Information and Inference: A Journal of the IMA (2023) **00**, 1–62
https://doi.org/DOI HERE

On adversarial robustness and the use of Wasserstein ascent-descent dynamics to enforce it

Camilo Andrés García Trillos
Department of Mathematics, University College London, Gower Street, WC1E 6BT, London, UK

and

Nicolás García Trillos*
Department of Statistics, University of Wisconsin Madison, 1300 University Avenue, 53706, Madison, WI, USA

We propose iterative algorithms to solve adversarial training problems in a variety of supervised learning settings of interest. Our algorithms, which can be interpreted as suitable ascent-descent dynamics in Wasserstein spaces, take the form of a system of interacting particles. These interacting particle dynamics are shown to converge toward appropriate mean-field limit equations in certain large number of particles regimes. In turn, we prove that, under certain regularity assumptions, these mean-field equations converge, in the large time limit, toward approximate Nash equilibria of the original adversarial learning problems. We present results for non-convex non-concave settings, as well as for non-convex concave ones. Numerical experiments illustrate our results.

Keywords: adversarial robustness, adversarial training, minmax games, Nash equilibrium, Wasserstein gradient flow, Wasserstein Fisher Rao metric, interacting particle system, mean-field limit

1. Introduction

In this paper, we propose and analyze ascent-descent dynamics to find approximate solutions (interpreted as Nash equilibria) to minmax problems of the form

$$\min_{\nu\in\mathcal{P}(\Theta)}\max_{\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z});\pi_z=\mu}\mathcal{U}(\pi,\nu), \tag{1.1}$$

where π_z is the first marginal of π and μ is a fixed probability measure. Our dynamics take the form of a system of finitely many interacting particles, which we will show converge, under suitable assumptions, toward a mean-field PDE as the number of particles in the system grows. We will also analyze the long-time behavior of the limiting mean-field dynamics and explore their ability to produce approximate Nash equilibria for (1.1). The studied dynamics are a version of gradient ascent-descent of the payoff function $\mathcal{U}$ under a convenient optimal transport geometric setting, and can be understood as analogous to dynamics studied in [12] and [17].

Through the paper we will think of Θ as the space of parameters of a learning model, e.g., a classifier or regression function; $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ is a space of input to output samples; $\mathcal{U}:\mathcal{P}(\mathcal{Z}^2)\times\mathcal{P}(\Theta)\to\mathbb{R}$ is a function representing a payoff defined over probability measures; and the inner maximization in (1.1) operates over *couplings where the first marginal is kept*

fixed and equal to the "clean" data distribution μ. As discussed in section 1.1 below and in Appendix A.1, (1.1) encompasses *distributionally robust optimization* (DRO) problems of the form

$$\min_{\nu\in\mathcal{P}(\Theta)}\max_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} R(\tilde{\mu},\nu)-C(\mu,\tilde{\mu}). \tag{1.2}$$

These are problems that in applications are used to enhance the robustness of learning models to adversarial perturbations of data.

In a nutshell, a DRO problem like the one above can be interpreted as a game played by a learner and an adversary: for the learner, the goal is to choose a distribution of learning parameters ν (implicitly inducing an input to output map) that is able to fare well when facing the attack of a reasonable adversary (reasonable as modeled by the cost function C) who can modify the distribution of clean data, here represented by the fixed probability measure μ; the functional R represents the *risk* of the classifier/regression function induced by ν relative to the data distribution $\tilde{\mu}$.

A brief discussion on adversarial training with pointers to relevant literature is presented in section 2.1.

Before we move on with the description of our algorithms and main theoretical results, it will be convenient to provide a concrete example of a payoff function $\mathcal{U}$ that is of interest in practical settings, in particular, in adversarial machine learning.

1.1. *Motivating example: robust supervised learning with shallow neural networks*

We examine a specific setting of (1.1) in which the variable ν can be directly related to a shallow, although possibly infinitely wide, neural network; see [18].

Let $\Theta\subseteq\mathbb{R}\times\mathbb{R}^{d'}$, $\mathcal{Z}=\mathbb{R}^{d'}\times\mathbb{R}$, and write $\theta=(a,b)$ and $z=(x,y)$. We consider the payoff function

$$\mathcal{U}(\pi,\nu):=\mathcal{R}(\pi,\nu)-\mathcal{C}(\pi), \tag{1.3}$$

with the following *risk* and *adversarial cost*:

$$\mathcal{R}(\pi,\nu):=\int_{\mathcal{Z}\times\mathcal{Z}}\ell(h_\nu(\tilde{x}),\tilde{y})d\pi(z,\tilde{z}),\quad h_\nu(x):=\int_\Theta af(b\cdot x)d\nu(a,b), \tag{1.4}$$

where $\ell:\mathbb{R}\times\mathbb{R}\to[0,\infty]$ is a loss function (e.g., squared-loss or logistics loss), f is an activation function (e.g., ReLu, sigmoid, or squared-ReLu), and

$$\mathcal{C}(\pi):=c_a\int_{\mathcal{Z}\times\mathcal{Z}}|z-\tilde{z}|^2d\pi(z,\tilde{z}), \tag{1.5}$$

for c_a a positive parameter. It is easy to verify that the case of an implementable finite neural network trained with a finite data-set is obtained by choosing discrete probability laws π and ν.

The square Euclidean distance case shown above is one of the many possible choices in (1.5). Notice that the risk function $\mathcal{R}$ only depends on π through its second marginal, $\pi_{\tilde{z}}$. This functional is thus the risk associated to the function h_ν when data points are assumed to be distributed according to $\pi_{\tilde{z}}$.

The parameter c_a in the cost $\mathcal{C}$ can be interpreted as reciprocal to an *adversarial budget* and determines the strength of adversarial attacks. In particular, if c_a is small, the attacker

can carry out stronger attacks, i.e., can propose new data points that are further away from clean data points $z \sim \mu$, while the opposite is true when c_a is large.

Remark 1 *As discussed in Appendix A.1, with the choices made above, problems* (1.1) *and* (1.2) *are equivalent if we set*

$$R(\tilde{\mu}, \nu) := \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}}[\ell(h_\nu(\tilde{x}), \tilde{y})], \quad C(\mu, \tilde{\mu}) := c_a W_2^2(\mu, \tilde{\mu}),$$

where $W_2(\mu, \tilde{\mu})$ is the standard 2-Wasserstein distance between μ and $\tilde{\mu}$. The resulting problem (1.2) *is a DRO version of adversarial training with an explicit penalty, as opposed to an explicit constraint; see [45]. One of the main outcomes of our work is precisely to propose an algorithm to solve this type of adversarial training problems. This intent is manifested in the marginal constraint we impose in the inner max in* (1.1).

Remark 2 *In adversarial training, in order to avoid enhancing robustness at the expense of a considerable loss in accuracy, it is important to tune the adversarial budget appropriately. Some papers that have studied the trade-off between robustness and accuracy include [47, 55].*

1.2. *Algorithm*

We introduce in Algorithm 1 a discrete in time particle-based scheme for solving the minmax problem (1.1).

Implicit in the definition of Algorithm 1 is the use of the *first variations* of the functional $\mathcal{U}$ in the directions ν and π.

Following Definition 7.12. in [43], we say that the measurable function $\mathcal{U}_\pi : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is the first variation of $\mathcal{U}$ in the direction π at the point (π, ν) if for any $\pi^* \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})$ we have

$$\frac{d}{d\epsilon}(\mathcal{U}(\pi + \epsilon(\pi^* - \pi), \nu))\Big|_{\epsilon=0} = \int_{\mathcal{Z}\times\mathcal{Z}} \mathcal{U}_\pi(\pi, \nu; z, \tilde{z}) d(\pi^* - \pi).$$

In general, $\mathcal{U}_\pi$ may depend on the point (π, ν) at which the first variation is being evaluated, but we will drop the explicit reference to this dependence whenever no confusion may arise from doing so, for otherwise we will write all of $\mathcal{U}_\pi$'s arguments like this: $\mathcal{U}_\pi(\pi, \nu; z, \tilde{z})$. Similarly, we say that the measurable function $\mathcal{U}_\nu : \Theta \to \mathbb{R}$ is the first variation of $\mathcal{U}$ in the direction ν at the point (π, ν) if for any $\nu^* \in \mathcal{P}(\Theta)$ we have

$$\frac{d}{d\epsilon}(\mathcal{U}(\pi, \nu + \epsilon(\nu^* - \nu)))\Big|_{\epsilon=0} = \int_{\Theta} \mathcal{U}_\nu(\pi, \nu; \theta) d(\nu^* - \nu).$$

Throughout the paper we will assume that the first variations of $\mathcal{U}$ are well-defined and satisfy regularity properties that are stated precisely in Assumptions 8.

In Algorithm 1 the term $\eta_t \Delta t$ can be interpreted as a time-dependent transport learning rate, and $\kappa \Delta t$ as a constant mass-transfer learning parameter. We expose explicitly the term Δt to facilitate all comparisons with the continuous-time dynamics below. The projection maps $P_{\mathcal{Z}}$ and P_Θ are introduced to ensure that iterations remain within the sets $\mathcal{Z}$ and Θ. The averaging steps in lines 18-19 will be discussed in section 6; Algorithm 1 is related to algorithms introduced in [17, 51], in turn related to [13]; a comparison between the content of these papers and ours is presented in section 2.1.

Algorithm 1 Wasserstein ascent-descent algorithm

Require: A collection $\{z_{i,0}, \omega_{i,0}\}_{i=1,\dots,n}$ such that $\frac{1}{n}\sum_{i=1}^n \omega_{i,0}\delta_{z_{i,0}}$ approximates μ.

Set $t = 0$

Choose $\{\vartheta_{k,0}\}_{k=1,\dots M}$, $\{\alpha_{k,0}\}_{k=1,\dots,M}$, $\{\tilde{z}_{ij,0}\}_{i=1,,\dots,n;\, j=1,\dots N}$, and $\{\omega_{ij,0}\}_{i=1,\dots,n;\ j=1,\dots N}$ with the constraint:

$$\sum_{j=1}^{N} \omega_{ij,0} = \omega_{i,0} \text{ for all } i = 1, \dots n.$$

while Stopping condition has not been satisfied **do**

Set

$$\pi_t^{n,N} := \sum_{i=1}^{n}\sum_{j=1}^{N} \omega_{ij,t}\delta_{(z_{i,0},\tilde{z}_{ij,t})} \qquad \nu_t^M := \sum_{k=1}^{M} \alpha_{k,t}\delta_{\vartheta_{k,t}}$$

for $i = 1$ to n ; $j = 1$ to N **do**

$\tilde{z}_{ij,t+1} = P_{\mathcal{Z}}\left(\tilde{z}_{ij,t} + (\eta_t \Delta t)\nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t, \nu_t; z_{i,0}\tilde{z}_{ij,t})\right)$

$\hat{\omega}_{ij,t+1} := \omega_{ij,t} \exp\left((\kappa \Delta t)\sum_{j'} \omega_{ij',t}\mathcal{U}_\pi(\pi_t, \nu_t; z_{i,0}, \tilde{z}_{ij,t})\right)$

$\omega_{ij,t+1} := \frac{\hat{\omega}_{ij,t+1}}{\sum_{j'} \hat{\omega}_{ij',t+1}}$

end for

for $k = 1$ to M **do**

$\vartheta_{k,t+1} = P_\Theta\left(\vartheta_{k,t} - (\eta_t \Delta t)\nabla_\theta \mathcal{U}_\nu(\pi_t, \nu_t; \vartheta_{k,t})\right)$

$\hat{\alpha}_{k,t+1} := \alpha_{k,t} \exp\left(-(\kappa \Delta t)\sum_{k'} \alpha_{k',t}\mathcal{U}_\nu(\pi_t, \nu_t; \vartheta_{k,t})\right)$

$\alpha_{k,t+1} := \frac{\hat{\alpha}_{k,t+1}}{\sum_{k'} \hat{\alpha}_{k',t+1}}$

end for

$t = t + 1$

end while

Calculate time-average

$\bar{z}_{ij} := \frac{1}{t}\sum_{s=0}^{t} \omega_{ij,s}\tilde{z}_{ij,s}$ for $i = 1, \dots, n; j = 1, \dots, N$

$\bar{\vartheta}_k := \frac{1}{t}\sum_{s=0}^{t} \alpha_{k,s}\tilde{\vartheta}_{k,s}$ for $k = 1, \dots, M$

2. Main theoretical results

We study the continuous-time version of the dynamics in Algorithm 1 and explore its ability to produce (approximate) Nash equilibria for the game (1.1). As in many works in the literature that study training dynamics of neural networks in overparameterized regimes (e.g., see [13, 52]) our analysis is split into two parts: 1) convergence of particle dynamics to a mean field in the large number of particles limit, and 2) analysis of the mean field equation in the long time horizon.

