ready to present the detailed proof of Theorem 3.

Proof [Proof of Theorem 3] Let $M := \max\{M_{\nabla L_1}, M_{\nabla L_2}\}$. Using Lemma 3 we get

$$\begin{aligned}\frac{d}{dt}\big(\mathcal{V}(\rho_t^1)+\mathcal{V}(\rho_t^2)\big) &\leq -(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2)\big(\mathcal{V}(\rho_t^1)+\mathcal{V}(\rho_t^2)\big)\\ &\quad + \sqrt{2}(\lambda_1 + d\sigma_1^2)\Big(\sqrt{\mathcal{V}(\rho_t^1)}|m_{L_1}^\alpha[\rho_t]-\theta_1^*| + \sqrt{\mathcal{V}(\rho_t^2)}|m_{L_2}^\alpha[\rho_t]-\theta_2^*|\Big)\\ &\quad + \frac{d\sigma_1^2}{2}\big(|m_{L_1}^\alpha[\rho_t]-\theta_1^*|^2 + |m_{L_2}^\alpha[\rho_t]-\theta_2^*|^2\big)\\ &\leq -(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2)\big(\mathcal{V}(\rho_t^1)+\mathcal{V}(\rho_t^2)\big)\\ &\quad + \sqrt{2}(\lambda_1 + d\sigma_1^2)\big(|m_{L_1}^\alpha[\rho_t]-\theta_1^*| + |m_{L_2}^\alpha[\rho_t]-\theta_2^*|\big)\sqrt{\mathcal{V}(\rho_t^1)+\mathcal{V}(\rho_t^2)}\\ &\quad + \frac{d\sigma_1^2}{2}\big(|m_{L_1}^\alpha[\rho_t]-\theta_1^*|^2 + |m_{L_2}^\alpha[\rho_t]-\theta_2^*|^2\big).\end{aligned} \tag{36}$$

Let $T_\alpha \geq 0$ be given by

$$T_\alpha := \sup\big\{t \geq 0 : \mathcal{V}(\rho_{t'}^1)+\mathcal{V}(\rho_{t'}^2) > \varepsilon, |m_{L_1}^\alpha[\rho_{t'}]-\theta_1^*| + |m_{L_2}^\alpha[\rho_{t'}]-\theta_2^*| < \widetilde{C}(t')\ \forall t' \in [0,t]\big\}, \tag{37}$$

where

$$\widetilde{C}(t) := C\sqrt{\mathcal{V}(\rho_t^1)+\mathcal{V}(\rho_t^2)} \tag{38}$$

with

$$C := \min\left\{\frac{\vartheta}{2}\frac{(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2)}{\sqrt{2}(\lambda_1 + d\sigma_1^2)}, \sqrt{\vartheta\frac{(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2)}{d\sigma_1^2}}\right\}. \tag{39}$$

Then, combining (36) with (37), for all $t \in [0, T_\alpha]$ we have

$$\frac{d}{dt}\big(\mathcal{V}(\rho_t^1)+\mathcal{V}(\rho_t^2)\big) \leq -(1-\vartheta)(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2)\big(\mathcal{V}(\rho_t^1)+\mathcal{V}(\rho_t^2)\big) < 0, \tag{40}$$

where the last inequality comes from the assumption $2\lambda_1 > 2\lambda_2 M + d\sigma_1^2 + d\sigma_2^2 M^2$. This implies that the sum $\mathcal{V}(\rho_t^1)+\mathcal{V}(\rho_t^2)$ is decreasing in time in the interval $[0, T_\alpha]$. Moreover, Grönwall's inequality implies the upper bound

$$\mathcal{V}(\rho_t^1)+\mathcal{V}(\rho_t^2) \leq \big(\mathcal{V}(\rho_0^1)+\mathcal{V}(\rho_0^2)\big)\exp\big(-(1-\vartheta)(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2)t\big), \quad \text{for } t \in [0, T_\alpha]. \tag{41}$$

Accordingly, the decay in time of the sum $\mathcal{V}(\rho_t^1)+\mathcal{V}(\rho_t^2)$ implies that the auxiliary function $\widetilde{C}(t)$ decreases as well. Hence, for $k = 1, 2$,

$$\max_{t\in[0,T_\alpha]}|m_{L_k}^\alpha[\rho_t]-\theta_k^*| \leq \max_{t\in[0,T_\alpha]}|m_{L_1}^\alpha[\rho_t]-\theta_1^*| + |m_{L_2}^\alpha[\rho_t]-\theta_2^*| \leq \max_{t\in[0,T_\alpha]}\widetilde{C}(t) \leq C\sqrt{\mathcal{V}(\rho_0^1)+\mathcal{V}(\rho_0^2)}. \tag{42}$$

Also, note that

$$\begin{aligned}\int |\theta-\theta_1^*| d\rho_{T_\alpha}(\theta) &= w_1 \int |\theta-\theta_1^*| d\rho^1_{T_\alpha}(\theta) + w_2 \int |\theta-\theta_1^*| d\rho^2_{T_\alpha}(\theta)\\ &\leq w_1\sqrt{2\mathcal{V}(\rho^1_{T_\alpha})} + w_2 \int |\theta-\theta_2^*| + |\theta_1^*-\theta_2^*| d\rho^2_{T_\alpha}(\theta)\\ &\leq \sqrt{2\mathcal{V}(\rho^1_{T_\alpha})} + \sqrt{2\mathcal{V}(\rho^2_{T_\alpha})} + |\theta_1^*-\theta_2^*|\\ &\leq 2\sqrt{\mathcal{V}(\rho^1_{T_\alpha}) + \mathcal{V}(\rho^2_{T_\alpha})} + |\theta_1^*-\theta_2^*|\\ &\leq 2\sqrt{\mathcal{V}(\rho^1_0) + \mathcal{V}(\rho^2_0)} + |\theta_1^*-\theta_2^*|,\end{aligned} \tag{43}$$

and, similarly,

$$\int |\theta-\theta_2^*| d\rho_{T_\alpha}(\theta) \leq 2\sqrt{\mathcal{V}(\rho^1_0) + \mathcal{V}(\rho^2_0)} + |\theta_1^*-\theta_2^*|. \tag{44}$$

To conclude that $\mathcal{V}(\rho^1_{T_\alpha}) + \mathcal{V}(\rho^2_{T_\alpha}) \leq \varepsilon$, it remains to analyze the following different cases.
Case $T_\alpha \geq T^*$: If $T_\alpha \geq T^*$, we can use the definition of T^* in (14) and the bound for $\mathcal{V}(\rho^1_t)+\mathcal{V}(\rho^2_t)$ in (41) to conclude that $\mathcal{V}(\rho^1_{T^*}) + \mathcal{V}(\rho^2_{T^*}) \leq \varepsilon$ and that $\mathcal{V}(\rho^1_t) + \mathcal{V}(\rho^2_t)$ decayed exponentially fast up to that point.
Case $T_\alpha < T^*$ and $\mathcal{V}(\rho^1_{T_\alpha}) + \mathcal{V}(\rho^2_{T_\alpha}) \leq \varepsilon$: Nothing needs to be discussed in this case.
Case $T_\alpha < T^*$, $\mathcal{V}(\rho^1_{T_\alpha}) + \mathcal{V}(\rho^2_{T_\alpha}) > \varepsilon$ and $|m^\alpha_{L_1}[\rho_{T_\alpha}] - \theta_1^*| + |m^\alpha_{L_2}[\rho_{T_\alpha}] - \theta_2^*| < \widetilde{C}(T_\alpha)$: This case actually doesn't arise, since, if it did, it would contradict the definition of T_α.
Case $T_\alpha < T^*$, $\mathcal{V}(\rho^1_{T_\alpha}) + \mathcal{V}(\rho^2_{T_\alpha}) > \varepsilon$, and $|m^\alpha_{L_1}[\rho_{T_\alpha}] - \theta_1^*| + |m^\alpha_{L_2}[\rho_{T_\alpha}] - \theta_2^*| \geq \widetilde{C}(T_\alpha)$: We will show there exists $\alpha_0 > 0$ so that for any $\alpha > \alpha_0$ we have

$$|m^\alpha_{L_1}[\rho_{T_\alpha}] - \theta_1^*| + |m^\alpha_{L_2}[\rho_{T_\alpha}] - \theta_2^*| < \widetilde{C}(T_\alpha), \tag{45}$$

which would contradict $|m^\alpha_{L_1}[\rho_{T_\alpha}] - \theta_1^*| + |m^\alpha_{L_2}[\rho_{T_\alpha}] - \theta_2^*| \geq \widetilde{C}(T_\alpha)$. In other words, we prove that the last case never happens if we choose α sufficiently large. To show (45), we define

$$q_1 := \frac{1}{2}\min\left\{ \left(\frac{\eta_1}{4} C\sqrt{\varepsilon}\right)^{\frac{1}{\nu_1}}, L^1_\infty \right\} \qquad \text{and} \qquad r_1 := \max_{s\in[0,R^1_0]} \left\{ \max_{\theta\in B_s(\theta_1^*)} L_1(\theta) - \underline{L_1} \leq q_1 \right\},$$

where $\underline{L_1} := \inf_{\theta\in\mathbb{R}^d} L_1(\theta)$, and $\eta_1, \nu_1, L^1_\infty$ come from assumption (II) and C is defined in (39). By construction, these choices satisfy $r_1 \leq R^1_0$ and $q_1 + \sup_{\theta\in B_{r_1}(\theta_1^*)} L_1(\theta) - \underline{L_1} \leq 2q_1 \leq L^1_\infty$. Furthermore, we note $q_1 > 0$, and by the continuity of L_1, there exists $s_{q_1} > 0$ such that $L_1(\theta) - \underline{L_1} \leq q_1$ for all $\theta \in B_{s_{q_1}}(\theta_1^*)$, thus yielding $r_1 > 0$. Therefore, we can apply Lemma 4 with q_1 and r_1 as above to get

$$\begin{aligned}|m^\alpha_{L_1}[\rho_{T_\alpha}] - \theta_1^*| &\leq \frac{\left(q_1 + \sup_{\theta\in B_{r_1}(\theta_1^*)} L_1(\theta) - \underline{L_1}\right)^{\nu_1}}{\eta_1} + \frac{\exp(-\alpha q_1)}{\rho_{T_\alpha}(B_{r_1}(\theta_1^*))} \int |\theta-\theta_1^*| d\rho_{T_\alpha}(\theta)\\ &\leq \frac{(2q_1)^{\nu_1}}{\eta_1} + \frac{\exp(-\alpha q_1)}{\rho_{T_\alpha}(B_{r_1}(\theta_1^*))} \int |\theta-\theta_1^*| d\rho_{T_\alpha}(\theta)\\ &\leq \frac{\left[\left(\frac{\eta_1}{4} C\sqrt{\varepsilon}\right)^{\frac{1}{\nu_1}}\right]^{\nu_1}}{\eta_1} + \frac{\exp(-\alpha q_1)}{\rho_{T_\alpha}(B_{r_1}(\theta_1^*))} \int |\theta-\theta_1^*| d\rho_{T_\alpha}(\theta)\\ &= \frac{C}{4}\sqrt{\varepsilon} + \frac{\exp(-\alpha q_1)}{\rho_{T_\alpha}(B_{r_1}(\theta_1^*))} \int |\theta-\theta_1^*| d\rho_{T_\alpha}(\theta).\end{aligned} \tag{46}$$

Similarly, by choosing

$$q_2 := \frac{1}{2}\min\left\{\left(\frac{\eta_2}{4}C\sqrt{\varepsilon}\right)^{\frac{1}{\nu_2}}, L_\infty^2\right\} \qquad \text{and} \qquad r_2 := \max_{s\in[0,R_0^2]}\left\{\max_{\theta\in B_s(\theta_1^*)} L_2(\theta) - \underline{L_2} \leq q_2\right\},$$

we have

$$|m_{L_2}^\alpha[\rho_{T_\alpha}] - \theta_2^*| \leq \frac{C}{4}\sqrt{\varepsilon} + \frac{\exp(-\alpha q_2)}{\rho_{T_\alpha}(B_{r_2}(\theta_2^*))}\int |\theta - \theta_2^*| d\rho_{T_\alpha}(\theta).$$