Following this general framework, in our first main result we describe the behaviour of the continuous-time version of Algorithm 1 (a system of coupled ODEs) as the number of particles grows. To be precise, the collection of iterates in Algorithm 1 can be thought of as a time

discretization of the system of ODEs:

$$\begin{aligned}
dZ_t^i &= 0\\
d\tilde{Z}_t^i &= \eta_t \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{Z}_t^i)dt\\
d\omega_t^i &= \kappa\omega_t^i \left(\mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{Z}_t^i) - \int \mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{z}')d\pi_t^N(\tilde{z}'|Z_t^i) \right) dt\\
d\vartheta_t^i &= -\eta_t \nabla_\theta \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \vartheta_t^i)dt\\
d\alpha_t^i &= -\kappa\alpha_t^i \left(\mathcal{U}_\nu(\pi_N^N, \nu_t^N; \vartheta_t^i) - \int \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \theta')d\nu_t^N(\theta') \right) dt,
\end{aligned} \tag{2.1}$$

with given initial condition $(Z_0^i, \tilde{Z}_0^i, \omega_0^i, \vartheta_0^i, \alpha_0^i)$ (possibly random) and

$$\pi_t^N := \frac{1}{N}\sum_{i=1}^N \omega_t^i \delta_{(Z_t^i, \tilde{Z}_t^i)}, \qquad \nu_t^N := \frac{1}{N}\sum_{i=1}^N \alpha_t^i \delta_{\vartheta_t^i}. \tag{2.2}$$

Here, as well as in our analysis in section 4, we have considered the same number of particles $Z, \tilde{Z}, \vartheta$ and we have eliminated the double indexes. This we do for simplicity and in order to reduce the burdensome notation throughout our analysis; we will only return to the double indexes when needed.

A simple computation reveals that the empirical measures (π_t^N, ν_t^N) in (2.2) satisfy the PDE (in weak form)

$$\begin{cases}
\partial_t \pi_t &= -\eta_t \mathrm{div}_{z,\tilde{z}}(\pi_t(0, \nabla_{\tilde{z}}\mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}))) + \kappa\pi_t \left(\mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}) - \int \mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}')d\pi_t(\tilde{z}'|z) \right)\\
\partial_t \nu_t &= \eta_t \mathrm{div}_\theta(\nu_t \nabla_\theta \mathcal{U}_\nu(\pi_t, \nu_t; \theta)) - \kappa\nu_t \left(\mathcal{U}_\nu(\pi_t, \nu_t; \theta) - \int \mathcal{U}_\nu(\pi_t, \nu_t; \theta')d\nu_t(\theta') \right),
\end{cases} \tag{2.3}$$

initialized at $\pi_0 = \pi_0^N$ and $\nu_0 = \nu_0^N$. In the above, $\pi_t(\cdot|z)$ must be interpreted as the conditional distribution of $\tilde{z}$ given z if the pair $(z, \tilde{z})$ is assumed to be distributed according to π_t. In **Theorem 12** we show that, under appropriate conditions, including a "well prepared initialization" assumption, the dynamics (2.3) converge to a mean-field system of non-local PDEs as $N \to \infty$. This mean field system is a solution to the exact same type of equation (2.3) except that initialized at different measures ν_0, π_0, formally, the limits of ν_0^N and π_0^N in suitable metrics. We will see that, in contrast to the consistency requirement for ν_0 in the standard 1-Wasserstein sense, the type of well-preparedness condition for the π_0^N variable is stronger and closely related to consistency in the Knothe-Rosenblatt optimal transport sense. The need for this stronger assumption is due to the presence of conditional distributions in the dynamics (2.3), which must be properly controlled with stronger metrics to carry out a propagation of chaos analysis. The proof of Theorem 12 thus requires a careful handling of new technical complications arising from the marginal constraint in (1.1) for the adversary. Our intermediate analysis will also help us establish the well-posedness of the system of PDEs (2.3) for arbitrary initializations, a result of interest in its own right.

In our second main result, **Theorem 35** (see also Theorem 36), we study the long-time behavior of the system of PDEs (2.3) when initialized appropriately. In particular, we prove

that, under suitable assumptions, the time-average of these dynamics eventually reaches an ε-Nash equilibrium of (1.1), a notion that we recall below.

Definition 3 (ε-Nash equilibrium) *Given $\varepsilon > 0$, we say that (π^*, ν^*) is an ε-Nash equilibrium for problem (1.1) if $\pi_z^* = \mu$ and*

$$\sup_{\pi \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z}) \ s.t. \ \pi_z = \mu} \{\mathcal{U}(\pi, \nu^*)\} - \inf_{\nu \in \mathcal{P}(\Theta)} \{\mathcal{U}(\pi^*, \nu)\} \leq \varepsilon. \tag{2.4}$$

Theorem 35 is proved under appropriate assumptions that include the concavity-convexity of $\mathcal{U}$ (in the linear interpolation sense) and a convenient, admittedly strong, assumption on the initializations of the variables π and ν. These strong assumptions, however, are analogous to ones considered in the study of training dynamics in overparameterized regimes in the non-adversarial setting. The strong assumption on the initialization of the variable π can be dropped under an additional Polyak-Lojasiewicz (PL) condition for $\mathcal{U}$ in the π variable (see Assumption 39). As discussed in Example 41, this assumption is not unreasonable in practical settings of interest, and in the scenario described in subsection 1.1 it is satisfied by assuming that the adversary has a sufficiently small budget (i.e., a sufficiently large c_a). Under this additional PL condition, **Theorem 42** (see also Theorem 43) states that it is possible to modify the dynamics in Algorithm 1 to create a gap in speed profiting from the additional concavity to obtain rapid convergence in the adversarial direction. Intuitively, in the modified dynamics one can quickly obtain good approximations for the inner maximization problems to obtain dynamics that resemble those of gradient descent for the outer minimization in (1.1). The effect of this analysis is that the requirements for convergence toward approximate Nash equilibria are relaxed.

2.1. *Literature review*

In this section we provide a brief literature review of the topic of adversarial robustness in supervised learning settings, focusing on some developments in recent years. Since the literature in this field has expanded very quickly and spans a variety of disciplines our review is necessarily non-exhaustive.

Many mathematical approaches that aim to enforce robustness in learning models can be categorized under the term "Distributionally Robust Optimization" (DRO), as formulated in (1.2). The DRO formulation has the advantage of clearly casting adversarial robustness in supervised learning as a minmax game. Several studies have explored adversarial training in the DRO framework, considering various learning models such as linear regression, neural networks, and other parametric settings [4, 5, 11, 29, 36, 45, 48]. Other works have focused on solving the problem by replacing the inner maximization associated with the adversary's actions with a regularized risk surrogate. For example, [19, 34, 37, 41, 42, 54] and [22] derived this surrogate when the adversary is restricted to positions within a distance ϵ from the training data, expanding the inner maximization objective around $\epsilon = 0$. A few recent studies have discussed adversarial robustness in the context of agnostic learners, where no modeling assumptions are made about the learner. This setting can be understood as a limiting case of a problem with a very expressive family of learning models and provides lower bounds for more general adversarial robustness problems. Some of these works include: [2, 3, 20, 39, 40].

Another approach is taken in [8, 9, 23, 24], where adversarial robustness in classification settings is linked to geometric variational problems.

There are other works in the literature that consider the computation of minmax problems over spaces of probability measures by using particle methods like we do in this paper. In particular, we would like to highlight our contributions in relation to two papers in this category that are closely related to ours [17, 51], both of which adapt ideas presented in [12] to minmax problems. In the work [17] and the very recent work [51] the authors consider minmax problems with a linear (with respect to the measures) payoff function. Our setting is broader as it covers not only non-linear objectives but also studies the effect on a coupling constrained by one of the components being pinned to an input function. This level of generality allows us to study broad cases of adversaries in the space of measures (DRO version of adversarial training). It is worth remarking that under the simpler setting in [51], the authors are able to show the exponential convergence (toward an actual Nash equilibrium), without assuming time separation of scales, of an algorithm with a similar geometric motivation than ours. In its practical implementation, both algorithms look very similar. The convergence in [51] is obtained under similar regularity hypotheses but assuming in addition that the (unique) solution is supported on a discrete set. Since we do not a priori assume the existence of a unique solution, our results are weaker in terms of convergence rate, as well as due to the fact that we can only recover approximate Nash equilibria. Other work of interest in the linear payoff setting, where a KL-regularization is introduced and then gradually turned off to deduce convergence of dynamics toward the Nash equilibrium of the original problem, is [33].

We emphasize that by considering the restriction $\pi_z = \mu$ in the minmax problem we can cover a wider variety of settings relevant to the study of adversarial machine learning than previous works in the literature. This gain in generality naturally comes at the expense of additional technical challenges. To point at some of these specific challenges, notice that when the payoff is non-linear its first variations are measure dependent, already suggesting the need for a more delicate analysis at the moment of proving the convergence of particle dynamics toward mean-field limits. The difficulties in our analysis are heightened by the presence of conditional distributions in the evolving systems. In order to handle these additional terms, we must recur to new ideas and constructions. In the end, the general mean-field analysis that we present can be also combined with lower-semicontinuity arguments to justify certain steps in the second part of the paper, i.e., the analysis of the long-time behavior of the mean-field system, providing in this way alternatives to approaches in the literature that may not be fully justified; we discuss this in section 5 below.

Moreover, we believe that some of the ancillary results we obtained to support the targeted level of generality of our model may be of interest in their own right.

We also highlight our study of the non-convex concave setting delineated in subsection 5.1. Indeed, we may exploit the additional strong concavity that is gained when considering adversaries with low budget to obtain stronger convergence results toward approximate Nash equilibria of the adversarial problem. Other papers that have explored this setting include [45], but the results presented there only guarantee, for the learner, convergence toward stationary points (although it is worth highlighting that they do not consider the mean-field regime).

In summary, our work is complementary to other papers such as [12, 17, 45, 51] (among others). Our results can be viewed as analogue to those in works such as [13, 52], which have studied the global convergence of (non-robust) training of shallow neural networks in the mean-field regime.