Combining with inequalities (43) and (44), we further obtain

$$\begin{aligned}
|m_{L_1}^\alpha[\rho_{T_\alpha}] - \theta_1^*| + |m_{L_2}^\alpha[\rho_{T_\alpha}] - \theta_2^*| &\leq \frac{C}{2}\sqrt{\varepsilon} + \frac{\exp(-\alpha q_1)}{\rho_{T_\alpha}(B_{r_1}(\theta_1^*))}\int |\theta - \theta_1^*| d\rho_{T_\alpha}(\theta) \\
&\quad + \frac{\exp(-\alpha q_2)}{\rho_{T_\alpha}(B_{r_2}(\theta_2^*))}\int |\theta - \theta_2^*| d\rho_{T_\alpha}(\theta) \\
&\leq \frac{C}{2}\sqrt{\varepsilon} + \left(\frac{\exp(-\alpha q_1)}{\rho_{T_\alpha}(B_{r_1}(\theta_1^*))} + \frac{\exp(-\alpha q_2)}{\rho_{T_\alpha}(B_{r_2}(\theta_2^*))}\right)\left(2\sqrt{\mathcal{V}(\rho_0^1) + \mathcal{V}(\rho_0^2)} + |\theta_1^* - \theta_2^*|\right) \\
\leq \frac{C}{2}\sqrt{\varepsilon} + \exp(-\alpha q)&\left(\frac{1}{\rho_{T_\alpha}(B_{r_1}(\theta_1^*))} + \frac{1}{\rho_{T_\alpha}(B_{r_2}(\theta_2^*))}\right)\left(2\sqrt{\mathcal{V}(\rho_0^1) + \mathcal{V}(\rho_0^2)} + |\theta_1^* - \theta_2^*|\right),
\end{aligned} \tag{47}$$

with $q := \min\{q_1, q_2\}$. By (42) we have the bound $G_{\alpha,k} := \max_{t\in[0,T_\alpha]} |m_{L_k}^\alpha[\rho_t] - \theta_k^*| \leq C\sqrt{\mathcal{V}(\rho_0^1) + \mathcal{V}(\rho_0^2)} := G$, which implies that all assumptions of Lemma 5 are satisfied. Therefore, by Lemma 5, for $k = 1, 2$ and mollifiers $\phi_{r_k}^k$ defined in (64), there exist $a_k := a_k^l + a_k^g > 0$ such that

$$\rho_{T_\alpha}^k\left(B_{r_k}(\theta_k^*)\right) \geq \int \phi_{r_k}^k(\theta) d\rho_0^k(\theta)\exp(-a_k T_\alpha),$$

where

$$a_k^l := \max\left\{h_1^l + h_2^l\frac{G_{\alpha,k}}{r_k} + h_3^l\frac{G_{\alpha,k}^2}{r_k^2}, h_4^l\right\} \qquad \text{and} \qquad a_k^g := \max\left\{h_1^g M_{\nabla L_k} + h_2^g M_{\nabla L_k}^2, h_3^g\right\},$$

with $h_1^l, h_2^l, h_3^l, h_4^l$ and h_1^g, h_2^g, h_3^g only depending on $\lambda_1, \lambda_2, \sigma_1, \sigma_2$ and d. Now we let $\widetilde{a} := \widetilde{a}^l + \widetilde{a}^g$, where

$$\widetilde{a}^l := \max\left\{h_1^l + h_2^l\frac{G}{r} + h_3^l\frac{G^2}{r^2}, h_4^l\right\} \qquad \text{and} \qquad \widetilde{a}^g := \max\left\{h_1^g M + h_2^g M^2, h_3^g\right\},$$

with $r := \min\{r_1, r_2\}$ and $M := \max\{M_{\nabla L_1}, M_{\nabla L_2}\}$. Then

$$\begin{aligned}
\rho_{T_\alpha}(B_{r_1}(\theta_1^*)) &= w_1\rho_{T_\alpha}^1(B_{r_1}(\theta_1^*)) + w_2\rho_{T_\alpha}^2(B_{r_1}(\theta_1^*)) \\
&\geq w_1\rho_{T_\alpha}^1(B_{r_1}(\theta_1^*)) \\
&\geq w_1\int \phi_{r_1}^1(\theta) d\rho_0^1(\theta)\exp(-a_1 T_\alpha) \\
&\geq w_1\int \phi_{r_1}^1(\theta) d\rho_0^1(\theta)\exp(-\widetilde{a}T^*) > 0,
\end{aligned}$$

since $\widetilde{a} \geq a_1$ and $T^* \geq T_\alpha$, and, similarly,

$$\rho_{T_\alpha}(B_{r_2}(\theta_2^*)) \geq w_2\int \phi_{r_2}^2(\theta) d\rho_0^2(\theta)\exp(-\widetilde{a}T^*) > 0.$$

Denote $K := \min\left\{w_1 \int \phi^1_{r_1}(\theta) d\rho^1_0(\theta), w_2 \int \phi^2_{r_2}(\theta) d\rho^2_0(\theta)\right\}$. Then by using $\alpha > \alpha_0$ with

$$\alpha_0 := \frac{\widetilde{a}T^* + \log\left(\frac{2\sqrt{\mathcal{V}(\rho^1_0)+\mathcal{V}(\rho^2_0)}+|\theta^*_1-\theta^*_2|}{CK\sqrt{\varepsilon}}\right)}{q}, \tag{48}$$

the second term in (47) is strictly smaller than $\frac{C}{2}\sqrt{\varepsilon}$. That is,

$$\begin{aligned}
&\exp(-\alpha q)\left(\frac{1}{\rho_{T_\alpha}(B_{r_1}(\theta^*_1))} + \frac{1}{\rho_{T_\alpha}(B_{r_2}(\theta^*_2))}\right)\left(2\sqrt{\mathcal{V}(\rho^1_0)+\mathcal{V}(\rho^2_0)} + |\theta^*_1-\theta^*_2|\right)\\
&< \exp(-\alpha_0 q)\frac{2}{K\exp(-\widetilde{a}T^*)}\left(2\sqrt{\mathcal{V}(\rho^1_0)+\mathcal{V}(\rho^2_0)} + |\theta^*_1-\theta^*_2|\right)\\
&= \frac{C}{2}\sqrt{\varepsilon}.
\end{aligned}$$

It follows from (47) that

$$|m^\alpha_{L_1}[\rho_{T_\alpha}] - \theta^*_1| + |m^\alpha_{L_2}[\rho_{T_\alpha}] - \theta^*_2| < C\sqrt{\varepsilon} < C\sqrt{\mathcal{V}(\rho^1_{T_\alpha}) + \mathcal{V}(\rho^2_{T_\alpha})},$$

contradicting in this way (45). ■

Remark 12 *Theorem 3 can be naturally extended to the case in which the number of underlying clusters K is strictly greater than 2. Here we highlight how the important constant α_0 in (48) would change in this case. Let us denote by ρ^k_t the distribution of the k-th cluster for $k \in [K]$, and let $\rho_t := \sum_{k=1}^K w_k \rho^k_t$ with $\sum_{k=1}^K w_k = 1$. Furthermore, we denote by θ^*_k the global minimizer of the objective function corresponding to cluster k. Similarly to the computations in (36), from Lemma 3 we obtain*

$$\begin{aligned}
\frac{d}{dt}\sum_{k=1}^K \mathcal{V}(\rho^k_t) \leq &-(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2)\sum_{k=1}^K \mathcal{V}(\rho^k_t)\\
&+ \sqrt{2}(\lambda_1 + d\sigma_1^2)\left(\sum_{k=1}^K |m^\alpha_{L_k}[\rho_t] - \theta^*_k|\right)\sqrt{\sum_{k=1}^K \mathcal{V}(\rho^k_t)} + \frac{d\sigma_1^2}{2}\left(\sum_{k=1}^K |m^\alpha_{L_k}[\rho_t] - \theta^*_k|\right).
\end{aligned}$$

Following the arguments in the proof of Theorem 3, it would remain to bound the term $\sum_{k=1}^K |m^\alpha_{L_k}[\rho_t] - \theta^*_k|$ as in (47). In particular, one can easily obtain

$$\sum_{k=1}^K |m^\alpha_{L_k}[\rho_t] - \theta^*_k| \leq \frac{C}{2}\sqrt{\varepsilon} + \sum_{k=1}^K \frac{\exp(-\alpha q_k)}{\rho_{T_\alpha}(B_{r_k}(\theta^*_k))}\int |\theta - \theta^*_k| d\rho_{T_\alpha}(\theta)$$

and

$$\begin{aligned}\int |\theta-\theta_k^*|d\rho_{T_\alpha}(\theta) &= \sum_{l=1}^K w_l \int |\theta-\theta_k^*|d\rho^l_{T_\alpha}(\theta)\\ &\leq \sum_{l=1}^K w_l \left(|\theta-\theta_l^*| + |\theta_l^*-\theta_k^*|\right) d\rho^l_{T_\alpha}(\theta)\\ &\leq \sum_{l=1}^K \sqrt{2\mathcal{V}(\rho^l_{T_\alpha})} + \sum_{l=1}^K |\theta_l^*-\theta_k^*|\\ &\leq 2\sqrt{\sum_{l=1}^K \mathcal{V}(\rho^l_{T_\alpha})} + \sum_{l=1}^{K-1}\sum_{h=l+1}^K |\theta_l^*-\theta_h^*|.\end{aligned}$$

Therefore, we obtain

$$\sum_{k=1}^K |m^\alpha_{L_k}[\rho_t]-\theta_k^*| \leq \frac{C}{2}\sqrt{\varepsilon} + \sum_{k=1}^K \frac{\exp(-\alpha q)}{\rho_{T_\alpha}(B_{r_k}(\theta_k^*))}\left(2\sqrt{\sum_{l=1}^K \mathcal{V}(\rho^l_{T_\alpha})} + \sum_{l=1}^{K-1}\sum_{h=l+1}^K |\theta_l^*-\theta_h^*|\right),$$

with $q := \min_{k\in[K]} q_k$*. Then again following the arguments in the proof of Theorem 3, we can estimate the important constant* α_0 *similarly to* (48) *and obtain:*

$$\alpha_0 := \frac{\widetilde{a}T^* + \log\left(\frac{2\sqrt{\sum_{l=1}^K \mathcal{V}(\rho_0^l)} + \sum_{l=1}^{K-1}\sum_{h=l+1}^K |\theta_l^*-\theta_h^*|}{CK\sqrt{\varepsilon}}\right)}{q}.$$

5. Large Time Behavior of Finite Particle Systems

In this section, we present the proof of Theorem 4 on the convergence of the finite particle system toward global minimizers of the objective functions. The proof is based on the combination of the mean-field convergence result established in Theorem 3 and a quantitative mean-field approximation result stated in Proposition 1 below. To get this quantitative approximation result, we work on a set with large probability in which the dynamics of the finite particle system stay withing a compact set. Let $\{\bar\theta_t^{1,i_1}\}_{i_1=1}^{N_1}, \{\bar\theta_t^{2,i_2}\}_{i_2=1}^{N_2}$ be independent copies of the solution to the mean-field dynamics (7a) and (7b), respectively. In what follows we use $(\Omega,\mathcal{F},\mathbb{P})$ to denote a common probability space over which all considered stochastic processes get their realizations. In this probability space, we consider the subset Ω_M of Ω defined according to:

$$\Omega_M := \left\{\omega\in\Omega : \sup_{t\in[0,\tilde T]} \frac{1}{N}\sum_{k=1,2}\sum_{i_k=1}^{N_k} \max\left\{\left|\theta_t^{k,i_k}(\omega)\right|^4, \left|\bar\theta_t^{k,i_k}(\omega)\right|^4\right\} \leq M\right\}.$$

In the following, $M>0$ is a constant that will be properly chosen in the proof of Theorem 4, and θ^{k,i_k} continues to denote the interacting particles of system (4). Finally, $\tilde T$ is a time horizon that will be chosen later on.

Before presenting the non-asymptotic mean-field approximation result, Proposition 1, we first prove that the stochastic processes of interest stay bounded with high probability (i.e., we estimate the probability of the set Ω_M). This is the content of Lemma 6. The proofs of these two statements are deferred to Appendix D.

Lemma 6 *Let $\tilde{T} > 0$, $\rho_0 := w_1\rho_0^1 + w_2\rho_0^2 \in \mathcal{P}_4(\mathbb{R}^d)$ and let $N = N_1 + N_2 \in \mathbb{N}$ be fixed. Moreover, let $\{\theta_t^{1,i_1}\}_{i_1=1}^{N_1}, \{\theta_t^{2,i_2}\}_{i_2=1}^{N_2}$ be the solution of the finite interacting particle system (4), and let $\{\bar{\theta}_t^{1,i_1}\}_{i_1=1}^{N_1}, \{\bar{\theta}_t^{2,i_2}\}_{i_2=1}^{N_2}$ denote independent copies of the solutions to the mean-field dynamics (4a) and (4b), respectively. Then, under Assumption 1, for any $M > 0$ we have*

$$\mathbb{P}(\Omega_M) = \mathbb{P}\left(\sup_{t\in[0,\tilde{T}]} \frac{1}{N}\sum_{k=1,2}\sum_{i_k=1}^{N_k} \max\left\{\left|\theta_t^{k,i_k}(\omega)\right|^4, \left|\bar{\theta}_t^{k,i_k}(\omega)\right|^4\right\} \leq M\right) \geq 1 - \frac{2C_{Bound}}{M}, \tag{49}$$

where $C_{Bound} = C_{Bound}(\lambda_1, \lambda_2, \sigma_1, \sigma_2, \tilde{T}, b_{11}, b_{12}, b_{21}, b_{22})$ is a constant that is independent of N and d. Here, b_{11}, b_{12}, b_{21} and b_{22} are problem-dependent constants defined in Lemma 10.