2.2. *Outline*

The remainder of the paper is organized as follows.

In section 3 we introduce required definitions and notation, and we briefly discuss the ascent-descent interpretation of Algorithm 1.

In section 4 we present the mean field analysis of the continuous-time version of our algorithm, i.e., the system of ODEs (2.1). That is, we state and prove our first main result, **Theorem 12**. We also discuss important auxiliary results that are used later in section 5.

In section 5 we discuss the long-time behavior of the mean-field system obtained in Theorem 12 and state conditions under which these mean-field dynamics produce approximate Nash equilibria for problem (1.1). In the first main result in section 5, **Theorem 35**, we assume strong conditions on the initialization of the mean field dynamics for both players. In **Theorem 42**, on the other hand, we drop the assumption on initialization for the variable π by imposing an additional PL condition on the payoff function $\mathcal{U}$ and by introducing a small modification to the dynamics discussed in the previous main result.

In section 6 we discuss some numerical results of an implementation of our algorithm when used in an actual machine learning task. Our main purpose with such an implementation is to illustrate that the algorithm is effective to obtain adversarially robust classifiers even away from the asymptotic regimes studied theoretically in the paper.

We wrap up the paper in section 7, where we present some conclusions and discuss future directions for research.

3. Preliminaries

Throughout this section we introduce some mathematical definitions and notation that we will use in the remainder. We will also briefly discuss the geometric motivation behind Algorithm 1.

In the sequel, we use the p-Wasserstein distance W_p (with $p \geq 1$) to compare probability distributions over a given metric space $(\mathbb{M}, d(\cdot,\cdot))$. The metric $d(\cdot,\cdot)$ that will be used in each instance will be specified in context. For example, in Assumption 8 the 1-Wasserstein distances considered are the ones relative to the Euclidean metric in each corresponding Euclidean space.

Definition 4 *Given two probability measures υ, υ' over $\mathbb{M}$, their p-Wasserstein distance $W_p(\upsilon, \upsilon')$ is defined according to*

$$W_p^p(\upsilon, \upsilon') := \inf_{\Upsilon \in \Gamma(\upsilon, \upsilon')} \int_{\mathbb{M}\times\mathbb{M}} (d(u,u'))^p d\Upsilon(u,u'),$$

where $\Gamma(\upsilon, \upsilon')$ is the set of couplings between υ and υ'. We will use $\mathcal{P}(\mathbb{M})$ to denote the space of Borel probability measures over $\mathbb{M}$.

3.1. *Gradient ascent-descent interpretation of Algorithm 1*

In this section we summarize the geometric motivation behind the system of equations (2.3) and its discretization in Algorithm 1. The interested reader can find a more detailed discussion in Appendix B, or consult several related references like [14, 21, 28, 32, 46]. In short, system (2.3) can be interpreted as the projection of a Wasserstein gradient flow in an appropriate lifted

space. It is also possible to interpret the resulting equations as gradient flows relative to a certain Wasserstein-Fisher-Rao metric over the original probability space (see our Remark 64). While both interpretations are valid, in the main text we avoid explicitly mentioning the WFR metric and stick with the Wasserstein interpretation given that several of our computations take place explicitly on the lifted spaces.

3.1.1. Lifted space

We introduce two projection maps between probability spaces that will play an important role in our derivations. We use the same name for both of them for convenience, and we expect no ambiguity given the context.

Let $\mathcal{M}_+(\Theta)$ (respectively $\mathcal{M}_+(\mathcal{Z}^2)$) denote the space of finite positive measures over Θ (respectively $\mathcal{Z}^2$). We consider the *projection map* from either $\mathcal{P}(\Theta \times \mathbb{R}_+)$ onto $\mathcal{M}_+(\Theta)$, or from $\mathcal{P}(\mathcal{Z}^2 \times \mathbb{R}_+)$ onto $\mathcal{M}_+(\mathcal{Z}^2)$, characterized by the respective identities

$$\int \varphi(\theta) d(\mathcal{F}\sigma)(\theta) = \int \alpha\varphi(\theta) d\sigma(\theta,\alpha); \qquad \int \varphi(z,\tilde{z}) d(\mathcal{F}\gamma)(z,\tilde{z}) = \int \alpha\varphi(z,\tilde{z}) d\gamma(z,\tilde{z},\alpha) \tag{3.1}$$

for all regular enough test functions φ from Θ or $\mathcal{Z}^2$ into $\mathbb{R}$. The map $\mathcal{F}$ allows us to lift an energy functional defined over $\mathcal{M}_+(\Theta)$ (or $\mathcal{M}_+(\mathcal{Z}^2)$): we can then consider gradient descent dynamics in the lifted space, and, in turn, these lifted dynamics can be projected down to the original space of measures to generate an evolution there.

Remark 5 *Notice that the function $\mathcal{F}$ is a surjection. Indeed, let $\nu \in \mathcal{M}_+(\Theta)$ and let $M = \nu(\Theta)$, which we first assume is non-zero. Consider the probability measure $\sigma = \frac{\mu}{M} \otimes \delta_M$. It is straightforward to show that $\mathcal{F}\sigma = \nu$. In case $M = 0$, which means ν is the measure that is identically equal to zero, we may take σ to be any probability measure over $\Theta \times [0,\infty)$ that satisfies $\sigma(\Theta \times \{0\}) = 1$ to conclude that $\mathcal{F}\sigma = \nu$. Clearly, the same argument holds for $\mathcal{F} : \mathcal{P}(\mathcal{Z}^2 \times \mathbb{R}_+) \to \mathcal{M}_+(\mathcal{Z}^2)$. Finally, while $\mathcal{F}$ is surjective, it is worth highlighting that it is far from being one to one.*

3.2. *Ascent-descent equations in the lifted space*

In an Euclidean space, where one has a target payoff function (say U) for which one wishes to find its saddles, one could consider a system of the form

$$\begin{cases} \dot{q}_t = -\nabla_q U(q_t, p_t) \\ \dot{p}_t = \nabla_p U(q_t, p_t), \end{cases}$$

or a projected version thereof in case additional constraints on the variables p, q are present. Analogous systems can be considered in more general Riemannian settings. In particular, by considering the Riemannian structure for the space $\mathcal{P}(\Theta \times [0,\infty))$ presented in Appendix B, one obtains the following gradient ascent-descent equations in the space of measures:

$$\begin{cases} \partial_t \gamma_t &= -\mathrm{div}_{(z,\tilde{z}),\omega}(\gamma_t v_\gamma(z,\tilde{z},\omega)), \\ \partial_t \sigma_t &= \mathrm{div}_{\theta,\alpha}(\sigma_t v_\sigma(\theta,\alpha)), \end{cases} \tag{3.2}$$

where

$$v_\gamma(z,\tilde{z},\omega) = \left(0, \eta_t \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t,\nu_t;z,\tilde{z}), \kappa\omega\left(\mathcal{U}_\pi(\pi_t,\nu_t;z,\tilde{z}) - \int \mathcal{U}_\pi(\pi_t,\nu_t;z,\tilde{z}') d\pi_t(\tilde{z}'|z)\right)\right),$$

$$v_\sigma(\theta,\alpha) = \left(\eta_t \nabla_\theta \mathcal{U}(\pi_t,\nu_t;\theta), \kappa\alpha(\mathcal{U}_\nu(\pi_t,\nu_t;\theta) - \int \mathcal{U}_\nu(\pi_t,\nu_t;\theta')d\nu_t(\theta'))\right),$$

and $\pi_t = \mathcal{F}\gamma_t, \quad \nu_t = \mathcal{F}\sigma_t$.

Notice that here we allow the scaling factor η to change in time. This change does not affect the above discussion but warrants us with additional flexibility that is used in the convergence analysis. Section 4 is devoted to studying equation (3.2). In particular, we prove well-posedness and show that system (3.2) can be recovered as a suitable limit of systems of interacting particles. Looking forward to applications in section 5, in section 4 we will actually study a slightly more general system than (3.2).

To finally return to the original system (2.3) it now suffices to project the dynamics (3.2) via the map $\mathcal{F}$ as stated in Proposition 6, whose proof is in appendix B.6.

Proposition 6 *Suppose that (γ,σ) solves the lifted dynamics (3.2). Then the pair $\pi_t = \mathcal{F}\gamma_t, \quad \nu_t = \mathcal{F}\sigma_t$ solves the system (2.3).*

The bottom line is that, by studying the system (3.2) and its approximation with particle systems, we will be implicitly studying the system (2.3) and its approximation with particle systems. System (3.2), however, has the advantage of having a direct Lagrangian interpretation that we exploit.

3.2.1. Conservation of mass

Let us now remark that the system (2.3) with arbitrary initialization satisfies certain conservation of mass properties.

Remark 7 *Note that the dynamics in (2.3) imply that the first marginal of π (π_z) remains constant. This can be verified by considering a test function $\phi : \mathcal{Z} \to \mathbb{R}$ and observing that*

$$\begin{aligned}
\frac{d}{dt}\int_{\mathcal{Z}\times\mathcal{Z}} \phi(z)d\pi_t(z,\tilde{z}) &= \eta_t \int_{\mathcal{Z}\times\mathcal{Z}} \nabla_{z,\tilde{z}}\phi(z)\cdot(0,\nabla_{\tilde{z}}\mathcal{U}_\pi)d\pi_t(z,\tilde{z}) \\
&\quad + \kappa\int_{\mathcal{Z}\times\mathcal{Z}} \phi(z)\left(\mathcal{U}_\pi(z,\tilde{z}) - \int \mathcal{U}_\pi(z,\tilde{z}')d\pi_t(\tilde{z}'|z)\right)d\pi_t(z,\tilde{z}) \\
&= \kappa\int_{\mathcal{Z}} \phi(z)\int_{\mathcal{Z}}\left(\mathcal{U}_\pi(z,\tilde{z}) - \int \mathcal{U}_\pi(z,\tilde{z}')d\pi_t(\tilde{z}'|x)\right)d\pi_t(\tilde{z}|z)d\pi_{t,z}(z) \\
&= 0.
\end{aligned}$$

Similarly, one can show that ν_t and π_t have a total mass equal to one for all times, provided ν_0,π_0 are probability measures.

3.3. *Notation*

In the sequel, we will use the following notation:

– $\mu,\tilde{\mu}$ probability measures over $\mathcal{Z}$. μ is the observed data distribution and $\tilde{\mu}$ represents an adversarial perturbation of μ.