Lemma 6 shows that all the considered processes are bounded uniformly in time with high probability. By restricting to the set Ω_M, we can obtain the following quantitative mean-field approximation.

Proposition 1 (Quantitative Mean-field Approximation) *Under the same assumptions as in Lemma 6, for $k = 1, 2$, if $(\theta_t^{k,i_k})_{t\geq 0}$ and $(\bar{\theta}_t^{k,i_k})_{t\geq 0}$ share the same initial data as well as the Brownian motion paths $(B_t^{k,i_k})_{t\geq 0}, (\tilde{B}_t^{k,i_k})_{t\geq 0}$ for all $i_k \in [N_k]$, then we have a probabilistic mean-field approximation of the form*

$$\max_{\substack{k=1,2,\\ i_k\in[N_k]}} \sup_{t\in[0,\tilde{T}]} \mathbb{E}\left[\left|\theta_t^{k,i_k} - \bar{\theta}_t^{k,i_k}\right|^2 \,\middle|\, \Omega_M\right] \leq C_{MFA}(N_1^{-1} + N_2^{-1}), \tag{50}$$

where $C_{MFA} := C_{MFA}\left(\alpha, C_{L_1}, C_{L_2}, C_{\nabla L_1}, C_{\nabla L_2}, M, \mathcal{M}_2, b_{11}, b_{12}, b_{21}, b_{22}\right)$, and $\mathcal{M}_2$ is an upper bound on the second moment of ρ_t^N uniformly over time $t \in [0, \tilde{T}]$.

Proposition 1 states that, for a fixed time horizon $\tilde{T}$, it is possible to take N large enough so that, on average, the trajectories of the *interacting* particles are close to the *independent* particles with the law specified by the mean field dynamics. It is is a quantitative propagation of chaos estimate.

Equipped with Lemma 6 and Proposition 1, we are now ready to prove Theorem 4.

Proof [Proof of Theorem 4] Let T be the first time in $[0, T^*]$ for which

$$\mathcal{V}(\rho_T^1) + \mathcal{V}(\rho_T^2) \leq \varepsilon,$$

where ρ^1, ρ^2 form the solution to the mean field Fokker-Planck equation. Notice that this T exists by the analysis in Theorem 3. Now, let us denote by $K_{\varepsilon_{\text{total}}}^N$ the subset of Ω where (16) does not hold, where T is chosen as above. Finally, let $\tilde{T} = T^*$ in the definition of the set Ω_M. Then one can estimate the measure of the set $K_{\varepsilon_{\text{total}}}^N$ as:

$$\begin{aligned}
\mathbb{P}\left(K_{\varepsilon_{\text{total}}}^N\right) &= \mathbb{P}\left(K_{\varepsilon_{\text{total}}}^N \cap \Omega_M\right) + \mathbb{P}\left(K_{\varepsilon_{\text{total}}}^N \cap \Omega_M^c\right) \\
&\leq \frac{\mathbb{P}\left(\Omega_M\right)}{\varepsilon_{\text{total}}} \mathbb{E}\left[\left|\frac{1}{N_1}\sum_{i_1=1}^{N_1} \theta_T^{1,i_1} - \theta_1^*\right|^2 + \left|\frac{1}{N_2}\sum_{i_2=1}^{N_2} \theta_T^{2,i_2} - \theta_2^*\right|^2 \,\middle|\, \Omega_M\right] + \mathbb{P}\left(\Omega_M^c\right),
\end{aligned}$$

where the last inequality comes from the conditional Markov's inequality.

By the triangle inequality, we have the error decomposition

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{N_1}\sum_{i_1=1}^{N_1}\theta_T^{1,i_1}-\theta_1^*\right|^2+\left|\frac{1}{N_2}\sum_{i_2=1}^{N_2}\theta_T^{2,i_2}-\theta_2^*\right|^2\,\middle|\,\Omega_M\right]\\
&\quad\le 2\mathbb{E}\left[\left|\frac{1}{N_1}\sum_{i_1=1}^{N_1}\left(\theta_T^{1,i_1}-\bar\theta_T^{1,i_1}\right)\right|^2+\left|\frac{1}{N_2}\sum_{i_2=1}^{N_2}\left(\theta_T^{2,i_2}-\bar\theta_T^{2,i_2}\right)\right|^2\,\middle|\,\Omega_M\right]\\
&\qquad+2\mathbb{E}\left[\left|\frac{1}{N_1}\sum_{i_1=1}^{N_1}\bar\theta_T^{1,i_1}-\theta_1^*\right|^2+\left|\frac{1}{N_2}\sum_{i_2=1}^{N_2}\bar\theta_T^{2,i_2}-\theta_2^*\right|^2\,\middle|\,\Omega_M\right],
\end{aligned}\tag{51}$$

which divides the overall error into the mean-field approximation error and the optimization error in the mean-field limit. The first term can be bounded using the quantitative mean-field approximation in Proposition 1, i.e.,

$$\mathbb{E}\left[\left|\frac{1}{N_1}\sum_{i_1=1}^{N_1}\left(\theta_T^{1,i_1}-\bar\theta_T^{1,i_1}\right)\right|^2+\left|\frac{1}{N_2}\sum_{i_2=1}^{N_2}\left(\theta_T^{2,i_2}-\bar\theta_T^{2,i_2}\right)\right|^2\,\middle|\,\Omega_M\right]\le C_{\mathrm{MFA}}(N_1^{-1}+N_2^{-1})\,.$$

For the second term, since the law of each $\bar\theta^{1,i_1}$ is ρ^1 and the law of each $\bar\theta^{2,i_2}$ is ρ^2 we conclude

$$\mathbb{E}\left[\left|\frac{1}{N_1}\sum_{i_1=1}^{N_1}\bar\theta_T^{1,i_1}-\theta_1^*\right|^2+\left|\frac{1}{N_2}\sum_{i_2=1}^{N_2}\bar\theta_T^{2,i_2}-\theta_2^*\right|^2\right]\le 2\left(\mathcal{V}(\rho_T^1)+\mathcal{V}(\rho_T^2)\right)\le 2\varepsilon,$$

from where it follows that

$$\mathbb{E}\left[\left|\frac{1}{N_1}\sum_{i_1=1}^{N_1}\bar\theta_T^{1,i_1}-\theta_1^*\right|^2+\left|\frac{1}{N_2}\sum_{i_2=1}^{N_2}\bar\theta_T^{2,i_2}-\theta_2^*\right|^2\,\middle|\,\Omega_M\right]\le\frac{2\varepsilon}{\mathbb{P}(\Omega_M)}.$$

Combining the above estimates with (51) gives the error bound

$$\mathbb{E}\left[\left|\frac{1}{N_1}\sum_{i_1=1}^{N_1}\theta_T^{1,i_1}-\theta_1^*\right|^2+\left|\frac{1}{N_2}\sum_{i_2=1}^{N_2}\theta_T^{2,i_2}-\theta_2^*\right|^2\,\middle|\,\Omega_M\right]\le 2C_{\mathrm{MFA}}(N_1^{-1}+N_2^{-1})+\frac{4\varepsilon}{\mathbb{P}(\Omega_M)}\,.\tag{52}$$

and thus

$$\mathbb{P}\left(K^N_{\varepsilon_{\mathrm{total}}}\right)\le\frac{1}{\varepsilon_{\mathrm{total}}}(2C_{\mathrm{MFA}}N^{-1}+4\varepsilon)+\mathbb{P}(\Omega_M^c)\le\frac{1}{\varepsilon_{\mathrm{total}}}(2C_{\mathrm{MFA}}(N_1^{-1}+N_2^{-1})+4\varepsilon)+\frac{C_{\mathrm{Bound}}}{M}.$$

Notice that, as discussed in Remark 3, we can make the right hand in the above display be smaller than some fix δ by first choosing ε to be small enough, for example so that $\frac{4\varepsilon}{\varepsilon_{\mathrm{total}}}\le\delta/3$; then taking M large enough so that $C_{\mathrm{bound}}/M\le\xi\le\delta/3$; and finally, setting N_1 and N_2 large enough so that $\frac{2C_{\mathrm{MFA}}}{\varepsilon_{\mathrm{total}}}(\frac{1}{N_1}+\frac{1}{N_2})$ is less than $\delta/3$.

■

6. Conclusions

This paper is a first step in bridging the consensus-based optimization literature and other PDE-based optimization methods with the federated learning problem. In particular, we have proposed a new CBO-type system of interacting particles that can be used to solve non-convex optimization problems arising in practical clustered federated learning settings. We prove that our particle system converges to a suitable mean-field limit when the number of interacting particles goes to infinity. In turn, we analyze the time evolution of the mean-field model and discuss how it forces particles within each cluster to reach consensus around a global minimizer of the cluster's objective function. This mean-field point of view may actually not be too far from reality, specially when dealing with cross-devices federated learning problems, where the number of users is indeed quite large. Motivated by our new CBO-type particle dynamics, we propose the FedCBO algorithm and empirically asses its performance. In our experiments, we show that our algorithm outperforms current state-of-the-art methods for federated learning.

Some important questions motivated by our work that deserve further investigation are the following. On the theoretical side, the long-term stability behavior of the mean-field system is still an open problem. In particular, it is unclear how the model behaves after the variance (defined in Theorem 3) reaches the prescribed tolerance level ε. In addition, it is of interest to analyze the more realistic setting where agents within the same cluster may have different, although related, loss functions. On the experimental side, we would like to investigate the robustness to adversarial attacks of the FedCBO algorithm. Indeed, given the weighted averaging mechanism in the model aggregation step (21) it is not unreasonable to expect that the FedCBO algorithm can offer some protection against adversarial attacks. Finally, exploring further strategies to reduce the communication cost of our algorithm is another research topic of interest.

Acknowledgments

Authors' names are listed in alphabetical order by family name. The authors are thankful to Hui Huang and Jinniao Qiu for enlightening discussions on mean-field limits of CBO. The authors would also like to thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. This work started while the authors were visiting the Simons Institute to participate in the program "Geometric Methods in Optimization and Sampling" during the Fall of 2021. The authors would like to thank the institute for hospitality and support. NGT was supported by NSF-DMS grants 2005797 and 2236447, and, together with SL would like to thank the IFDS at UW-Madison and NSF through TRIPODS grant 2023239 for their support. JAC was supported by the Advanced Grant Nonlocal-CPD (Nonlocal PDEs for Complex Particle Dynamics: Phase Transitions, Patterns and Synchronization) of the European Research Council Executive Agency (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 883363). JAC was partially supported by the EPSRC grant numbers EP/T022132/1 and EP/V051121/1. JAC was also partially supported by the "Maria de Maeztu" Excellence Unit IMAG, reference CEX2020-001105-M, funded by MCIN/AEI/10.13039/501100011033/. YZ was supported by NSF grant DMS-2411396.

Appendix A. Moment Estimates for the Stochastic Empirical Measures

For the solution $\boldsymbol{\theta}^{1,N} \in \mathcal{C}([0,T],\mathbb{R}^d)^{N_1}, \boldsymbol{\theta}^{2,N} \in \mathcal{C}([0,T],\mathbb{R}^d)^{N_2}$ of the particle system (4), we denote by

$$\rho_t^{1,N} = \frac{1}{N_1}\sum_{i_1=1}^{N_1} \delta_{\theta_t^{1,(i_1,N)}} \qquad \rho_t^{2,N} = \frac{1}{N_2}\sum_{i_2=1}^{N_2} \delta_{\theta_t^{2,(i_2,N)}} \qquad \rho_t^{N} = \frac{N_1}{N}\rho_t^{1,N} + \frac{N_2}{N}\rho_t^{2,N}$$

the empirical measures corresponding to $\boldsymbol{\theta}^{1,N}, \boldsymbol{\theta}^{2,N}$ for each $t \in [0,T]$.

Lemma 7 (Moment Estimates) *Let L_1, L_2 satisfy Assumption 1 and either (i) boundedness, or (ii) quadratic growth at infinity, and $\rho_0 \in \mathcal{P}_{2p}(\mathbb{R}^d), p \geq 1$. Further, let $\boldsymbol{\theta}^{1,N}, \boldsymbol{\theta}^{2,N}$ be the solution of the particle system (4) with $\rho_0^{\otimes N}$-distributed initial data $\boldsymbol{\theta}_0^{1,N}, \boldsymbol{\theta}_0^{2,N}$ and $\rho^{1,N_1}, \rho^{2,N_2}$ and ρ^N the corresponding empirical measures. Then, there exists a constant $K > 0$, independent of N, such that*

$$\sup_{t\in[0,T]} \mathbb{E}\int |\theta|^{2p} d\rho_t^N, \quad \sup_{t\in[0,T]} \mathbb{E}|m_{L_1}^\alpha[\rho_t^N]|^{2p}, \quad \sup_{t\in[0,T]} \mathbb{E}|m_{L_2}^\alpha[\rho_t^N]|^{2p} \leq K, \tag{53}$$

and consequently also the estimates

$$\sup_{t\in[0,T]} \mathbb{E}\int |\theta|^2 d\eta_{1,t}^{\alpha,N}, \qquad \sup_{t\in[0,T]} \mathbb{E}|\theta_t^{1,i_1}|^{2p} \leq K,$$

$$\sup_{t\in[0,T]} \mathbb{E}\int |\theta|^2 d\eta_{2,t}^{\alpha,N}, \qquad \sup_{t\in[0,T]} \mathbb{E}|\theta_t^{2,i_2}|^{2p} \leq K,$$

for $i_1 = 1, 2, \dots, N_1$ and $i_2 = 1, 2, \dots, N_2$.