- π is a measure over $\mathcal{Z}\times\mathcal{Z}$, and we write points in the support of π as $(z,\tilde{z})$. z can be interpreted as an observed data point, while $\tilde{z}$ corresponds to a perturbed data point.
- π_z will be used to denote the first marginal of π, whenever π is a probability measure. $\pi(\cdot|z)$ will be used to denote the conditional distribution of the second variable given that the first one is equal to z.
- γ will represent a probability measure over the lifted space $\mathcal{Z}^2\times\mathbb{R}_+$.
- ν will represent a measure over Θ.
- σ will denote measures over the lifted space $\Theta\times\mathbb{R}_+$.
- $\mathcal{F}$ is the projection map from either $\mathcal{P}(\Theta\times\mathbb{R}_+)$ onto $\mathcal{M}_+(\Theta)$, or from $\mathcal{P}(\mathcal{Z}^2\times\mathbb{R}_+)$ onto $\mathcal{M}_+(\mathcal{Z}^2)$.
- $\boldsymbol{\gamma}$ will denote a probability measure over the space $\mathcal{C}([0,T],\mathcal{Z}\times\mathbb{R}_+)$, and $\boldsymbol{\sigma}$ will be used to denote probability measures over the space $\mathcal{C}([0,T],\Theta\times\mathbb{R}_+)$. The space $\mathcal{C}([0,T],\Theta\times\mathbb{R}_+)$ is the space of continuous functions from the interval $[0,T]$ into $\Theta\times\mathbb{R}_+$ and $\mathcal{C}([0,T],\mathcal{Z}^2\times\mathbb{R}_+)$ is defined analogously. These spaces will be endowed with the metric of uniform convergence.
- We will use $\check{\gamma}$ to represent probability measures over the lifted space $\mathcal{Z}^2\times\mathbb{R}_+^2$ (notice the additional coordinate), and $\check{\sigma}$ will be used to represent probability measures over the lifted space $\Theta\times\mathbb{R}_+^2$.
- $\check{\boldsymbol{\gamma}}$ will denote a probability measure over the space $\mathcal{C}([0,T],\mathcal{Z}\times\mathbb{R}_+^2)$, and $\check{\boldsymbol{\sigma}}$ will be used to denote probability measures over the space $\mathcal{C}([0,T],\Theta\times\mathbb{R}_+^2)$.
- $\mathcal{U}(\pi,\nu)$ denotes the payoff associated to the measures π and ν, and $\mathcal{U}_\pi$ and $\mathcal{U}_\nu$ denote the first variations of $\mathcal{U}$ in the coordinates π and ν, respectively.
- We will use $\mathcal{H}(\cdot||\cdot)$ to denote the KL-divergence, or Shannon relative entropy, between two arbitrary probability measures defined over the same space. That is, given υ,υ' probability measures, $\mathcal{H}(\upsilon'||\upsilon)$ is defined as $\int\log(\frac{d\upsilon'}{d\upsilon})d\upsilon'$, if $\upsilon'\ll\upsilon$, and $+\infty$ otherwise.

4. From particle system to mean-field PDE

Our first result, which describes the large number of particles limit ($N\to\infty$) of the system (2.3) when initialized at π_0^N and ν_0^N, is deduced under the following assumptions on $\mathcal{U}$ and its first variations.

Assumption 8 *We assume that there exist constants $M,L>0$ such that*

- *$\mathcal{U}$ is bounded and Lipschitz with respect to the 1-Wasserstein distance. That is,*

$$|\mathcal{U}(\pi,\nu)|\le M;\qquad \mathcal{U}(\pi^1,\nu^1)-\mathcal{U}(\pi^2,\nu^2)\le L(W_1(\pi^1,\pi^2)+W_1(\nu^1,\nu^2)).$$

- *The first variations of $\mathcal{U}$ are bounded and Lipschitz, i.e.,*

$$\begin{aligned}
&|\mathcal{U}_\pi(\pi,\nu;z,\tilde{z})|+|\mathcal{U}_\nu(\pi,\nu;\theta)|\le M\\
&|\mathcal{U}_\pi(\pi^1,\nu^1;z^1,\tilde{z}^1)-\mathcal{U}_\pi(\pi^2,\nu^2;z^2,\tilde{z}^2)|\le L(W_1(\pi^1,\pi^2)+W_1(\nu^1,\nu^2)+|z^1-z^2|+|\tilde{z}^1-\tilde{z}^2|)\\
&|\mathcal{U}_\nu(\pi^1,\nu^1;\theta^1)-\mathcal{U}_\nu(\pi^2,\nu^2;\theta^2)|\le L(W_1(\pi^1,\pi^2)+W_1(\nu^1,\nu^2)+|\theta^1-\theta^2|).
\end{aligned}$$

- *The gradients of the first variations of $\mathcal{U}$ are bounded and Lipschitz, i.e.,*

$$|\nabla_{\tilde{z}}\mathcal{U}_\pi(\pi,\nu;z,\tilde{z})| + |\nabla_\theta \mathcal{U}_\nu(\pi,\nu;\theta)| \leq M$$
$$|\nabla_{\tilde{z}}\mathcal{U}_\pi(\pi^1,\nu^1;z^1,\tilde{z}^1) - \nabla_{\tilde{z}}\mathcal{U}_\pi(\pi^2,\nu^2;z^2,\tilde{z}^2)| \leq L(W_1(\pi^1,\pi^2)+W_1(\nu^1,\nu^2)+|z^1-z^2|+|\tilde{z}^1-\tilde{z}^2|)$$
$$|\nabla_\theta \mathcal{U}_\nu(\pi^1,\nu^1;\theta^1) - \nabla_\theta \mathcal{U}_\nu(\pi^2,\nu^2;\theta^2)| \leq L(W_1(\pi^1,\pi^2)+W_1(\nu^1,\nu^2)+|\theta^1-\theta^2|).$$

In the above, $\pi,\pi^i \in \mathcal{P}(\mathcal{Z}^2)$, $\nu,\nu^i \in \mathcal{P}(\Theta)$, $(z^i,\tilde{z}^i) \in \mathcal{Z}^2$, and $\theta^i \in \Theta$. The sets Θ and $\mathcal{Z}^2$ are compact subsets of the Euclidean spaces $\mathbb{R}^d$ and $\mathbb{R}^{2d'}$, respectively. We assume that these sets have Lipschitz boundaries.

Since the sets Θ and $\mathcal{Z}^2$ have been assumed to be bounded, in order to simplify the writing of our proofs and guarantee that all the dynamics to be studied in the paper stay within the domains Θ and $\mathcal{Z}^2$ we make the following technical assumption:

Assumption 9 *At all points $\tilde{z}$ at the boundary of $\mathcal{Z}$ and at all points θ at the boundary of Θ, it holds that the vector $\nabla_{\tilde{z}}\mathcal{U}_\pi(\pi,\nu;z,\tilde{z})$ points toward the interior of $\mathcal{Z}$, regardless of π,ν,z; and the vector $\nabla_\theta \mathcal{U}_\nu(\pi,\nu;\theta)$ points toward the interior of Θ, regardless of π,ν.*

By restricting our attention to compact sets $\Theta,\mathcal{Z}^2$ we make it simpler to verify the boundedness and Lipschitz conditions in Assumption 8 as these conditions reduce to weaker properties like local-Lipschitzness. Notice that in many applications there are natural bounds on the supports of the desired solution[1]. Assumption 9, on the other hand, guarantees that all dynamics considered in the paper remain in the domains Θ and $\mathcal{Z}^2$ (e.g., the ODE dynamics (4.1)) below. For Assumption 9 to make sense one requires the Lipschitz assumption on the boundary of $\Theta,\mathcal{Z}^2$ (indeed, the reader is invited to consider the case of Θ or $\mathcal{Z}^2$ being a Cantor set). Now, in order to satisfy the constraint imposed by this assumption, we can work with a modified functional $\mathcal{U}$ that strongly penalizes leaving the domains as we move closer to their borders. In particular, to a given $\mathcal{U}$ satisfying Assumptions 8 we can add, if needed, an exogenous term of the form $\int \varphi_2(\theta)d\nu(\theta) - \int \varphi_1(\tilde{z})d\pi_{\tilde{z}}$, where the φ_1,φ_2 are confining potentials: they are zero away from the boundary of the domains and grow as one approaches the boundaries. We reiterate that this assumption is made to simplify the writing of our proofs by sparing us from introducing additional terms like projection operators. We emphasize that Assumption 9 does not have an effect on the convexity properties (in linear or Wasserstein sense) of our loss function and the addition of confining potentials as described does not play any role in our analysis. Throughout the entire paper we adopt Assumption 9, even if not mentioned explicitly.

Example 10 *In the context of the motivating example in subsection 1.1 we see that when Θ and $\mathcal{Z}$ are bounded balls with respect to the ℓ_p, $p \geq 1$, norm the required conditions on the spaces would be satisfied, and so all conditions in Assumption 8 are satisfied when one considers a loss function that is twice differentiable and an activation function whose first derivative is Lipschitz. This is the case, for example, for the squared-loss and the squared ReLu or sigmoid activations.*

[1] To give only one example, images are typically represented by pixels which have a lower and upper values

Let us now introduce an enlarged system of ODEs closely related to the system (2.1). For $i = 1, \dots, N$, let

$$\begin{aligned}
dZ_t^i &= 0 \\
d\tilde{Z}_t^i &= \eta_t \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{Z}_t^i) dt \\
d\omega_t^i &= \kappa \omega_t^i \left(\mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{Z}_t^i) - \int \mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{z}') d\pi_t^N(\tilde{z}' | Z_t^i) \right) dt \\
d\vartheta_t^i &= -\eta_t \nabla_\theta \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \vartheta_t^i) dt \\
d\alpha_t^i &= -\kappa \alpha_t^i \left(\mathcal{U}_\nu(\pi_N^N, \nu_t^N; \vartheta_t^i) - \int \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \theta') d\nu_t^N(\theta') \right) dt \\
d\beta_t^i &= 0 \\
d\varrho_t^i &= 0;
\end{aligned} \tag{4.1}$$

with given initial condition $(Z_0^i, \tilde{Z}_0^i, \omega_0^i, \vartheta_0^i, \alpha_0^i, \beta_0^i, \varrho_0^i)$ (possibly random) and

$$\begin{aligned}
\check{\gamma}_t^N &:= \frac{1}{N} \sum_{i=1}^N \delta_{(Z_t^i, \tilde{Z}_t^i), \omega_t^i, \beta_t^i}, \quad \gamma_t^N := \frac{1}{N} \sum_{i=1}^N \delta_{(Z_t^i, \tilde{Z}_t^i), \omega_t^i}, \quad \pi_t^N := \mathcal{F}[\gamma_t^N] = \frac{1}{N} \sum_{i=1}^N \omega_t^i \delta_{(Z_t^i, \tilde{Z}_t^i)}, \\
\check{\sigma}_t^N &:= \frac{1}{N} \sum_{i=1}^N \delta_{\vartheta_t^i, \omega_t^i, \varrho_t^i}, \quad \sigma_t^N := \frac{1}{N} \sum_{i=1}^N \delta_{\vartheta_t^i, \alpha_t^i}, \quad \nu_t^N := \mathcal{F}[\sigma_t^N] = \frac{1}{N} \sum_{i=1}^N \alpha_t^i \delta_{\vartheta_t^i}.
\end{aligned} \tag{4.2}$$

The new variables β and ϱ have been added to the system for convenience: in particular, the extra degrees of freedom that come from the different ways to initialize these variables will come in useful in the second half of section 4.3. However, as can be seen from (4.1), these variables do not affect the evolution of the remaining variables, which follow the dynamics (2.1).

Before stating the main result of this section, it is worth introducing one last definition that we use to characterize the type of consistency requirement for the initialization in the particle system in the $N \to \infty$ limit.

Definition 11 *Given two probability measures γ, γ' over $\mathcal{Z} \times \mathcal{Z} \times [0, \infty)$, we define*

$$W_1^{KR}(\gamma, \gamma') := \inf_{\upsilon \in \Gamma_{Opt}(\mathcal{F}[\gamma]_z, F[\gamma']_z)} \int W_1(\gamma(\cdot|z), \gamma'(\cdot|z')) d\upsilon(z, z').$$

In the above, $\mathcal{F}[\gamma]_z$ is the first marginal of $\mathcal{F}[\gamma]$ (we recall $\mathcal{F}$ was introduced in (3.1)) and $\mathcal{F}[\gamma']_z$ is interpreted analogously; $\Gamma_{Opt}(\mathcal{F}[\gamma]_z, F[\gamma']_z)$ stands for the set of optimal couplings between $\mathcal{F}[\gamma]_z$ and $\mathcal{F}[\gamma']_z$ that realize the 1-Wasserstein distance between $\mathcal{F}[\gamma]_z$ and $\mathcal{F}[\gamma']_z$; finally, $\gamma(\cdot|z)$ (likewise for $\gamma'(\cdot|z)$) is the conditional of the second and third variables given the first one has been fixed.

The above construction is related to the notion of Knothe-Rosenblatt rearrangement (see chapter 2.3 in [43] and also [6]), and to the notion of fibered optimal transport introduced in [38].

We are ready to state our first main result precisely.

Theorem 12 (Convergence particle system) *Let $T>0$, and suppose that Assumptions 8 and 9 hold. Let $\bar{\pi}_0,\bar{\nu}_0$ be probability measures with $\bar{\pi}_{0,z}=\mu$ and suppose that γ_0 and σ_0 are probability measures satisfying $\mathcal{F}\gamma_0=\bar{\pi}_0$ and $\mathcal{F}\sigma_0=\bar{\nu}_0$, where $\mathcal{F}$ is defined in (3.1).*

Let γ_t^N,σ_t^N

$$\gamma_t^N := \frac{1}{N}\sum_{i=1}^N \delta_{(Z_t^i,\tilde{Z}_t^i),\omega_t^i}, \qquad \sigma_t^N := \frac{1}{N}\sum_{i=1}^N \delta_{\vartheta_t^i,\alpha_t^i},$$

for initial values ω_0^i,α_0^i bounded from above by a constant D (uniformly over N) and Z_0^i in the support of μ, and evolutions as in (4.1).

Finally, suppose that, as $N\to\infty$,

$$W_1^{KR}(\gamma_0^N,\gamma_0)\to 0, \text{ and } W_1(\sigma_0^N,\sigma_0)\to 0, \tag{4.3}$$

where W_1^{KR} was introduced in Definition 11.

Then, as $N\to\infty$,

$$\sup_{t\in[0,T]}\{W_1(\pi_t^N,\pi_t)+W_1(\nu_t^N,\nu_t)\}\to 0,$$

where π_t,ν_t solve (2.3) with initializations $\bar{\pi}_0$ and $\bar{\nu}_0$.

In simple terms, the above theorem states that our particle dynamics are consistent when their initializations are consistent in a suitable sense. This theorem is a consequence of a propagation of chaos result that we will develop gradually. Indeed, the structure of the dynamics in (2.3) involving a conditional contractive term escapes the scope of established results in mean-field analysis with deterministic trajectories (like Dobrushin's analysis, see [15]). Our own analysis revisits and goes beyond the underlying argument behind these known results. In particular, our propagation of chaos result imposes stronger initialization assumptions, ultimately reflected in the

stronger consistency guarantee required for the initialization of the variable γ (and thus also π) in Theorem 12.

Remark 13 (Constructing approximate initializations in Theorem 12) *Fix $\overline{\pi}_0$ and $\overline{\nu}_0$ and define $\gamma_0=\overline{\pi}_0\otimes\delta_1$. That is, γ_0 is the product of $\overline{\pi}_0$ with a Dirac delta at 1. Likewise, let $\sigma_0=\overline{\nu}_0\otimes\delta_1$. Evidently, $\mathcal{F}\gamma_0=\overline{\pi}_0,\mathcal{F}\sigma_0=\overline{\nu}_0$.*

We use randomization to construct approximate initializations satisfying the assumptions in Theorem 12. Let $\xi_1,\dots,\xi_n,\dots$ be a sequence of i.i.d. samples from $\overline{\pi}_{0,z}$, and for each $i\in\mathbb{N}$ let $\tilde{Z}_{i1},\dots,\tilde{Z}_{im},\dots$, be i.i.d. samples from $\overline{\pi}_0(\cdot|\xi_i)$. Let $\theta_1,\dots,\theta_n,\dots$ be i.i.d. samples from $\overline{\nu}_0$.

For fixed n,m and $i\le n$ and $j\le m$, set $\omega_{ij}=\alpha_{ij}=1$, $Z_{ij}=\xi_i$, and $\vartheta_{ij}=\theta_i$. Consider the measures

$$\pi_0^{n,m} := \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \delta_{(Z_{ij},\tilde{Z}_{ij})}, \quad \gamma_0^{n,m} := \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \delta_{(Z_{ij},\tilde{Z}_{ij},\omega_{ij})}$$

and

$$\nu_0^{n,m} := \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \delta_{\vartheta_{ij}}, \quad \sigma_0^{n,m} := \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \delta_{(\vartheta_{ij},\alpha_{ij})}.$$

Evidently, $\mathcal{F}\gamma_0^{n,m} = \pi_0^{n,m}$ *and* $\mathcal{F}\sigma_0^{n,m} = \nu_0^{n,m}$*, and the* Z_{ij} *can be assumed to belong to the support of* $\pi_{0,z}$*. It is also clear that the measure* $\gamma_0^{n,m}$ *has support in* $\mathcal{Z}^2 \times [0,1]$ *and* $\sigma_0^{n,m}$ *has support in* $\Theta \times [0,1]$*.*

By Lemma 59 in Appendix A.3, we can conclude that there exists a sequence $\{(n_k, m_k)\}_{k\in\mathbb{N}}$ *such that, as* $k \to \infty$*, the measures* $\sigma_0^{N_k} := \sigma_0^{n_k,m_k}$ *and* $\gamma_0^{N_k} := \gamma_0^{n_k,m_k}$ *satisfy (4.3) with probability one.*

Remark 14 *We highlight that in order to satisfy the first condition in (4.3) we need to consider the iterative sampling for the variables* $Z_{ij}, \tilde{Z}_{ij}$ *illustrated in Remark 13, while in general i.i.d. sampling from* $\overline{\pi}_0$ *does not provide a valid initialization for the particle system. This is because the first condition in (4.3) is a stronger condition than simply requiring* $W_1(\gamma_0^N, \gamma_0) \to 0$*; see Remark 60 in Appendix A.*

Finally, we highlight that the assumption on the conditional distributions at initialization imposed in (4.3) is used to control the conditional distributions of π *as the systems evolve in time.*

As mentioned earlier, Theorem 12 relies on the fact that, in the large N limit, system (4.1) is expected to behave like a system where the interactions in (4.1) have been replaced by *mean-field* dynamics. Such a system reads as follows. For $i = 1, \dots, N$, let

$$
\begin{aligned}
dZ_t^{mf,i} &= 0 \\
d\tilde{Z}_t^{mf,i} &= \eta_t \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t^{mf}, \nu_t^{mf}; Z_t^{mf,i}, \tilde{Z}_t^{mf,i}) dt \\
d\omega_t^{mf,i} &= \kappa \omega_t^{mf,i} \left(\mathcal{U}_\pi(\pi_t^{mf}, \nu_t^{mf}; Z_t^{mf,i}, \tilde{Z}_t^{mf,i}) - \int \mathcal{U}_\pi(\pi_t^{mf}, \nu_t^{mf}; Z_t^{mf,i}, \tilde{z}') d\pi_t^{mf}(\tilde{z}' | Z_t^{mf,i}) \right) dt \\
d\vartheta_t^{mf,i} &= -\eta_t \nabla_\theta \mathcal{U}_\nu(\pi_t^{mf}, \nu_t^{mf}; \vartheta_t^{mf,i}) dt \\
d\alpha_t^{mf,i} &= -\kappa \alpha_t^{mf,i} \left(\mathcal{U}_\nu(\pi_N^{mf}, \nu_t^{mf}; \vartheta_t^{mf,i}) - \int \mathcal{U}_\nu(\pi_t^{mf}, \nu_t^{mf}; \theta') d\nu_t^{mf}(\theta') \right) dt \\
d\beta^{mf,i} &= 0 \\
d\varrho^{mf,i} &= 0;
\end{aligned}
\tag{4.4}
$$

with the same initial conditions as in (4.1), and where $\pi_t^{mf} = \mathcal{F}(\boldsymbol{\gamma}_t), \nu_t^{mf} = \mathcal{F}(\boldsymbol{\sigma}_t)$ and $(\boldsymbol{\gamma}, \boldsymbol{\sigma})$ solves (3.2) with initial condition γ_0, σ_0 (in section 4.1 we prove the well-posedness of this equation); we recall that the map $\mathcal{F}$ has been introduced in Section 3.1. The fundamental difference between the mean-field system (4.4) and the original particle system (4.1) is that the measures determining the dynamics in (4.4) can be treated as fixed and independent of the evolving particles, while in system (4.1) there is an explicit dependence of the driving dynamics on the empirical measures π^N, ν^N associated to the underlying evolving particles.

In order to deduce a *propagation of chaos* result for the system (4.1), we need to show that the mean-field system (4.4) is well-defined and that we can control how far the evolutions (4.1) and (4.4) are from each other; we show this under Assumptions 8. To this aim, it is convenient to introduce some extra mathematical structure that will allow us to use standard analytical arguments: we will work on spaces of measures over continuous paths on $\mathcal{Z}^2 \times \mathbb{R}_+^2$ and $\Theta \times \mathbb{R}_+^2$

and eventually use a fixed point argument to establish well-posedness of a mean-field equation. We start by introducing a family of particle evolutions that will play an important role in our analysis.

Let us fix $T>0$, and let $\mathcal{A}_T$ be the set of pairs $(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}) \in \mathcal{P}(\mathcal{C}([0,T],\mathcal{Z}^2\times\mathbb{R}^2_+))\times\mathcal{P}(\mathcal{C}([0,T],\Theta\times\mathbb{R}^2_+))$ such that:

1. $\mathcal{F}\boldsymbol{\sigma}_t$ and $\mathcal{F}\boldsymbol{\gamma}_t$ are probability measures for all $t\in[0,T]$.
2. $\mathcal{F}[\boldsymbol{\gamma}_t](\cdot\times\mathcal{Z}) = \mathcal{F}[\boldsymbol{\gamma}_0](\cdot\times\mathcal{Z}) \quad \forall t\in[0,T]$.

Here, as well as in the remainder, for a given $\check{\boldsymbol{\gamma}}$ we denote by $\boldsymbol{\gamma}$ the pushforward of $\check{\boldsymbol{\gamma}}$ by the map $\{(z_t,\tilde{z}_t,\omega_t,\varrho_t)\}\mapsto\{(z_t,\tilde{z}_t,\omega_t)\}$, and abusing notation slightly, in the remainder we may use $\mathcal{F}\check{\boldsymbol{\gamma}}_t$ and $\mathcal{F}\boldsymbol{\gamma}_t$ indistinctly; we can analogously relate $\check{\boldsymbol{\sigma}}$ and $\boldsymbol{\sigma}$.