Proof Let $\boldsymbol{\theta}^{1,N}, \boldsymbol{\theta}^{2,N}$ be the solution of the particle system (4). Using the inequality $(a+b)^q \leq 2^{q-1}(a^q + b^q), q \geq 1$ and the Itô isometry and Jensen's inequality yields

$$\begin{aligned}
\mathbb{E}|\theta_t^{1,i_1}|^{2p} &= \mathbb{E}\bigg|\theta_0^{1,i_1} + \int_0^t \big(-\lambda_1(\theta_s^{1,i_1} - m_{L_1}^\alpha[\rho_s^N]) - \lambda_2 \nabla L_1(\theta_s^{1,i_1})\big)ds \\
&\quad + \int_0^t \sigma_1|\theta_s^{1,i_1} - m_{L_1}^\alpha[\rho_s^N]|dB_s^{1,i_1} + \int_0^t \sigma_2|\nabla L_1(\theta_s^{1,i_1})|d\widetilde{B}_s^{1,i_1}\bigg|^{2p} \\
&\leq 2^{2p-1}\mathbb{E}|\theta_0^{1,i_1}|^{2p} + 2^{3(2p-1)}\mathbb{E}\left|\int_0^t \big(-\lambda_1(\theta_s^{1,i_1} - m_{L_1}^\alpha[\rho_s^N]) - \lambda_2\nabla L_1(\theta_s^{1,i_1})\big)ds\right|^{2p} \\
&\quad + 2^{3(2p-1)}\mathbb{E}\left|\int_0^t \sigma_1|\theta_s^{1,i_1} - m_{L_1}^\alpha[\rho_s^N]|dB_s^{1,i_1}\right|^{2p} + 2^{3(2p-1)}\mathbb{E}\left|\int_0^t \sigma_2|\nabla L_1(\theta_s^{1,i_1})|d\widetilde{B}_s^{1,i_1}\right|^{2p} \\
&\leq 2^{2p-1}\mathbb{E}|\theta_0^{1,i_1}|^{2p} + 2^{4(2p-1)}\lambda_1^{2p}T^{2p-1}\int_0^t \big(\mathbb{E}|\theta_s^{1,i_1}|^{2p}ds + \mathbb{E}|m_{L_1}^\alpha[\rho_s^N]|^{2p}\big)ds \\
&\quad + 2^{3(2p-1)}\lambda_2^{2p}T^{2p-1}\int_0^t \mathbb{E}|\nabla L_1(\theta_s^{1,i_1})|^{2p}ds \qquad \text{(by Hölder's inequality)} \\
&\quad + 2^{4(2p-1)}T^{p-1}\sigma_1^{2p}p(2p-1)^p\int_0^t \big(\mathbb{E}|\theta_s^{1,i_1}|^{2p}ds + \mathbb{E}|m_{L_1}^\alpha[\rho_s^N]|^{2p}\big)ds \\
&\quad + 2^{3(2p-1)}T^{p-1}\sigma_2^{2p}p(2p-1)^p\int \mathbb{E}|\nabla L_1(\theta_s^{1,i_1})|^{2p}ds \\
&= 2^{2p-1}\mathbb{E}|\theta_0^{1,i_1}|^{2p} + 2^{4(2p-1)}\big(\lambda_1^{2p}T^p + p(2p-1)^p\sigma_1^{2p}\big)T^{p-1}\int_0^t \big(\mathbb{E}|\theta_s^{1,i_1}|^{2p}ds + \mathbb{E}|m_{L_1}^\alpha[\rho_s^N]|^{2p}\big)ds \\
&\quad + 2^{3(2p-1)}\big(\lambda_2^{2p}T^p + \sigma_2^{2p}p(2p-1)^p\big)T^pC_{\nabla L_1}^{2p},
\end{aligned} \tag{55}$$

for $i_1 \in [N_1]$. Similar computations give for $i_2 \in [N_2]$

$$\mathbb{E}|\theta_t^{2,i_2}|^{2p} \le 2^{2p-1}\mathbb{E}|\theta_0^{2,i_2}|^{2p} + 2^{4(2p-1)}\big(\lambda_1^{2p}T^p + p(2p-1)^p\sigma_1^{2p}\big)T^{p-1}\int_0^t \big(\mathbb{E}|\theta_s^{2,i_2}|^{2p} + \mathbb{E}|m_{L_2}^\alpha[\rho_s^N]|^{2p}\big)ds$$
$$+ 2^{3(2p-1)}\big(\lambda_2^{2p}T^p + \sigma_2^{2p}p(2p-1)^p\big)T^pC_{\nabla L_2}^{2p}.$$

Summing above two inequalities over $i_1 \in [N_1]$ and $i_2 \in [N_2]$, dividing by N, we have

$$\begin{aligned}\mathbb{E}\int|\theta|^{2p}d\rho_t^N \le{}& 2^{2p-1}\mathbb{E}\int|\theta|^{2p}d\rho_0^N \\ &+ 2^{4(2p-1)}\big(\lambda_1^{2p}T^p + p(2p-1)^p\sigma_1^{2p}\big)T^{p-1}\int_0^t\Big[\mathbb{E}\int|\theta|^{2p}d\rho_s^N \\ &\qquad + \frac{N_1}{N}\mathbb{E}|m_{L_1}^\alpha[\rho_s^N]|^{2p} + \frac{N_2}{N}\mathbb{E}|m_{L_2}^\alpha[\rho_s^N]|^{2p}\Big]ds \\ &+ 2^{3(2p-1)}\big(\lambda_2^{2p}T^p + \sigma_2^{2p}p(2p-1)^p\big)T^p\big(\frac{N_1}{N}C_{\nabla L_1}^{2p} + \frac{N_2}{N}C_{\nabla L_2}^{2p}\big).\end{aligned} \tag{56}$$

As shown in Lemma 10, for loss functions L_1, L_2 satisfying Assumption 1 and either bounded above or growing quadratically at infinity, we have

$$|m_{L_k}^\alpha[\rho_s^N]|^2 \le \int|\theta|^2d\eta_{k,s}^{\alpha,N} \le c_1 + c_2\int|\theta|^2d\rho_s^N, \tag{57}$$

for $k = 1, 2$ and appropriate constants c_1, c_2 independent of N, where by construction $m_{L_k}^\alpha[\rho_s^N] = \int\theta d\eta_{k,t}^{\alpha,N}$ with $\eta_{k,t}^{\alpha,N} = \frac{w_{L_k}^\alpha\rho_t^N}{\|w_{L_k}^\alpha\|_{\mathbb{L}^1(\rho_t^N)}}$. Therefore, we further obtain

$$|m_{L_k}^\alpha[\rho_s^N]|^{2p} \le \big(c_1 + c_2\int|\theta|^2d\rho_s^N\big)^p \le 2^{p-1}\big(c_1^p + c_2^p\int|\theta|^{2p}d\rho_s^N\big).$$

Inserting above inequalities into (56) and applying the Grönwall's inequality provides a constant $K_p > 0$, independent of N, such that $\sup_{t\in[0,T]}\mathbb{E}\int|\theta|^{2p}d\rho_t^N \le K_p$ holds, and consequently also

$$\sup_{t\in[0,T]}\mathbb{E}|m_{L_k}^\alpha[\rho_t^N]|^{2p} \le 2^{p-1}(c_1^p + c_2^pK_p),$$

which concludes the proof of the estimates in (53) by choosing K sufficiently large. The other two estimates easily follow by (57) and by applying the Grönwall's inequality on (55), respectively. ■

Appendix B. Auxiliary Propositions and Lemma for Well-posedness of System

Lemmas 8, 9 and 10 are direct consequences of Lemmas 3.1, 3.2 and 3.3 in (Carrillo et al., 2018). We state these results here for completeness and refer the reader to (Carrillo et al., 2018) for detailed proofs.

B.1 Loss Functions Bounded Above

Lemma 8 *Let L_1, L_2 satisfy Assumption 1 and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\int|\theta|^2d\mu \le K$. Then*

$$\frac{e^{-\alpha\underline{L_1}}}{\|w_{L_1}^\alpha\|_{\mathbb{L}^1(\mu)}} \le \exp(\alpha C_{L_1}(1+K)) =: C_{K_1}, \qquad \frac{e^{-\alpha\underline{L_2}}}{\|w_{L_2}^\alpha\|_{\mathbb{L}^1(\mu)}} \le \exp(\alpha C_{L_2}(1+K)) =: C_{K_2}$$

Lemma 9 *Let L_1, L_2 satisfy Assumption 1 and $\mu, \hat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ with $\int |\theta|^4 d\mu, \int |\hat{\theta}|^4 d\hat{\mu} \leq K$. Then for $k = 1, 2$, the following stability estimates hold*

$$|m^\alpha_{L_k}[\mu] - m^\alpha_{L_k}[\hat{\mu}]| \leq c_{0,k} W_2(\mu, \hat{\mu}),$$

with constants $c_{0,k} > 0$ depending only on $\alpha, C_{\nabla L_k}$ and K.

Proof By the assumption (10c), we know that the objective functions L_k is Lipschitz, i.e.

$$|L_k(\theta) - L_k(\hat{\theta})| \leq M_{\nabla L_k} |\theta - \hat{\theta}|$$

Then the proof follows (Carrillo et al., 2018, Lemma 3.2). ■

Now we are ready to prove the Lipschitz property of the consensus operator $m^\alpha_{L_k}[\nu]$ stated in Lemma 2.

Lemma 2 *Let objective functions L_1, L_2 satisfy Assumption 1 and let $\nu^1, \nu^2 \in \mathcal{C}([0,T], \mathcal{P}_2(\mathbb{R}^d))$ be such that $\sup_{t\in[0,T]} \int |\theta|^4 d\nu^1_t \leq K$, $\sup_{t\in[0,T]} \int |\theta|^4 d\nu^2_t \leq K$. Let us denote by $\overline{L}_1, \overline{L}_2$ the supremum of each of the loss functions. Let $w_1, w_2 > 0$ be such that $w_1 + w_2 = 1$, and let $\nu := w_1\nu^1 + w_2\nu^2$. Then, for all $s, t \in (0, T)$, the following stability estimates hold*

$$\left|m^\alpha_{L_k}[\nu_t] - m^\alpha_{L_k}[\nu_s]\right| \leq C\left(\sqrt{w_1} W_2\left(\nu^1_t, \nu^1_s\right) + \sqrt{w_2} W_2\left(\nu^2_t, \nu^2_s\right)\right), \tag{29}$$

for $k = 1, 2$ and for a constant C that depends only on $\alpha, C_{\nabla L_k}$ and K.

Proof By Lemma 9, we have

$$|m^\alpha_{L_k}[\nu_t] - m^\alpha_{L_k}[\nu_s]| \leq c_{0,k} W_2(\nu_t, \nu_s). \tag{58}$$

On the other hand, we know that

$$\begin{aligned}
W_2(\nu_t, \nu_s) &= \left(\inf_{\pi\in\Gamma(\nu_t,\nu_s)} \int_{\mathbb{R}^d\times\mathbb{R}^d} |\theta - \hat{\theta}|^2 \pi(d\theta, d\hat{\theta})\right)^{1/2} \\
&\leq \left(\inf_{\substack{\pi_1\in\Gamma(\nu^1_t,\nu^1_s),\\ \pi_2\in\Gamma(\nu^2_t,\nu^2_s)}} \int_{\mathbb{R}^d\times\mathbb{R}^d} |\theta - \hat{\theta}|^2 (w_1\pi_1 + w_2\pi_2)(d\theta, d\hat{\theta})\right)^{1/2} \\
&= \left(w_1 \inf_{\pi_1\in\Gamma(\nu^1_t,\nu^1_s)} \int_{\mathbb{R}^d\times\mathbb{R}^d} |\theta - \hat{\theta}|^2 \pi_1(d\theta, d\hat{\theta}) + w_2 \inf_{\pi_2\in\Gamma(\nu^2_t,\nu^2_s)} \int_{\mathbb{R}^d\times\mathbb{R}^d} |\theta - \hat{\theta}|^2 \pi_2(d\theta, d\hat{\theta})\right)^{1/2} \\
&\leq \sqrt{w_1} W_2(\nu^1_t, \nu^1_s) + \sqrt{w_2} W_2(\nu^2_t, \nu^2_s)
\end{aligned} \tag{59}$$

We conclude the proof by combining inequalities (58) and (59). ■

B.2 Loss Functions with Quadratic Growth at Infinity

In this section, we will prove the Theorem 2 for the case that the objective functions have quadratic growth at infinity, i.e. there exist constants $M > 0$ and $c_{q_k} > 0$ such that $L_k(\theta) - \underline{L_k} \geq c_{q_k}|\theta|^2$ for all $|\theta| \geq M$.

Lemma 10 *Let L_1, L_2 satisfy Assumption 1 and have quadratic growth at infinity, and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Then for $k = 1, 2$*

$$\int |\theta|^2 d\eta_k^\alpha \leq b_{k,1} + b_{k,2} \int |\theta|^2 d\mu, \qquad \eta_k^\alpha = \frac{w_{L_k}^\alpha \mu}{\|w_{L_k}^\alpha\|_{\mathbb{L}^1(\mu)}},$$

with constants

$$b_{k,1} := M^2 + b_{k,2}, \qquad b_{k,2} = 2\frac{C_{L_k}}{C_{q_k}} \left(1 + \frac{1}{\alpha C_{q_k} M^2}\right).$$

Proof [Proof of Theorem 2] Here we provide the proof for the case of quadratic growth at infinity. Since steps 1, 2 and 4 remain the same, we only show step 3.
Step 3: Let $(u^1, u^2) \in \mathcal{C}([0,T], \mathbb{R}^d) \times \mathcal{C}([0,T], \mathbb{R}^d)$ satisfy $(u^1, u^2) = \tau \mathcal{T}(u^1, u^2)$ for $\tau \in [0,1]$. In particular, there exists $\nu_1, \nu_2 \in \mathcal{C}([0,T], \mathcal{P}_2(\mathbb{R}^d))$ satisfying (31a) and (31b) respectively such that $(u^1, u^2) = \tau(m_{L_1}^\alpha[\nu], m_{L_2}^\alpha[\nu])$, where $\nu = w_1\nu_1 + w_2\nu_2$. From Lemma 10 and Jensen's inequality, we have

$$|u_t^1|^2 = \tau^2 |m_{L_1}^\alpha[\nu_t]|^2 \leq \tau^2 \left(b_{1,1} + b_{1,2} \int |\theta|^2 d\nu_t\right) \leq \tau^2 \left[b_{1,1} + b_{1,2} \left(w_1 \int |\theta|^2 d\nu_t^1 + w_2 \int |\theta|^2\right) d\nu_t^2\right]. \tag{60}$$

Therefore, a similar computation of the second moment estimate as in bounded case gives

$$\begin{aligned}
\frac{d}{dt} \int |\theta|^2 d\nu_t^1 &\leq (d\sigma_1^2 - 2\lambda_1 + |\gamma| + \lambda_2) \int |\theta|^2 d\nu_t^1 + (d\sigma_1^2 + |\gamma|)|u_t^1|^2 + (\lambda_2 + d\sigma_2^2) C_{\nabla L_1}^2 \\
&\leq (d\sigma_1^2 + |\gamma| + \lambda_2)(1 + b_{1,2}) \int |\theta|^2 d\nu_t^1 + (d\sigma_1^2 + |\gamma| + \lambda_2) b_{1,2} \int |\theta|^2 d\nu_t^2 \\
&\quad + (d\sigma_1^2 + |\gamma| + \lambda_2) b_{1,1} + (\lambda_2 + d\sigma_2^2) C_{\nabla L_1}^2.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\frac{d}{dt} \int |\theta|^2 d\nu_t^2 &\leq (d\sigma_1^2 + |\gamma| + \lambda_2) b_{2,2} \int |\theta|^2 d\nu_t^1 + (d\sigma_1^2 + |\gamma| + \lambda_2)(1 + b_{2,2}) \int |\theta|^2 d\nu_t^2 \\
&\quad + (d\sigma_1^2 + |\gamma| + \lambda_2) b_{2,1} + (\lambda_2 + d\sigma_2^2) C_{\nabla L_2}^2.
\end{aligned}$$

Adding above two inequalities gives

$$\frac{d}{dt}\left(\int |\theta|^2 d\nu_t^1 + \int |\theta|^2 d\nu_t^2\right) \leq C_1 \left(\int |\theta|^2 d\nu_t^1 + \int |\theta|^2 d\nu_t^2\right) + C_2.$$

Then the Grönwall's inequality yields

$$\int |\theta|^2 d\nu_t^1 + \int |\theta|^2 d\nu_t^2 \leq \exp(C_1 t)\left(\int |\theta|^2 d\nu_t^1 + \int |\theta|^2 d\nu_t^2\right) + \frac{C_2}{C_1}\big(\exp(C_1 t) - 1\big).$$

Consequently, we know that $\|u^1\|_\infty$ is bounded via (60). Similar bound also hold for $\|u^2\|_\infty$. Then we conclude the proof by the argument as in **Step 3** for the bounded case. ■

Appendix C. Auxiliary Lemma for Large-time Behavior of the Mean-field Particle System

In this section, we prove Lemmas 3, 4 and 5 that we used to derive the convergence to the global minimizers in the mean-field law as stated in Section 4.

Lemma 3 (Evolution of energy functional $\mathcal{V}$) *For $k = 1, 2$, let objective functions $L_k : \mathbb{R}^d \to \mathbb{R}$ and fix $\alpha, \lambda_1, \lambda_2, \sigma_1, \sigma_2 > 0$. Moreover, let $T > 0$ and $\rho^1, \rho^2 \in \mathcal{C}([0,T], \mathcal{P}_4(\mathbb{R}))$ form the weak solution to the Fokker-Planck equations* (8a) *and* (8b)*. Then the functional $\mathcal{V}(\rho_t^k)$ satisfies*

$$\begin{aligned}\frac{d}{dt}\mathcal{V}(\rho_t^k) \leq &-\big(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2\big)\mathcal{V}(\rho_t^k)\\ &+ \sqrt{2}(\lambda_1 + d\sigma_1^2)|m_{L_k}^\alpha[\rho_t] - \theta_k^*|\sqrt{\mathcal{V}(\rho_t^k)} + \frac{d\sigma_1^2}{2}|m_{L_k}^\alpha[\rho_t^k] - \theta_k^*|^2,\end{aligned}$$

with $M := \max\{M_{\nabla L_1}, M_{\nabla L_2}\}$.

Proof Since test function $\phi(\theta) := \frac{1}{2}|\theta - \theta^*|^2$ is in $\mathcal{C}_*^2(\mathbb{R}^d)$ and ρ^1 is the weak solution of the Fokker-Planck equation (8a) (see Remark 1), then the evolution of $\mathcal{V}(\rho_t^1)$ reads

$$\begin{aligned}\frac{d}{dt}\mathcal{V}(\rho_t^1) &= \frac{1}{2}\frac{d}{dt}\int |\theta - \theta_1^*|^2 d\rho_t^1(\theta)\\ &= -\lambda_1 \int (\theta - m_{L_1}^\alpha[\rho_t]) \cdot (\theta - \theta_1^*) d\rho_t^1 - \lambda_2 \int \nabla L_1(\theta) \cdot (\theta - \theta_1^*) d\rho_t^1\\ &\quad + \frac{d\sigma_1^2}{2}\int |\theta - m_{L_1}^\alpha[\rho_t]|^2 d\rho_t^1 + \frac{d\sigma_2^2}{2}\int |\nabla L_1(\theta)|^2 d\rho_t^1\\ &=: T_1 + T_2 + T_3 + T_4.\end{aligned}$$

Expanding the right-hand side of the inner product in the integral of T_1 by subtracting and adding θ_1^* yields

$$\begin{aligned}T_1 &= -\lambda \int |\theta - \theta_1^*|^2 d\rho_t^1(\theta) - \lambda \int (\theta_1^* - m_{L_1}^\alpha[\rho_t]) \cdot (\theta - \theta_1^*) d\rho_t^1(\theta)\\ &\leq -2\lambda \mathcal{V}(\rho_t^1) + \lambda |\mathbb{E}[\rho_t^1] - \theta_1^*||m_{L_1}^\alpha[\rho_t] - \theta_1^*|,\end{aligned} \tag{61}$$

where the last step is due to Cauchy-Schwarz inequality. Also note that

$$|\mathbb{E}[\rho_t^1] - \theta_1^*| \leq \int |\theta - \theta_1^*| d\rho_t^1(\theta) \leq \sqrt{\int |\theta - \theta_1^*|^2 d\rho_t^1(\theta)} = \sqrt{2\mathcal{V}(\rho_t^1)}. \tag{62}$$

Hence, we get

$$T_1 \leq -2\lambda_1 \mathcal{V}(\rho_t^1) + \lambda_1 \sqrt{2\mathcal{V}(\rho_t)}|m_{L_1}^\alpha[\rho_t] - \theta_1^*|$$

For term T_2, by the fact $\nabla L_1(\theta_1^*) = 0$ and Assumption (10b) one can compute

$$T_2 = -\lambda_2 \int \nabla L_1(\theta) \cdot (\theta - \theta_1^*) d\rho_t^1 \leq \lambda_2 M_{\nabla L_1} \int |\theta - \theta_1^*|^2 d\rho_t^1 = 2\lambda_2 M_{\nabla L_1} \mathcal{V}(\rho_t^1).$$

For term T_3, again by subtracting and adding θ_1^*, we have

$$\begin{aligned}T_2 &= \frac{d\sigma^2}{2}\int |\theta - m_{L_1}^\alpha[\rho_t]|^2 d\rho_t^1(\theta)\\ &= \frac{d\sigma^2}{2}\bigg(\int |\theta - \theta_1^*|^2 d\rho_t^1(\theta) - 2\int (\theta - \theta_1^*) \cdot (m_{L_1}^\alpha[\rho_t] - \theta_1^*) d\rho_t^1(\theta) + |m_{L_1}^\alpha[\rho_t] - \theta_1^*|^2\bigg)\\ &\leq d\sigma^2\bigg(\mathcal{V}(\rho_t^1) + |m_{L_1}^\alpha[\rho_t] - \theta_1^*| \int |\theta - \theta_1^*| d\rho_t^1(v) + \frac{1}{2}|m_{L_1}^\alpha[\rho_t] - \theta_1^*|^2\bigg),\end{aligned} \tag{63}$$

with Cauchy-Schwarz inequality being used in the last step. For term T_4, again by $\nabla L_1(\theta_1^*) = 0$ and Assumption (10b) one can compute

$$T_4 = \frac{d\sigma_2^2}{2}\int |\nabla L_1(\theta)|^2 d\rho_t^1(\theta) \leq \frac{d\sigma_2^2}{2}\int M_{\nabla L_1}^2 |\theta - \theta_1^*|^2 d\rho_t^1(\theta) = d\sigma_2^2 M_{\nabla L_1}^2 \mathcal{V}(\rho_t^1).$$

Therefore, combining the estimations of T_1, T_2, T_3 and T_4, we get

$$\begin{aligned}\frac{d}{dt}\mathcal{V}(\rho_t^1) \leq &-\big(2\lambda_1 - 2\lambda_2 M_{\nabla L_1} - d\sigma_1^2 - d\sigma_2^2 M_{\nabla L_1}^2\big)\mathcal{V}(\rho_t^1)\\ &+ \sqrt{2}(\lambda_1 + d\sigma_1^2)|m_{L_1}^\alpha[\rho_t] - \theta_1^*|\sqrt{\mathcal{V}(\rho_t^1)} + \frac{d\sigma_1^2}{2}|m_{L_1}^\alpha[\rho_t^1] - \theta_1^*|^2.\end{aligned}$$

The computations for $\mathcal{V}(\rho_t^2)$ are similar and hence we get the upper bound as in the statement. ■

Lemma 4 (Quantitative Laplace principle) *For $k = 1, 2$, denote $\underline{L_k} := \inf_{\theta\in\mathbb{R}^d} L_k(\theta)$. Let $\rho \in \mathcal{P}(\mathbb{R}^d)$ and fix $\alpha > 0$. For any $r > 0$, we define $L_r^k := \sup_{\theta\in B_r(\theta_k^*)} L_k(\theta)$. Then, under Assumption 2, for any $r \in (0, \min\{R_0^1, R_0^2\}]$ and $q_k > 0$ such that $q_k + L_r^k - \underline{L_k} \leq L_\infty^k$, we have*

$$|m_{L_k}^\alpha[\rho] - \theta_k^*| \leq \frac{(q_k + L_r^k - \underline{L_k})^{\nu_k}}{\eta_k} + \frac{\exp(-\alpha q_k)}{\rho(B_r(\theta_k^*))}\int |\theta - \theta_k^*| d\rho(\theta).$$

Proof The same as the proof of Proposition 21 in (Fornasier et al., 2024a). ■

Definition 2 (Mollifier) *For $k = 1, 2$, $r > 0$, we define the mollifiers $\phi_r^k : \mathbb{R}^d \to \mathbb{R}$ by*

$$\phi_r^k(\theta) := \begin{cases}\exp\left(-\frac{r^2}{r^2 - |\theta - \theta_k^*|^2}\right), & \text{if } \|\theta - \theta_k^*\|_2 < r,\\ 0, & \text{else}\end{cases} \tag{64}$$