Associated to $(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}})\in\mathcal{A}_T$, we consider the multidimensional ODE:

$$
\begin{aligned}
dZ_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}} &= 0\\
d\tilde{Z}_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}} &= \eta_t\nabla_{\tilde{z}}\mathcal{U}_\pi(\pi_t,\nu_t;Z_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\tilde{Z}_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}})dt\\
d\omega_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}} &= \kappa\omega_t\left(\mathcal{U}_\pi(\pi_t,\nu_t;Z_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\tilde{Z}_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}) - \int\mathcal{U}_\pi(\pi_t,\nu_t;Z_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\tilde{z}')d\pi_t(\tilde{z}'|Z_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}})\right)dt\\
d\vartheta_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}} &= -\eta_t\nabla_\theta\mathcal{U}_\nu(\pi_t,\nu_t;\vartheta_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}})dt\\
d\alpha_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}} &= -\kappa\alpha_t\left(\mathcal{U}_\nu(\pi_t,\nu_t;\vartheta_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}) - \int\mathcal{U}_\nu(\pi_t,\nu_t;\theta')d\nu_t(\theta')\right)dt\\
d\beta_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}} &= 0\\
d\varrho_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}} &= 0\\
\pi_t = \mathcal{F}(\boldsymbol{\gamma}_t),&\qquad \nu_t = \mathcal{F}(\boldsymbol{\sigma}_t),
\end{aligned}
\tag{4.5}
$$

with initial conditions

$$
((Z_0^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\tilde{Z}_0^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}),\omega_0^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\varrho_0^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}) = ((\xi,\tilde{\xi}),\omega_0,\varrho_0)\sim\check{\boldsymbol{\gamma}}_0,\qquad (\vartheta_0^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\alpha_0^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\beta_0^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}) = (\vartheta,\alpha_0,\beta_0)\sim\check{\boldsymbol{\sigma}}_0. \tag{4.6}
$$

Note that with the condition $(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}})\in\mathcal{A}_T$ we can make sense of the term $\pi_t(\cdot|Z_t^{\check{\gamma},\check{\sigma}})$ in equation (4.5). Indeed, let us denote by $\pi_{t,z}$ the marginal on the z coordinate of π_t. By assumption, $\pi_{t,z}=\pi_{0,z}$, while $Z_t^{\check{\gamma},\check{\sigma}} = Z_0^{\check{\gamma},\check{\sigma}}$ can be assumed to be in the support of $\pi_{0,z}$ without the loss of generality. The conditional distribution $\pi_t(\cdot|Z_t^{\check{\gamma},\check{\sigma}})$ is thus well-defined thanks to the disintegration theorem.

Equation (4.5) is a multidimensional classical ODE describing an isolated particle following dynamics driven by an exogenous measure. A key observation is that, under Assumptions 8, equation (4.5) is driven by Lipschitz coefficients and so is well-posed by Caratheodory's existence theorem (see Theorem 5.3 in [27]). Assumption 8 and Gronwall's inequality further imply a bound on $\omega^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}$ and $\alpha^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}$. We summarize these observations in the next proposition for easy reference.

Proposition 15 *Under Assumption 8, there exists a unique solution to (4.5) for any fixed initialization. Moreover, we have*

$$\omega_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}} \in [0,\omega_0 e^{2\kappa M t}], \quad \alpha_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}} \in [0,\alpha_0 e^{2\kappa M t}]; \qquad \forall T \geq t > 0.$$

For a given $T > 0$, let us now consider the map:

$$\Psi_T : \mathcal{A}_T \mapsto \mathcal{P}(\mathcal{C}([0,T], \mathcal{Z}^2 \times \mathbb{R}_+^2)) \times \mathcal{P}(\mathcal{C}([0,T], \Theta \times \mathbb{R}_+^2))$$

defined by

$$\Psi_T(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}) = (\Psi_T^1(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}), \Psi_T^2(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}})) := (\mathrm{Law}[(Z^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}, \tilde{Z}^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}), \omega^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}, \varrho^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}], \mathrm{Law}[\vartheta^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}, \alpha^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}, \beta^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}]),$$

i.e., Ψ_T maps paths in the space of measures in the lifted space to itself. Moreover, Ψ_T maps $\mathcal{A}_T$ into itself, as we state in the next lemma.

Lemma 16 *Under Assumption 8, it follows*

$$\Psi_T(\mathcal{A}_T) \subseteq \mathcal{A}_T.$$

Moreover, for every $(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}) \in \mathcal{A}_T$ we have

$$\mathcal{F}[(\Psi_T^1(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}))_t](\cdot \times \mathcal{Z}) = \mathcal{F}[\check{\boldsymbol{\gamma}}_t](\cdot \times \mathcal{Z}).$$

Proof This result is immediate from Remark 7 and the fact that $dZ_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}} = 0$. □

For technical reasons, it will be convenient to introduce a version of the set $\mathcal{A}_T$ whose elements have supports satisfying a certain boundedness condition. Precisely, for a given $T > 0$ and $D > 0$, we let $\mathcal{A}_{T,D}$ be the set

$$\mathcal{A}_{T,D} := \{(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}) \in \mathcal{A}_T \text{ s.t. } \check{\boldsymbol{\gamma}}_t(\mathcal{Z}^2 \times [0, De^{2\kappa M t}] \times [0,D]) = 1, \quad \check{\boldsymbol{\sigma}}_t(\Theta \times [0, De^{2\kappa M t}] \times [0,D]) = 1, \quad \forall t \in [0,T]\}.$$

In particular, for $(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}) \in \mathcal{A}_{T,D}$, the weights (ω_0, ϱ_0) and (α_0, β_0) obtained as in (4.6) can be assumed to belong to $[0,D]^2$. Combining with Proposition 15, we can deduce that for $(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}) \in \mathcal{A}_{T,D}$ the weights $\omega_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}, \alpha_t^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}}$ in the dynamics (4.5) can be bounded above by $De^{2\kappa M t}$. We summarize this in the following lemma.

Lemma 17 *For every $T, D > 0$ we have $\Psi_T(\mathcal{A}_{T,D}) \subseteq \mathcal{A}_{T,D}$.*

Under Assumptions 8, we prove that we can control the distance between the image Ψ_T of two pairs $(\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\sigma}}^1)$ and $(\check{\boldsymbol{\gamma}}^2,\check{\boldsymbol{\sigma}}^2)$ in $\mathcal{A}_{T,D}$ with their own distance. Given $p \geq 1$, we use

$$W_{t,p}^p(\upsilon,\upsilon') := \inf_{\Upsilon \in \Gamma(\upsilon,\upsilon')} \int \sup_{s \in [0,t]} |u_s - v_s|^p d\Upsilon(x,y), \tag{4.7}$$

to compare two probability measures over the same path space. In particular, we will use these distances to compare measures over paths in any of the lifted spaces.

First, we show a continuity property for the map $\mathcal{F}$ when considering a restriction of its domain.

Lemma 18 *Let σ, σ' be two probability measures over $\Theta \times [0, D]$, where D is a fixed constant, and suppose that $\mathcal{F}\sigma$ and $\mathcal{F}\sigma'$ are also probability measures.*

Then

$$W_p^p(\mathcal{F}(\sigma), \mathcal{F}(\sigma')) \leq C_{\Theta,p,D} W_p(\sigma, \sigma'),$$

where the constant $C_{\Theta,p,D}$ can be written as $C_{\Theta,p,D} = \mathrm{diam}(\Theta)^{p-1}(\mathrm{diam}(\Theta) + D)$.

In particular, when its domain has been restricted, the map $\mathcal{F}$ is Lipschitz in the 1-Wasserstein sense.

Proof Let us start by noticing that the measures σ for which $\mathcal{F}\sigma$ is a probability measure are precisely the measures satisfying $\int \alpha d\sigma(\theta, \alpha) = 1$.

We first prove the result for $p = 1$.

Assume that σ and σ' take the form $\sigma = \sigma_n = \frac{1}{n}\sum_{i=1}^n \delta_{(\theta_i, \alpha_i)}$ and $\sigma' = \sigma'_n = \frac{1}{n}\sum_{i=1}^n \delta_{(\theta'_i, \alpha'_i)}$. It is well known that in that case there exists a permutation $T : \{1, \dots, n\} \mapsto \{1, \dots, n\}$ such that $W_1(\sigma_n, \sigma'_n) = \frac{1}{n}\sum_{i=1}^n |(\theta_i, \alpha_i) - (\theta'_{T(i)}, \alpha'_{T(i)})|$. Now, we can write the measures $\mathcal{F}\sigma_n$ and $\mathcal{F}\sigma'_n$ as

$$\mathcal{F}\sigma_n = \frac{1}{n}\sum_{i=1}^n \min\{\alpha_i, \alpha'_{T(i)}\}\delta_{\theta_i} + \frac{1}{n}\sum_{i=1}^n (\alpha_i - \min\{\alpha_i, \alpha'_{T(i)}\})\delta_{\theta_i}$$

and

$$\mathcal{F}\sigma'_n = \frac{1}{n}\sum_{i=1}^n \min\{\alpha_i, \alpha'_{T(i)}\}\delta_{\theta'_i} + \frac{1}{n}\sum_{i=1}^n (\alpha'_{T(i)} - \min\{\alpha_i, \alpha'_{T(i)}\})\delta_{\theta'_i}.$$

Notice that the mass from $\frac{1}{n}\sum_{i=1}^n \min\{\alpha_i, \alpha'_{T(i)}\}\delta_{\theta_i}$ can be used to cover for the mass demanded in $\frac{1}{n}\sum_{i=1}^n \min\{\alpha_i, \alpha'_{T(i)}\}\delta_{\theta'_i}$. We carry out the following mass transfer: for each i, we send $\min\{\alpha_i, \alpha'_{T(i)}\}$ units of mass from θ_i to $\theta'_{T(i)}$. The total cost of this mass transfer is $\frac{1}{n}\sum_{i=1}^n \min\{\alpha_i, \alpha'_{T(i)}\}|\theta_i - \theta'_{T(i)}| \leq D W_1(\sigma_n, \sigma'_n)$. Finally, the mass $\frac{1}{n}\sum_{i=1}^n (\alpha_i - \min\{\alpha_i, \alpha'_{T(i)}\})\delta_{\theta_i}$ can be used to cover for the mass demanded in $\frac{1}{n}\sum_{i=1}^n (\alpha'_{T(i)} - \min\{\alpha_i, \alpha'_{T(i)}\})\delta_{\theta'_i}$. This mass transfer can be carried out in any way, the important point being that the total cost of such a mass transfer will not be larger than the total amount of mass to be transferred $\frac{1}{n}\sum_{i=1}^n (\alpha_i - \min\{\alpha_i, \alpha'_{T(i)}\})$ (which is less than $W_1(\sigma_n, \sigma'_n)$) times the diameter of the set Θ. The bottom line is that $W_1(\mathcal{F}\sigma_n, \mathcal{F}\sigma'_n) \leq (D + \mathrm{diam}(\Theta)) W_1(\sigma_n, \sigma_n)$.

We can extend to arbitrary probability measures σ, σ' by noticing that: 1) any probability measure σ can be approximated in the weak sense by empirical probability measures σ_n for growing n; 2) the map $\mathcal{F}$ is continuous in the weak sense (as can be verified directly); 3) given that all measures are supported on a fixed bounded set, Wasserstein metrics are continuous with respect to weak convergence.

Finally, to extend to arbitrary $p \geq 1$, notice that $W_1(\sigma, \sigma') \leq W_p(\sigma, \sigma')$, while $W_p^p(\mathcal{F}\sigma, \mathcal{F}\sigma') \leq \mathrm{diam}(\Theta)^{p-1} W_1(\mathcal{F}\sigma, \mathcal{F}\sigma')$. □

Remark 19 *Lemma 18 also holds, mutatis mutandis, for $\mathcal{F}$ when it acts on measures $\gamma \in \mathcal{Z}^2 \times [0, D]$.*

We now deduce an a priori control on the difference between solutions to (4.5) for two different pairs of measures $(\check{\boldsymbol{\gamma}}^i, \check{\boldsymbol{\sigma}}^i)$, $i = 1, 2$.

Lemma 20 *Let $T, D > 0$. Suppose that Assumption 8 holds. For $i = 1, 2$, let $(\check{\boldsymbol{\gamma}}^i, \check{\boldsymbol{\sigma}}^i) \in \mathcal{A}_{T,D}$, and denote by*

$$\zeta^i = (Z^{\check{\boldsymbol{\gamma}}^i, \check{\boldsymbol{\sigma}}^i}, \tilde{Z}^{\check{\boldsymbol{\gamma}}^i, \check{\boldsymbol{\sigma}}^i}, \omega^{\check{\boldsymbol{\gamma}}^i, \check{\boldsymbol{\sigma}}^i}, \vartheta^{\check{\boldsymbol{\gamma}}^i, \check{\boldsymbol{\sigma}}^i}, \alpha^{\check{\boldsymbol{\gamma}}^i, \check{\boldsymbol{\sigma}}^i}, \beta^{\check{\boldsymbol{\gamma}}^i, \check{\boldsymbol{\sigma}}^i}, \varrho^{\check{\boldsymbol{\gamma}}^i, \check{\boldsymbol{\sigma}}^i})$$

the corresponding evolution determined by (4.5). We assume that $Z_0^{\check{\boldsymbol{\gamma}}^i, \check{\boldsymbol{\sigma}}^i}$ (although possibly random) belongs to the support of $\pi^i_{0,z}$, the marginal on the z coordinate of π^i_0. We also assume that $\omega^i_0, \varrho^i_0, \alpha^i_0, \beta^i_0 \in [0, D]$.

Then there exists a constant $K_{T,D}$, depending only on T, D, the function η, κ, and on the constants in Assumption 8, such that for all $t \in [0, T]$ we have

$$\mathbb{E}|\zeta^1_t - \zeta^2_t| \leq \mathbb{E}[\sup_{0 \leq s \leq t} |\zeta^1_s - \zeta^2_s|]$$

$$\leq K_{T,D} \left(\mathbb{E}|\zeta^1_0 - \zeta^2_0| + \int_0^t \{ W_1(\check{\boldsymbol{\gamma}}^1_s, \check{\boldsymbol{\gamma}}^2_s) + W_1(\check{\boldsymbol{\sigma}}^1_s, \check{\boldsymbol{\sigma}}^2_s) + \mathbb{E}\left(W_1(\pi^2_s(\cdot | Z_0^{\check{\boldsymbol{\gamma}}^2, \check{\boldsymbol{\sigma}}^2}), \pi^1_s(\cdot | Z_0^{\check{\boldsymbol{\gamma}}^1, \check{\boldsymbol{\sigma}}^1})) \right) \} ds \right).$$

In the above, the expectation is taken over the prescribed (joint) initializations of the two systems.

Proof From (4.5) and the Lipschitzness and boundedness conditions in Assumption 8 we get

$$\begin{aligned} |\frac{d}{dt}(\tilde{Z}_t^{\check{\boldsymbol{\gamma}}^1, \check{\boldsymbol{\sigma}}^1} - \tilde{Z}_t^{\check{\boldsymbol{\gamma}}^2, \check{\boldsymbol{\sigma}}^2})| &= \eta_t \left| \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi^1_t, \nu^1_t; Z_t^{\check{\boldsymbol{\gamma}}^1, \check{\boldsymbol{\sigma}}^1}, \tilde{Z}_t^{\check{\boldsymbol{\gamma}}^1, \check{\boldsymbol{\sigma}}^1}) - \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi^2_t, \nu^2_t; Z_t^{\check{\boldsymbol{\gamma}}^2, \check{\boldsymbol{\sigma}}^2}, \tilde{Z}_t^{\check{\boldsymbol{\gamma}}^2, \check{\boldsymbol{\sigma}}^2}) \right| \\ &\leq \eta_t L \{ |Z_t^{\check{\boldsymbol{\gamma}}^1, \check{\boldsymbol{\sigma}}^1} - Z_t^{\check{\boldsymbol{\gamma}}^2, \check{\boldsymbol{\sigma}}^2}| + |\tilde{Z}_t^{\check{\boldsymbol{\gamma}}^1, \check{\boldsymbol{\sigma}}^1} - \tilde{Z}_t^{\check{\boldsymbol{\gamma}}^2, \check{\boldsymbol{\sigma}}^2}| + W_1(\pi^1_t, \pi^2_t) + W_1(\nu^1_t, \nu^2_t) \}. \end{aligned}$$

By performing a similar analysis on the other components of the systems, and using the assumption $(\check{\boldsymbol{\gamma}}^i, \check{\boldsymbol{\sigma}}^i) \in \mathcal{A}_{T,D}$ and Assumption 8, we deduce that we can find a constant $C_{T,D}$ such that for all $t \in [0, T]$

$$\begin{aligned} |\zeta^1_t - \zeta^2_t| &\leq |\zeta^1_0 - \zeta^2_0| \\ &\quad + C_{T,D} \int_0^t \left\{ |\zeta^1_s - \zeta^2_s| + W_1(\pi^1_s, \pi^2_s) + W_1(\nu^1_s, \nu^2_s) + W_1(\pi^1_s(\cdot | Z_s^{\check{\boldsymbol{\gamma}}^1, \check{\boldsymbol{\sigma}}^1}), \pi^2_s(\cdot | Z_s^{\check{\boldsymbol{\gamma}}^2, \check{\boldsymbol{\sigma}}^2})) \right\} ds. \end{aligned}$$

Thus, using Gronwall's inequality, we get that for all $t \in [0, T]$

$$\begin{aligned} |\zeta^1_t - \zeta^2_t| &\leq \sup_{0 \leq s \leq t} |\zeta^1_s - \zeta^2_s| \qquad (4.8) \\ &\leq e^{C_{T,D} T} \left(|\zeta^1_0 - \zeta^2_0| + C_{T,D} \int_0^t \left\{ W_1(\pi^1_s, \pi^2_s) + W_1(\nu^1_s, \nu^2_s) + W_1(\pi^1_s(\cdot | Z_s^{\check{\boldsymbol{\gamma}}^1, \check{\boldsymbol{\sigma}}^1}), \pi^2_s(\cdot | Z_s^{\check{\boldsymbol{\gamma}}^2, \check{\boldsymbol{\sigma}}^2})) \right\} ds \right). \end{aligned}$$

Now, from the fact that $Z_s^{\check{\boldsymbol{\gamma}}^i,\check{\boldsymbol{\sigma}}^i} = Z_0^{\check{\boldsymbol{\gamma}}^i,\check{\boldsymbol{\sigma}}^i}$, it follows

$$\mathbb{E}\left(W_1(\pi_s^2(\cdot|Z_s^{\check{\boldsymbol{\gamma}}^2,\check{\boldsymbol{\sigma}}^2}),\pi_s^1(\cdot|Z_s^{\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\sigma}}^1}))\right) = \mathbb{E}\left(W_1(\pi_s^2(\cdot|Z_0^{\check{\boldsymbol{\gamma}}^2,\check{\boldsymbol{\sigma}}^2}),\pi_s^1(\cdot|Z_0^{\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\sigma}}^1}))\right).$$

From this and (4.8) it then follows that for all $t\in[0,T]$

$$\mathbb{E}[|\zeta_t^1-\zeta_t^2|] \le \mathbb{E}[\sup_{0\le s\le t}|\zeta_s^1-\zeta_s^2|]$$
$$\le e^{C_{T,D}T}\left(|\zeta_0^1-\zeta_0^2| + C_{T,D}\int_0^t W_1(\pi_s^1,\pi_s^2)+W_1(\nu_s^1,\nu_s^2)+\mathbb{E}\left(W_1(\pi_s^2(\cdot|Z_0^{\check{\boldsymbol{\gamma}}^2,\check{\boldsymbol{\sigma}}^2}),\pi_s^1(\cdot|Z_0^{\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\sigma}}^1}))\right)ds\right).$$

To conclude, we apply Proposition 15 and Lemma 18. □

In general, the terms $\mathbb{E}\left(W_1(\pi_s^2(\cdot|Z_0^{\boldsymbol{\gamma}^2,\boldsymbol{\sigma}^2}),\pi_s^1(\cdot|Z_0^{\boldsymbol{\gamma}^1,\boldsymbol{\sigma}^1}))\right)$ cannot be bounded above by the Wasserstein distance between π_s^2 and π_s^1. We thus use a similar construction as in Definition 11. Given two collections $\boldsymbol{\pi}^1 := \{\pi_s^1\}_{0\le s\le T}$ and $\boldsymbol{\pi}^2 := \{\pi_s^2\}_{0\le s\le T}$, we define their cost $\tilde{W}_{t,1}$ by

$$\tilde{W}_{t,1}(\boldsymbol{\pi}^1,\boldsymbol{\pi}^2) := \sup_{s\in[0,t]}\inf_{\upsilon_s\in\Gamma_{\mathrm{Opt}}(\pi^1_{s,z},\pi^2_{s,z})}\left\{\int W_1(\pi_s^2(\cdot|z^2),\pi_s^1(\cdot|z^1))d\upsilon_s(z^1,z^2)\right\}; \tag{4.9}$$

in the above, we interpret $\pi^i_{s,z}$ as the marginal in the z coordinate of the measure π^i_s, and $\Gamma_{\mathrm{Opt}}(\pi^1_{s,z},\pi^2_{s,z})$ is the set of optimal couplings realizing the 1-Wasserstein distance between $\pi^1_{s,z}$ and $\pi^2_{s,z}$.

With this notion in hand, we can state and prove the following corollary of Lemma 20.

Corollary 21 *Suppose the assumptions in Lemma 20 hold. Assume further that $\check{\boldsymbol{\gamma}}_0^1 = \check{\boldsymbol{\gamma}}_0^2$ and $\check{\boldsymbol{\sigma}}_0^1 = \check{\boldsymbol{\sigma}}_0^2$. Then there exists a constant $\tilde{C}_{T,D}$ depending only on M,L,T,D,κ,η such that for all $t\in[0,T]$*

$$W_{t,1}(\Psi_T^1(\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\sigma}}^1),\Psi_T^1(\check{\boldsymbol{\gamma}}^2,\check{\boldsymbol{\sigma}}^2)) + W_{t,1}(\Psi_T^2(\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\sigma}}^1),\Psi_T^2(\check{\boldsymbol{\gamma}}^2,\check{\boldsymbol{\sigma}}^2)) + \tilde{W}_{t,1}(\Psi_T^1(\boldsymbol{\pi}^1),\Psi_T^1(\boldsymbol{\pi}^2))$$
$$\le t\tilde{C}_{T,D}\{W_{t,1}(\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\gamma}}^2)+W_{t,1}(\check{\boldsymbol{\sigma}}^1,\check{\boldsymbol{\sigma}}^2)+\tilde{W}_{t,1}(\boldsymbol{\pi}^1,\boldsymbol{\pi}^2)\}.$$

Here we are abusing notation slightly to denote the collection of measures $\{\mathcal{F}((\Psi_T^1(\check{\boldsymbol{\gamma}}^i,\check{\boldsymbol{\sigma}}^i))_s)\}_{s\in[0,T]}$ by $\Psi_T^1(\boldsymbol{\pi}^i)$.