We have $\phi_t^k(\theta_k^*) = 1$, $Im(\phi_t^k) = [0, 1]$, $supp(\phi_r^k) = B_r(\theta_k^*), \phi_r^k \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ and

$$\begin{aligned}\nabla\phi_r^k(\theta) &= -2r^2\frac{\theta - \theta_k^*}{(r^2 - |\theta - \theta_k^*|^2)^2}\phi_r^k(\theta),\\ \Delta\phi_r^k(\theta) &= 2r^2\left(\frac{2(2|\theta - \theta_k^*|^2 - r^2)|\theta - \theta_k^*|^2 - d(r^2 - |\theta - \theta_k^*|^2)^2}{(r^2 - |\theta - \theta_k^*|^2)^4}\right)\phi_r^k(\theta).\end{aligned}$$

Lemma 5 *For $k = 1, 2$, let $T > 0, r > 0$, and fix parameters $\alpha, \lambda_1, \lambda_2, \sigma_1, \sigma_2 > 0$. Assume $\rho^1, \rho^2 \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ weakly solve the Fokker-Planck equations (8a) and (8b) respectively with initial conditions $\rho_0^1, \rho_0^2 \in \mathcal{P}(\mathbb{R}^d)$. Furthermore, denote $B_k := \sup_{t\in[0,T]} |m_{L_k}^\alpha[\rho_t] - \theta_k^*|$. Then for all $t \in [0, T]$ we have*

$$\rho_t^k(B_r(\theta_k^*)) \geq \big(\int \phi_r^k(\theta) d\rho_0^k(\theta)\big)\exp\big(-(q_k^l + q_k^g)t\big),$$

with the functions ϕ_r^1 and ϕ_r^2 as in (64) and

$$\begin{aligned}q_k^l &:= \max\left\{\frac{2\lambda_1(\sqrt{c}r + B_k)\sqrt{c}}{(1-c)^2 r} + \frac{2\sigma_1^2(cr^2 + B_k^2)(2c + d)}{(1-c)^4 r^2}, \frac{4\lambda_1^2}{(2c-1)\sigma_1^2}\right\},\\ q_k^g &:= \max\left\{\frac{2\lambda_2 c M_{\nabla L_k}}{(1-c)^2} + \frac{\sigma_2^2 M_{\nabla L_k}^2 c(2c + d)}{(1-c)^4}, \frac{4\lambda_2^2}{(2c-1)\sigma_2^2}\right\},\end{aligned}$$

where $c \in (\frac{1}{2}, 1)$ can be any constant that satisfies the inequality

$$(2c - 1)c \geq d(1 - c)^2. \tag{35}$$

Proof Here we will prove the case for $\rho_t^k(B_r(\theta_1^*))$, the computation for the other one is similar. By the properties of the mollifier in Definition 2 we have $0 \leq \phi_r^1(\theta) \leq 1$ and $\mathrm{supp}(\phi_r^1) = B_r(\theta_1^*)$. This implies $\rho_t^1(B_r(\theta_1^*)) = \rho_t^1\big(\{\theta \in \mathbb{R}^d : |\theta - \theta_1^*| \leq r\}\big) \geq \int \phi_r^1(\theta) d\rho_t^1(\theta)$. Similar to the proof in ((Fornasier et al., 2024a; Riedl, 2023)), we will derive a lower bound for the right-hand side of this inequality. Since ρ^1 is the weak solution of (8a) and $\phi_r^1 \in \mathcal{C}_c^\infty(\mathbb{R}^d)$, we have

$$\frac{d}{dt}\int \phi_r^1(\theta) d\rho_t^1(\theta) = \int \big(T_1(\theta) + T_2(\theta) + T_3(\theta) + T_4(\theta)\big) d\rho_t^1(\theta),$$

with

$$\begin{aligned} T_1(\theta) &:= -\lambda_1(\theta - m_{L_1}^\alpha[\rho_t]) \cdot \nabla \phi_r^1(\theta), \qquad T_2(\theta) := -\lambda_2 \nabla L_1(\theta) \cdot \nabla \phi_r^1(\theta), \\ T_3(\theta) &:= \frac{\sigma_1^2}{2} |\theta - m_{L_1}^\alpha[\rho_t]|^2 \Delta \phi_r^1(\theta), \qquad T_4(\theta) := \frac{\sigma_2^2}{2} |\nabla L_1(\theta)|^2 \Delta \phi_r^1(\theta). \end{aligned}$$

From the proof in ((Fornasier et al., 2024a; Riedl, 2023)), we know that

$$T_1(\theta) + T_3(\theta) \geq -q_1^l \phi_r^1(\theta) \qquad \text{for all } \theta \in \mathbb{R}^d, \tag{65}$$

where

$$q_1^l := \max\left\{ \frac{2\lambda_1(\sqrt{c}r + B_1)\sqrt{c}}{(1-c)^2 r} + \frac{2\sigma_1^2(cr^2 + B_1^2)(2c+d)}{(1-c)^4 r^2}, \frac{4\lambda_1^2}{(2c-1)\sigma_1^2} \right\}.$$

Now we aim to show that $T_2(\theta) + T_4(\theta) \geq -q_1^g \phi_r^1(\theta)$ holds for all $\theta \in \mathbb{R}^d$ and some constants $q_1^g > 0$. Since the mollifier ϕ_r^1 and its first and second derivatives vanish outside of $\Omega_r := \{\theta \in \mathbb{R}^d : |\theta - \theta_1^*| < r\}$ we can restrict our attention to the open ball Ω_r. To achieve the lower bound over Ω_r, we introduce the subsets

$$\begin{aligned} K_1 &:= \left\{ \theta \in \mathbb{R}^d : |\theta - \theta_1^*| > \sqrt{c}r \right\}, \\ K_2 &:= \left\{ \theta \in \mathbb{R}^d : -\lambda_2 \nabla L_1(\theta) \cdot (\theta - \theta_1^*)(r^2 - |\theta - \theta_1^*|^2)^2 > (2c-1) r^2 \frac{\sigma_2^2}{2} |\nabla L_1(\theta)|^2 |\theta - \theta_1^*|^2 \right\}, \end{aligned}$$

where c is the constant adhering to (35). We now decompose Ω_r according to

$$\Omega_r = (K_1^c \cap \Omega_r) \cup (K_1 \cap K_2^c \cap \Omega_r) \cup (K_1 \cap K_2 \cap \Omega_r).$$

In the following we treat each of these three subsets respectively.
Subset $K_1^c \cap \Omega_r$: On this subset we have $|\theta - \theta_1^*| \leq \sqrt{c}r$, then one can compute

$$\begin{aligned} T_2(\theta) &= 2\lambda_2 r^2 \frac{\nabla L_1(\theta) \cdot (\theta - \theta_1^*)}{(r^2 - |\theta - \theta_1^*|^2)^2} \phi_r^1(\theta) \\ &\geq -2\lambda_2 r^2 \frac{M_{\nabla L_1} |\theta - \theta_1^*|^2}{(r^2 - |\theta - \theta_1^*|^2)^2} \phi_r^1(\theta) \\ &\geq -2\lambda_2 r^2 \frac{M_{\nabla L_1} c r^2}{(1-c)^2 r^4} \phi_r^1(\theta) \\ &= \frac{-2c\lambda_2 M_{\nabla L_1}}{(1-c)^2} \phi_r^1(\theta) =: -q^{g,1} \phi_r^1(\theta). \end{aligned}$$

For term T_4, we deduce

$$\begin{aligned}T_4(\theta) &= \frac{\sigma_2^2}{2}|\nabla L_1|^2 2r^2\left(\frac{2(2|\theta-\theta_1^*|^2-r^2)|\theta-\theta_1^*|^2 - d(r^2-|\theta-\theta_1^*|^2)^2}{(r^2-|\theta-\theta_1^*|^2)^4}\right)\phi_r^1(\theta)\\ &\leq -\sigma_2^2 M_{\nabla L_1}^2|\theta-\theta_1^*|^2 r^2\left(\frac{2c(2c-1)r^4 - d(1-c)^2r^4}{(1-c)^4 r^8}\right)\phi_r^1(\theta)\\ &\leq -\frac{\sigma_2^2 M_{\nabla L_1}^2 c(2c+d)}{(1-c)^4}\phi_r^1(\theta) =: -q^{g,2}\phi_r^1(\theta).\end{aligned}$$

Subset $K_1 \cap K_2^c \cap \Omega_r$: By the definition of K_1 and K_2 we have $|\theta-\theta_1^*| > \sqrt{c}r$ and

$$-\lambda_2 \nabla L_1(\theta)\cdot(\theta-\theta_1^*)(r^2-|\theta-\theta_1^*|^2)^2 \leq (2c-1)r^2\frac{\sigma_2^2}{2}|\nabla L_1|^2|\theta-\theta_1^*|^2,$$

respectively. Our goal now is to show that $T_2(\theta)+T_4(\theta)\geq 0$ for all θ in this subset. We first compute

$$\begin{aligned}\frac{T_2(\theta)+T_4(\theta)}{2r^2\phi_r^1(\theta)} &= \frac{\lambda_2\nabla L_1(\theta)\cdot(\theta-\theta_1^*)(r^2-|\theta-\theta_1^*|^2)^2}{(r^2-|\theta-\theta_1^*|^2)^4}\\ &\quad + \frac{\sigma_2^2}{2}|\nabla L_1|^2\frac{2(2|\theta-\theta_1^*|^2-r^2)|\theta-\theta_1^*|^2 - d(r^2-|\theta-\theta_1^*|^2)^2}{(r^2-|\theta-\theta_1^*|^2)^4}.\end{aligned}$$

Therefore, we have $T_2(\theta)+T_4(\theta)\geq 0$ whenever we can show

$$\left(-\lambda_2\nabla L_1(\theta)\cdot(\theta-\theta_1^*)+\frac{d\sigma_2^2}{2}|\nabla L_1|^2\right)(r^2-|\theta-\theta_1^*|^2)^2 \leq \sigma_2^2|\nabla L_1|^2\big(2|\theta-\theta_1^*|^2-r^2\big)|\theta-\theta_1^*|^2.$$

The first term on the left-hand side can be bounded above by

$$\begin{aligned}-\lambda_2\nabla L_1(\theta)\cdot(\theta-\theta_1^*)(r^2-|\theta-\theta_1^*|^2)^2 &\leq (2c-1)r^2\frac{\sigma_2^2}{2}|\nabla L_1(\theta)|^2|\theta-\theta_1^*|^2\\ &\leq \frac{\sigma_2^2}{2}|\nabla L_1|^2(2|\theta-\theta_1^*|^2-r^2)|\theta-\theta_1^*|^2.\end{aligned}$$

For the second term on the left-hand side, we can use $d(1-c)^2\leq(2c-1)c$ to get

$$\begin{aligned}\frac{d\sigma_2^2}{2}|\nabla L_1|^2(r^2-|\theta-\theta_1^*|^2)^2 &\leq \frac{d\sigma_2^2}{2}|\nabla L_1|^2(1-c)^2r^4\\ &\leq \frac{\sigma_2^2}{2}|\nabla L_1|^2(2c-1)r^2cr^2\\ &\leq \frac{\sigma_2^2}{2}|\nabla L_1|^2(2|\theta-\theta_1^*|^2-r^2)|\theta-\theta_1^*|^2.\end{aligned}$$

Hence we have $T_2(\theta)+T_4(\theta)\geq 0$ uniformly on this subset.
Subset $K_1\cap K_2\cap\Omega_r$: On this subset we have $|\theta-\theta_1^*|>\sqrt{c}r$ and

$$-\lambda_2\nabla L_1(\theta)\cdot(\theta-\theta_1^*)(r^2-|\theta-\theta_1^*|^2)^2 > (2c-1)r^2\frac{\sigma_2^2}{2}|\nabla L_1(\theta)|^2|\theta-\theta_1^*|^2.$$

We first note that $T_2(\theta) = 0$ whenever $\lambda_2^2|\nabla L_1(\theta)|^2 = 0$ provided that $\lambda_2 > 0$. On the other hand, if $\lambda_2^2|\nabla L_1(\theta)|^2 > 0$, one can compute

$$\begin{aligned}
T_2(\theta) &= 2\lambda_2 r^2 \frac{\nabla L_1(\theta)\cdot(\theta-\theta_1^*)}{(r^2-|\theta-\theta_1^*|^2)^2}\phi_r^1(\theta)\\
&\geq 2\lambda_2 r^2 \frac{-\lambda_2(\nabla L_1(\theta)\cdot(\theta-\theta_1^*))^2}{(2c-1)r^2\frac{\sigma_2^2}{2}|\nabla L_1(\theta)|^2|\theta-\theta_1^*|^2}\phi_r^1(\theta)\\
&\geq -2\lambda_2^2 r^2 \frac{|\nabla L_1(\theta)|^2|\theta-\theta_1^*|^2}{(2c-1)r^2\frac{\sigma_2^2}{2}|\nabla L_1(\theta)|^2|\theta-\theta_1^*|^2}\phi_r^1(\theta)\\
&= -\frac{4\lambda_2^2}{(2c-1)\sigma_2^2}\phi_r^1(\theta) =: -q^{g,3}\phi_r^1(\theta).
\end{aligned}$$

For term T_4, we deduce

$$\begin{aligned}
T_4(\theta) &= \frac{\sigma_2^2}{2}|\nabla L_1|^2 2r^2\left(\frac{2(2|\theta-\theta_1^*|^2-r^2)|\theta-\theta_1^*|^2 - d(r^2-|\theta-\theta_1^*|^2)^2}{(r^2-|\theta-\theta_1^*|^2)^4}\right)\phi_r^1(\theta)\\
&\geq \frac{\sigma_2^2}{2}|\nabla L_1|^2 2r^2\left(\frac{2(2c-1)r^2cr^2 - d(1-c)^2r^4}{(r^2-|\theta-\theta_1^*|^2)^4}\right)\phi_r^1(\theta)\\
&= \frac{\sigma_2^2}{2}|\nabla L_1|^2 2r^2\left(\frac{[2(2c-1)c - d(1-c)^2]r^4}{(r^2-|\theta-\theta_1^*|^2)^4}\right)\phi_r^1(\theta) \geq 0,
\end{aligned}$$

provided c satisfies $2(2c-1)c \geq d(1-c)^2$.

Concluding the proof: From the above computations, one can obtain

$$\begin{aligned}
\int \big(T_2(\theta)+T_4(\theta)\big)d\rho_t^1(\theta) &= \int_{K_1^c\cap\Omega_r} \underbrace{\big(T_2(\theta)+T_4(\theta)\big)}_{\geq -(q^{g,1}+q^{g,2})\phi_r^1(\theta)} d\rho_t^1(\theta) + \int_{K_1\cap K_2^c\cap\Omega_r} \underbrace{\big(T_2(\theta)+T_4(\theta)\big)}_{\geq 0} d\rho_t^1(\theta)\\
&\quad + \int_{K_1\cap K_2\cap\Omega_r} \underbrace{\big(T_2(\theta)+T_4(\theta)\big)}_{-q^{g,3}\phi_r^1(\theta)} d\rho_t^1(\theta)\\
&\geq \int -q_1^g\phi_r^1(\theta)d\rho_t^1(\theta),
\end{aligned}$$

where

$$q_1^g := \max\Big\{q^{g,1}+q^{g,2}, q^{g,3}\Big\} = \max\left\{\frac{2\lambda_2 c M_{\nabla L_1}}{(1-c)^2} + \frac{\sigma_2^2 M_{\nabla L_1}^2 c(2c+d)}{(1-c)^4}, \frac{4\lambda_2^2}{(2c-1)\sigma_2^2}\right\}.$$

Combining above estimation with (65), we get

$$\frac{d}{dt}\int \phi_r^1(\theta)d\rho_t^1(\theta) \geq -(q_1^l+q_1^g)\int \phi_r^1(\theta)d\rho_t^1(\theta).$$

By applying Grönwall's inequality and multiplying both sides (-1) gives

$$\int \phi_r^1(\theta)d\rho_t^1(\theta) \geq \left(\int \phi_r^1(\theta)d\rho_0^1(\theta)\right)\exp\big(-(q_1^l+q_1^g)t\big)$$

. Hence, we conclude

$$\rho_t^1(B_r(\theta_1^*)) \geq \left(\int \phi_r^1(\theta)d\rho_0^1(\theta)\right)\exp\big(-(q_1^l+q_1^g)t\big).$$

■

Appendix D. Auxiliary Lemmas for Large Time Behavior of Finite Particle System

In this section, we present the detailed proof of the non-asymptotic mean-field approximation of FedCBO system stated in Proposition 1. Before that, we first prove the auxiliary Lemma 6, which ensures that all considered stochastic processes stay bounded with high probability.

Lemma 6 *Let $\tilde{T} > 0$, $\rho_0 := w_1\rho_0^1 + w_2\rho_0^2 \in \mathcal{P}_4(\mathbb{R}^d)$ and let $N = N_1 + N_2 \in \mathbb{N}$ be fixed. Moreover, let $\{\theta_t^{1,i_1}\}_{i_1=1}^{N_1}, \{\theta_t^{2,i_2}\}_{i_2=1}^{N_2}$ be the solution of the finite interacting particle system* (4)*, and let $\{\bar{\theta}_t^{1,i_1}\}_{i_1=1}^{N_1}, \{\bar{\theta}_t^{2,i_2}\}_{i_2=1}^{N_2}$ denote independent copies of the solutions to the mean-field dynamics* (4a) *and* (4b)*, respectively. Then, under Assumption 1, for any $M > 0$ we have*

$$\mathbb{P}(\Omega_M) = \mathbb{P}\left(\sup_{t\in[0,\tilde{T}]} \frac{1}{N}\sum_{k=1,2}\sum_{i_k=1}^{N_k} \max\left\{\left|\theta_t^{k,i_k}(\omega)\right|^4, \left|\bar{\theta}_t^{k,i_k}(\omega)\right|^4\right\} \leq M\right) \geq 1 - \frac{2C_{Bound}}{M}, \tag{49}$$

where $C_{Bound} = C_{Bound}(\lambda_1, \lambda_2, \sigma_1, \sigma_2, \tilde{T}, b_{11}, b_{12}, b_{21}, b_{22})$ is a constant that is independent of N and d. Here, b_{11}, b_{12}, b_{21} and b_{22} are problem-dependent constants defined in Lemma 10.

Proof Denote $\rho_t^N := \frac{1}{N}\sum_{k=1,2}\sum_{i_k=1}^{N_k} \delta_{\theta_t^{k,i_k}}$ as the empirical measure associated to the processes $\{(\theta_t^{1,i_1})_{t\geq 0}\}_{i_1=1}^{N_1}, \{(\theta_t^{2,i_2})_{t\geq 0}\}_{i_2=1}^{N_2}$. Similarly, we let $\bar{\rho}_t^N := \frac{1}{N}\sum_{k=1,2}\sum_{i_k=1}^{N_k} \delta_{\bar{\theta}_t^{k,i_k}}$ to be the empirical measure corresponding to the processes $\{(\bar{\theta}_t^{1,i_1})_{t\geq 0}\}_{i_1=1}^{N_1}, \{(\bar{\theta}_t^{2,i_2})_{t\geq 0}\}_{i_2=1}^{N_2}$. By Markov's inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[0,T]} \frac{1}{N}\sum_{k=1,2}\sum_{i_k=1}^{N_k} \max\left\{\left|\theta_t^{k,i_k}\right|^4, \left|\bar{\theta}_t^{k,i_k}\right|^4\right\} > M\right) \\
&\quad\leq \frac{\mathbb{E}\left[\sup_{t\in[0,T]} \frac{1}{N}\sum_{k=1,2}\sum_{i_k=1}^{N_k} \max\left\{\left|\theta_t^{k,i_k}\right|^4, \left|\bar{\theta}_t^{k,i_k}\right|^4\right\}\right]}{M} \\
&\quad\leq \frac{\mathbb{E}\left[\sup_{t\in[0,T]} \int |\theta|^4 d\rho_t^N(\theta)\right] + \mathbb{E}\left[\sup_{t\in[0,T]} \int |\theta|^4 d\bar{\rho}_t^N(\theta)\right]}{M}
\end{aligned}$$

In the following, we estimate the two terms in the numerator respectively. For $k = 1, 2$, $i_k \in [N_k]$ one can compute that

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in[0,T]} \left|\theta_t^{k,i_k}\right|^4\right] \lesssim\ & \mathbb{E}\left[\left|\theta_0^{k,i_k}\right|^4\right] + \lambda_1^4\mathbb{E}\left[\sup_{t\in[0,T]} \left|\int_0^t \left(\theta_t^{k,i_k} - m_{L_k}^\alpha[\rho_t^N]\right) d\tau\right|^4\right] \\
&+ \lambda_2^4\mathbb{E}\left[\sup_{t\in[0,T]} \left|\int_0^t \nabla L_k(\theta_\tau^{k,i_k}) d\tau\right|^4\right] \\
&+ \sigma_1^4\mathbb{E}\left[\sup_{t\in[0,T]} \left|\int_0^t \left|\theta_t^{k,i_k} - m_{L_k}^\alpha[\rho_t^N]\right| dB_t^{k,i_k}\right|^4\right] \\
&+ \sigma_2^4\mathbb{E}\left[\sup_{t\in[0,T]} \left|\int_0^t \nabla L_k(\theta_\tau^{k,i_k}) d\tilde{B}_\tau^{k,i_k}\right|^4\right]
\end{aligned} \tag{66}$$

By the regularity established in Lemma 7 and boundedness of $|\nabla L_k(\theta)|$ in Assumption 1, the last two terms in the above inequalities are martingale, which is a direct consequence of (Oksendal,

2013, Corollary 3.2.6). Then applying the Burkholder-Davis-Gurdy inequality yields

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_0^t\left|\theta_t^{k,i_k}-m_{L_k}^{\alpha}[\rho_t^N]\right|dB_t^{k,i_k}\right|^4\right]\lesssim\mathbb{E}\left[\int_0^T\left|\theta_\tau^{k,i_k}-m_{L_k}^{\alpha}[\rho_t^N]\right|^2d\tau\right]^2,$$
$$\mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_0^t\nabla L_k(\theta_\tau^{k,i_k})d\tilde{B}_\tau^{k,i_k}\right|^4\right]\lesssim\mathbb{E}\left[\int_0^T C_{\nabla L_k}^2d\tau\right]^2=T^2C_{\nabla L_k}^4\,.$$

For the third term in the inequality (66), we again use Assumption 1 to get

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_0^t\nabla L_k(\theta_\tau^{k,i_k})d\tau\right|^4\right]\le T^4C_{\nabla L_k}^4\,.$$

For the remaining terms, one may apply the same technique as in (Fornasier et al., 2022, Lemma 15) and obtain

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left|\theta_t^{k,i_k}\right|^4\right]\le C_k\left(1+\mathbb{E}\left|\theta_0^{k,i_k}\right|^4+\mathbb{E}\left[\int_0^T\left(\left|\theta_\tau^{k,i_k}\right|^4+\int|\theta|^4d\rho_\tau^N(\theta)\right)d\tau\right]\right),\tag{67}$$

with constants $C_k=C_k\left(\lambda_1,\lambda_2,\sigma_1,\sigma_2,T,C_{\nabla L_k},b_{11},b_{12},b_{21},b_{22}\right)$ for $k=1,2$. Averaging (67) over $k=1,2$ and $i_k\in[N_k]$ yields the estimate

$$\mathbb{E}\left[\sup_{t\in[0,T]}\int|\theta|^4d\rho_t^N(\theta)\right]\le C\left(1+\mathbb{E}\int|\theta|^4d\rho_0^N(\theta)+2\int_0^T\mathbb{E}\sup_{\hat{\tau}\in[0,\tau]}\int|\theta|^4d\rho_{\hat{\tau}}^N(\theta)d\tau\right).$$

Then by Grönwall's inequality, $\mathbb{E}\left[\sup_{t\in[0,T]}\int|\theta|^4d\rho_t^N(\theta)\right]$ is bounded by a constant K that is independent from number of particles N and dimension d. Similar arguments allow to show $\mathbb{E}\left[\sup_{t\in[0,T]}\int|\theta|^4d\bar{\rho}_t^N(\theta)\right]\le K$. Then we conclude the proof by the Markov's inequality. ■

We now present the proof of Proposition 1, which mainly follows the idea in (Fornasier et al., 2022, Proposition 16).

Proposition 1 (Quantitative Mean-field Approximation) *Under the same assumptions as in Lemma 6, for $k=1,2$, if $(\theta_t^{k,i_k})_{t\ge0}$ and $(\bar{\theta}_t^{k,i_k})_{t\ge0}$ share the same initial data as well as the Brownian motion paths $(B_t^{k,i_k})_{t\ge0},(\tilde{B}_t^{k,i_k})_{t\ge0}$ for all $i_k\in[N_k]$, then we have a probabilistic mean-field approximation of the form*

$$\max_{\substack{k=1,2,\\ i_k\in[N_k]}}\sup_{t\in[0,\tilde{T}]}\mathbb{E}\left[\left|\theta_t^{k,i_k}-\bar{\theta}_t^{k,i_k}\right|^2\,\middle|\,\Omega_M\right]\le C_{MFA}(N_1^{-1}+N_2^{-1}),\tag{50}$$

where $C_{MFA}:=C_{MFA}\left(\alpha,C_{L_1},C_{L_2},C_{\nabla L_1},C_{\nabla L_2},M,\mathcal{M}_2,b_{11},b_{12},b_{21},b_{22}\right)$, and $\mathcal{M}_2$ is an upper bound on the second moment of ρ_t^N uniformly over time $t\in[0,\tilde{T}]$.

Proof Let us define the cutoff process

$$I_M(t):=\begin{cases}1, & \text{if }\frac{1}{N}\sum_{k=1,2}\sum_{i_k=1}^{N_k}\max\left\{\left|\theta_\tau^{k,i_k}\right|^4,\left|\bar{\theta}_\tau^{k,i_k}\right|^4\right\}\le M\text{ for all }\tau\in[0,t],\\ 0, & \text{else,}\end{cases}$$

which is adapted to the natural filtration of the underlying Brownian motions and has the property $I_M(t) = I_M(t)I_M(\tau)$ for all $\tau \in [0,t]$. By Jensen's inequality and Itô isometry, we have for $k = 1, 2$, $i_k \in [N_k]$,

$$\begin{aligned}
\mathbb{E}\left[\left|\theta_t^{k,i_k} - \bar{\theta}_t^{k,i_k}\right|^2 I_M(t)\right] = \mathbb{E}\Bigg[\Bigg| & \lambda_1 \int_0^t \left(\left(\theta_\tau^{k,i_k} - m_{L_k}^\alpha[\rho_\tau^N]\right) - \left(\bar{\theta}_\tau^{k,i_k} - m_{L_k}^\alpha[\rho_\tau]\right)\right) d\tau \\
& + \lambda_2 \int_0^t \left(\nabla L_k(\theta_\tau^{k,i_k}) - \nabla L_k(\bar{\theta}_\tau^{k,i_k})\right) d\tau \\
& + \sigma_1 \int_0^t \left(\left(\theta_\tau^{k,i_k} - m_{L_k}^\alpha[\rho_\tau^N]\right) - \left(\bar{\theta}_\tau^{k,i_k} - m_{L_k}^\alpha[\rho_\tau]\right)\right) dB_\tau^{k,i_k} \\
& + \sigma_2 \int_0^t \left(\nabla L_k(\theta_\tau^{k,i_k}) - \nabla L_k(\bar{\theta}_\tau^{k,i_k})\right) d\tilde{B}_\tau^{k,i_k} \Bigg|^2 I_M(t)\Bigg] \\
\lesssim C_k \int_0^t \mathbb{E}\left[\left|\theta_\tau^{k,i_k} - \bar{\theta}_\tau^{k,i_k}\right|^2 + \left|m_{L_k}^\alpha[\rho_\tau^N] - m_{L_k}^\alpha[\rho_\tau]\right|^2 I_M(\tau) d\tau\right],
\end{aligned} \tag{68}$$

where $C_k := \left(\lambda_1^2 + \lambda_2^2 M_{\nabla L_k}^2\right) T + \sigma_1^2 + \sigma_2^2$. In what follows, let us denote the empirical measure of the processes $\{\bar{\theta}_\tau^{1,i_1}\}, \{\bar{\theta}_\tau^{2,i_2}\}$ by $\bar{\rho}_\tau^N$. Then by the same arguments as in the proofs of Lemma 2 and (Fornasier et al., 2021, Lemma 3.1), we have for $k = 1, 2$

$$\begin{aligned}
\mathbb{E}\left[\left|m_{L_k}^\alpha[\rho_\tau^N] - m_{L_k}^\alpha[\rho_\tau]\right|^2 I_M(\tau)\right] &\lesssim \mathbb{E}\left[\left|m_{L_k}^\alpha[\rho_\tau^N] - m_{L_k}^\alpha[\bar{\rho}_\tau^N]\right|^2 I_M(\tau)\right] \\
&\quad + \mathbb{E}\left[\left|m_{L_k}^\alpha[\bar{\rho}_\tau^N] - m_{L_k}^\alpha[\rho_\tau]\right|^2 I_M(\tau)\right] \\
&\leq \widetilde{C}_k \left(\max_{\substack{k=1,2,\\ i_k \in [N_k]}} \mathbb{E}\left[\left|\theta_\tau^{k,i_k} - \bar{\theta}_\tau^{k,i_k}\right|^2 I_M(\tau)\right] + \frac{1}{N_1} + \frac{1}{N_2} \right),
\end{aligned} \tag{69}$$

with constants $\widetilde{C}_k = \widetilde{C}_k\left(\alpha, C_{L_1}, C_{L_2}, C_{\nabla L_1}, C_{\nabla L_2}, M, \mathcal{M}_2, b_{11}, b_{12}, b_{21}, b_{22}\right)$, where $\mathcal{M}_2$ is a constant upper bound of the second moment of ρ_t^N uniformly in time $t \in [0,T]$. Note that $\mathbb{1}_{\Omega_M} \leq I_M(t)$ pointwise and for all $t \in [0,T]$. Then by plugging the above estimates into (68) and taking the maximum over $k = 1, 2$, $i_k \in [N_k]$, and further applying the Grönwall's inequality, we obtain the non-asymptotic mean-field approximation result in (50). ■

Appendix E. Additional Experiments

In this section, we present the results of an additional experiment where we compare the differences in performance of the dynamics with an isotropic noise term as originally introduced in (4) versus dynamics with an anisotropic noise term as proposed in (Carrillo et al., 2021). In particular, in Algorithm 2, we add the noise term to the aggregation step of local agents in (22) to yield a modified update rule: [6]

$$\theta_{n+1}^j \leftarrow \theta_n^j - \lambda_1 \gamma \left(\theta_n^j - m_j\right) + \sigma_1 \sqrt{\gamma} \left|\theta_n^j - m_j\right| z_j, \qquad z_j \sim \mathcal{N}(0, I_{d\times d}) \qquad \text{(isotropic noise)} \tag{70}$$

$$\theta_{n+1}^j \leftarrow \theta_n^j - \lambda_1 \gamma \left(\theta_n^j - m_j\right) + \sigma_1 \sqrt{\gamma} \left(\theta_n^j - m_j\right) \circ z_j, \quad z_j \sim \mathcal{N}(0, I_{d\times d}) \qquad \text{(anisotropic noise)} \tag{71}$$

6. We didn't include the noise term $\widetilde{B}^{k,i_k}$ as introduced in (4) since we use stochastic gradient descent in the local agent update in Algorithm 1, which already has added noise.

Table 2: Test accuracy ± standard deviation % on rotated MNIST using FedCBO with different noise terms (levels).

FedCBO (no noise)	Isotropic ($\sigma_1 = 0.01$)	Isotropic ($\sigma_1 = 0.05$)	Anisotropic ($\sigma_1 = 0.01$)	Anisotropic ($\sigma_1 = 0.05$)
96.51 ± 0.04	96.41 ± 0.02	10.49 ± 0.05	96.33 ± 0.01	96.34 ± 0.02

We follow the same experimental setting as described in Section 2.4, and use same hyperparameters in FedCBO as before. We then vary the value of the hyperparameter σ_1 in the noise term. We present the results in Table 2. We can observe that 1) when the noise level is small ($\sigma_1 = 0.01$), the performance of FedCBO with or without noise is similar; 2) when the noise level is relatively large ($\sigma_1 = 0.05$), FedCBO with isotropic noise fails while the anisotropic version behaves as in the small noise level case. This observation aligns with the findings in (Carrillo et al., 2021), as the impact of noise terms on the algorithm's performance is not significant when training a neural network on the MNIST dataset.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Ludwig Arnold. Stochastic differential equations. *New York*, 1974.

Hyeong-Ohk Bae, Seung-Yeal Ha, Myeongju Kang, Hyuncheul Lim, Chanho Min, and Jane Yoo. A constrained consensus based optimization algorithm and its application to finance. *Applied Mathematics and Computation*, 416:126726, 2022.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

Giacomo Borghi, Michael Herty, and Lorenzo Pareschi. A consensus-based algorithm for multi-objective optimization and its mean-field description. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 4131–4136. IEEE, 2022.

Giacomo Borghi, Michael Herty, and Lorenzo Pareschi. An adaptive consensus based method for multi-objective optimization with uniform pareto front approximation. *Applied Mathematics & Optimization*, 88(2):58, 2023a.

Giacomo Borghi, Michael Herty, and Lorenzo Pareschi. Constrained consensus-based optimization. *SIAM Journal on Optimization*, 33(1):211–236, 2023b.

Leon Bungert, Tim Roith, and Philipp Wacker. Polarized consensus-based dynamics for optimization and sampling. *Mathematical Programming*, pages 1–31, 2024.

Gon Buzaglo, Niv Haim, Gilad Yehudai, Gal Vardi, Yakir Oz, Yaniv Nikankin, and Michal Irani. Deconstructing data reconstruction: Multiclass, weight decay and general losses. *Advances in Neural Information Processing Systems*, 2023.

José A Carrillo, Young-Pil Choi, Claudia Totzeck, and Oliver Tse. An analytical framework for consensus-based global optimization method. *Mathematical Models and Methods in Applied Sciences*, 28(06):1037–1066, 2018.

José A Carrillo, Shi Jin, Lei Li, and Yuhua Zhu. A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 27:S5, 2021.

José A Carrillo, Franca Hoffmann, Andrew M Stuart, and Urbain Vaes. Consensus-based sampling. *Studies in Applied Mathematics*, 148(3):1069–1140, 2022.

José A Carrillo, Claudia Totzeck, and Urbain Vaes. Consensus-based optimization and Ensemble Kalman Inversion for global optimization problems with constraints. 40, 2023.

Enis Chenchene, Hui Huang, and Jinniao Qiu. A consensus-based algorithm for non-convex multiplayer games. *arXiv preprint arXiv:2311.08270*, 2023.

Amir Dembo. *Large deviations techniques and applications*. Springer, 2009.

Richard Durrett. *Stochastic calculus: a practical introduction*. CRC press, 2018.

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Soc., 2010.

Massimo Fornasier and Lukang Sun. A pde framework of consensus-based optimization for objectives with multiple global minimizers. *arXiv preprint arXiv:2403.06662*, 2024.

Massimo Fornasier, Hui Huang, Lorenzo Pareschi, and Philippe Sünnen. Consensus-based optimization on the sphere: Convergence to global minimizers and machine learning. *The Journal of Machine Learning Research*, 22(1):10722–10776, 2021.

Massimo Fornasier, Timo Klock, and Konstantin Riedl. Convergence of anisotropic consensus-based optimization in mean-field law. In *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*, pages 738–754. Springer, 2022.

Massimo Fornasier, Timo Klock, and Konstantin Riedl. Consensus-based optimization methods converge globally, 2024a. URL `https://arxiv.org/abs/2103.15130`.

Massimo Fornasier, Peter Richtárik, Konstantin Riedl, and Lukang Sun. Consensus-based optimisation with truncated noise. *European Journal of Applied Mathematics*, page 1–24, 2024b. doi: 10.1017/S095679252400007X.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. *Advances in Neural Information Processing Systems*, 35: 22911–22924, 2022.

Hui Huang and Jinniao Qiu. On the mean-field limit for the consensus-based optimization. *Mathematical Methods in the Applied Sciences*, 45(12):7814–7831, August 2022. doi: 10.1002/mma.8279.

Hui Huang, Jinniao Qiu, and Konstantin Riedl. Consensus-based optimization for saddle point problems. *SIAM Journal on Control and Optimization*, 62(2):1093–1121, 2024.

Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in Neural Information Processing Systems*, 34:7232–7241, 2021.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Zhuohang Li, Jiaxin Zhang, Luyang Liu, and Jian Liu. Auditing privacy defenses in federated learning via generative gradient leakage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10132–10142, 2022.

Guodong Long, Ming Xie, Tao Shen, Tianyi Zhou, Xianzhi Wang, and Jing Jiang. Multi-center federated learning: clients clustering for better personalization. *World Wide Web*, 26(1):481–500, 2023.

Jie Ma, Guodong Long, Tianyi Zhou, Jing Jiang, and Chengqi Zhang. On the convergence of clustered federated learning. *arXiv preprint arXiv:2202.06187*, 2022.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Peter David Miller. *Applied asymptotic analysis*, volume 75. American Mathematical Soc., 2006.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

René Pinnau, Claudia Totzeck, Oliver Tse, and Stephan Martin. A consensus-based model for global optimization and its mean-field limit. *Mathematical Models and Methods in Applied Sciences*, 27 (01):183–204, 2017.

Konstantin Riedl. Leveraging memory effects and gradient information in consensus-based optimisation: On global convergence in mean-field law. *European Journal of Applied Mathematics*, page 1–32, 2023. doi: 10.1017/S0956792523000293.

Konstantin Riedl, Timo Klock, Carina Geldhauser, and Massimo Fornasier. Gradient is all you need? *arXiv preprint arXiv:2306.09778*, 2023.

Yichen Ruan and Carlee Joe-Wong. Fedsoft: Soft clustered federated learning with proximal local updating. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8124–8131, 2022.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

Claudia Totzeck. Trends in consensus-based optimization. In *Active Particles, Volume 3: Advances in Theory, Models, and Applications*, pages 201–226. Springer, 2021.

Yi Xu, Qihang Lin, and Tianbao Yang. Adaptive svrg methods under error bound conditions with unknown growth parameter. *Advances in Neural Information Processing Systems*, 30, 2017.

Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16337–16346, 2021.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M. Alvarez. Personalized federated learning with first order model optimization, 2021. URL `https://arxiv.org/abs/2012.08565`.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. iDLG: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

Ligeng Zhu and Song Han. Deep leakage from gradients. In *Federated learning*, pages 17–31. Springer, 2020.