Proof Take ζ^i for $i=1,2$ in Lemma 20 with identical initial conditions, sampling $(Z_0^1,\tilde{Z}_0^1,\omega_0^1,\varrho_0^1)$ from $\check{\boldsymbol{\gamma}}_0$ and $(\vartheta_0^1,\alpha_0^1,\beta_0^1)$ from $\check{\boldsymbol{\sigma}}_0$. By the definition of the Wasserstein distance it follows

$$W_{t,1}(\Psi_T^1(\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\sigma}}^1),\Psi_T^1(\check{\boldsymbol{\gamma}}^2,\check{\boldsymbol{\sigma}}^2)) + W_{t,1}(\Psi_T^2(\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\sigma}}^1),\Psi_T^2(\check{\boldsymbol{\gamma}}^2,\check{\boldsymbol{\sigma}}^2)) \le \mathbb{E}\sup_{0\le s\le t}|\zeta_s^1-\zeta_s^2|.$$

Now, since $\check{\boldsymbol{\gamma}}_0^1 = \check{\boldsymbol{\gamma}}_0^2$ and $(\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\sigma}}^1),(\check{\boldsymbol{\gamma}}^2,\check{\boldsymbol{\sigma}}^2)\in\mathcal{A}_T$, the optimal couplings υ_s in $\tilde{W}_{t,1}(\boldsymbol{\pi}^1,\boldsymbol{\pi}^2)$ are all the identity coupling for the measure $\pi^1_{0,z}$; the exact same is true for $\tilde{W}_{t,1}(\Psi_T^1(\boldsymbol{\pi}^1),\Psi_T^1(\boldsymbol{\pi}^2))$. It

follows that

$$
\begin{aligned}
&W_{t,1}(\Psi_T^1(\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\sigma}}^1),\Psi_T^1(\check{\boldsymbol{\gamma}}^2,\check{\boldsymbol{\sigma}}^2))+W_{t,1}(\Psi_T^2(\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\sigma}}^1),\Psi_T^2(\check{\boldsymbol{\gamma}}^2,\check{\boldsymbol{\sigma}}^2))\\
&\leq \mathbb{E}\sup_{0\leq s\leq t}|\zeta_s^1-\zeta_s^2|\leq K_{T,D}\int_0^t\{W_1(\check{\boldsymbol{\gamma}}_s^1,\check{\boldsymbol{\gamma}}_s^2)+W_1(\check{\boldsymbol{\sigma}}_s^1,\check{\boldsymbol{\sigma}}_s^2)+\tilde{W}_{s,1}(\boldsymbol{\pi}^1,\boldsymbol{\pi}^2)\}ds\\
&\leq K_{T,D}\int_0^t\{W_{s,1}(\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\gamma}}^2)+W_{s,1}(\check{\boldsymbol{\sigma}}^1,\check{\boldsymbol{\sigma}}^2)+\tilde{W}_{s,1}(\boldsymbol{\pi}^1,\boldsymbol{\pi}^2)\}ds\\
&\leq tK_{T,D}\{W_{t,1}(\check{\boldsymbol{\gamma}}^1,\check{\boldsymbol{\gamma}}^2)+W_{t,1}(\check{\boldsymbol{\sigma}}^1,\check{\boldsymbol{\sigma}}^2)+\tilde{W}_{t,1}(\boldsymbol{\pi}^1,\boldsymbol{\pi}^2)\}.
\end{aligned}
$$

Likewise,

$$
\tilde{W}_{t,1}(\Psi_T^1(\boldsymbol{\pi}^1),\Psi_T^1(\boldsymbol{\pi}^2))\leq\mathbb{E}\sup_{0\leq s\leq t}|\zeta_s^1-\zeta_s^2|.
$$

Putting together the above estimates we obtain the desired result. □

4.1. *Well-posedness of mean-field PDE*

We now look for a system of ODEs characterizing the solution of the system (3.2). The natural candidate is given by the mean-field equation

$$
\begin{aligned}
&(Z^{mf},\tilde{Z}^{mf},\omega^{mf},\vartheta^{mf},\alpha^{mf},\beta^{mf},\varrho^{mf}):=(Z^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\tilde{Z}^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\omega^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\vartheta^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\alpha^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\beta^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}},\varrho^{\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}}});\\
&\qquad\text{with}\quad \check{\boldsymbol{\gamma}}=\mathrm{Law}[((Z^{mf},\tilde{Z}^{mf}),\omega^{mf},\varrho^{mf})],\quad \check{\boldsymbol{\sigma}}=\mathrm{Law}[(\vartheta^{mf},\alpha^{mf},\beta^{mf})].
\end{aligned}
\tag{4.10}
$$

Indeed, assuming that such mean-field equation exists, we verify that setting

$$
\boldsymbol{\gamma}=\mathrm{Law}[((Z^{mf},\tilde{Z}^{mf}),\omega^{mf})],\ \text{and}\ \boldsymbol{\sigma}=\mathrm{Law}[(\vartheta^{mf},\alpha^{mf})]
$$

we satisfy (3.2). Consider two arbitrary testing functions ϕ,φ. We have

$$
\frac{d}{dt}\mathbb{E}[\phi(Z_t^{mf},\tilde{Z}_t^{mf}),\omega_t^{mf})]=\mathbb{E}\left[\nabla_{(z,\tilde{z}),\omega}\phi((Z_t^{mf},\tilde{Z}_t^{mf}),\omega_t^{mf})\cdot[(\frac{d}{dt}Z_t^{mf},\frac{d}{dt}\tilde{Z}_t^{mf}),\frac{d}{dt}\omega_t^{mf}]^\top\right],
$$

and

$$
\frac{d}{dt}\mathbb{E}[\varphi(\vartheta_t^{mf},\omega_t^{mf})]=\mathbb{E}\left[\nabla_{\theta,\alpha}\varphi(\vartheta_t^{mf},\alpha_t^{mf})\cdot[\frac{d}{dt}\vartheta_t^{mf},\frac{d}{dt}\alpha_t^{mf}]^\top\right].
$$

Using the dynamics (4.5) with π,ν as defined, we obtain precisely (3.2) in a weak sense. Note that, since the measure driving the dynamics comes from the distribution of the dynamics itself, our previous argument does not immediately apply. However, the matter is settled by establishing the well-posedness of the system of ODEs (4.10). The proof is based on Banach's fixed-point theorem, which simultaneously guarantees the existence and uniqueness of the solution to this system.

Theorem 22 *Let $D>0$, and suppose that $\check{\gamma}_0$ and $\check{\sigma}_0$ are two probability measures such that $\check{\gamma}_0(\mathcal{Z}^2\times[0,D]^2)=1,\quad \check{\sigma}_0(\Theta\times[0,D]^2)=1$, and such that $\mathcal{F}\check{\sigma}_0$ and $\mathcal{F}\check{\gamma}_0$ are probability measures.*

Then, under Assumption 8, there exists a unique solution to the mean-field system (4.10) with initial distributions $(\check{\gamma}_0, \check{\sigma}_0)$.

Proof Consider the set $\mathcal{A}_{T,D}(\check{\gamma}_0, \check{\sigma}_0) := \{(\check{\boldsymbol{\gamma}}, \check{\boldsymbol{\sigma}}) \in \mathcal{A}_{T,D} \text{ s.t. } \check{\boldsymbol{\gamma}}_0 = \check{\gamma}_0, \quad \check{\boldsymbol{\sigma}}_0 = \check{\sigma}_0\}$. As shown, for example, in [7], we can deduce that the set $\mathcal{A}_T$ endowed with the metric $W_{T,1}(\check{\boldsymbol{\gamma}}^1, \check{\boldsymbol{\gamma}}^2) + W_{T,1}(\check{\boldsymbol{\sigma}}^1, \check{\boldsymbol{\sigma}}^2)$ is a complete metric space given that $\mathcal{C}([0,T], \mathcal{Z}^2 \times \mathbb{R}^2_+)$ (respectively $\mathcal{C}([0,T], \Theta \times \mathbb{R}^2_+)$) is complete with respect to the distance function $d(u,v) := \sup_{s\in[0,T]} |u_s - v_s|$. It is straightforward to see that this property is inherited by $\mathcal{A}_{T,D}(\check{\gamma}_0, \check{\sigma}_0)$. Note also that by Corollary 21 one can find $T > 0$ small enough so that Ψ contracts the quantity $W_{T,1}(\check{\boldsymbol{\gamma}}^1, \check{\boldsymbol{\gamma}}^2) + W_{T,1}(\check{\boldsymbol{\sigma}}^1, \check{\boldsymbol{\sigma}}^2) + \tilde{W}_{T,1}(\boldsymbol{\pi}^1, \boldsymbol{\pi}^2)$ in the space $\mathcal{A}_{T,D}(\check{\gamma}_0, \check{\sigma}_0)$. Now, the latter quantity dominates the metric in the space $\mathcal{A}_{T,D}(\check{\gamma}_0, \check{\sigma}_0)$. Hence, there is a unique solution $(\check{\boldsymbol{\gamma}}, \check{\boldsymbol{\sigma}}) \in \mathcal{A}_{T,D}(\check{\gamma}_0, \check{\sigma}_0)$ to the fixed point equation

$$\Psi(\check{\boldsymbol{\gamma}}, \check{\boldsymbol{\sigma}}) = (\check{\boldsymbol{\gamma}}, \check{\boldsymbol{\sigma}}).$$

By definition, the mean-field system (4.10) is then satisfied and is well-posed in the interval $[0,T]$. By continuation, well-posedness can be arbitrarily extended. □

Remark 23 *Since (3.2) is well-defined, we can also conclude that the system of mean-field particles (4.4) is well-defined given that it can be obtained by plugging in the mean-field law, except that the initial condition is not sampled from* $\check{\gamma}_0, \check{\sigma}_0$ *but taken as in the system (4.1).*

4.2. *Propagation of chaos*

Before stating our propagation of chaos result we first present a lemma.

Lemma 24 *Let* $(\check{\gamma}_0, \check{\sigma}_0)$ *be such that* $\check{\gamma}_0(\mathcal{Z}^2 \times [0,D]^2) = 1$, $\check{\sigma}_0(\Theta \times [0,D]^2) = 1$, *and such that* $\mathcal{F}\check{\sigma}_0$ *and* $\mathcal{F}\check{\gamma}_0$ *are probability measures. Let* $(\check{\boldsymbol{\gamma}}, \check{\boldsymbol{\sigma}})$ *be the law of (4.10) with* $(\check{\boldsymbol{\gamma}}_0, \check{\boldsymbol{\sigma}}_0) = (\check{\gamma}_0, \check{\sigma}_0)$. *Let* z_0' *and* z_0 *be two arbitrary points in the support of* $\pi_{0,z}$. *Then for every* $t \in [0,T]$ *we have*

$$\sup_{s\in[0,t]} W_1(\pi_s(\cdot|z_0'), \pi_s(\cdot|z_0)) \leq K_{T,D}(W_1(\pi_0(\cdot|z_0'), \pi_0(\cdot|z_0)) + |z_0 - z_0'|). \tag{4.11}$$

Proof Consider one particle as in (4.4) that we denote by ζ and that we initialize at $Z_0 = z_0$ and $(\tilde{Z}_0, \omega_0, \varrho_0) \sim \check{\boldsymbol{\gamma}}_0(\cdot|z_0)$ and $(\vartheta_0, \alpha_0, \beta_0) \sim \check{\boldsymbol{\sigma}}_0$. Likewise, consider another particle as in (4.4) that we denote by ζ' and that we initialize at $Z_0' = z_0'$, $(\tilde{Z}_0', \omega_0', \varrho_0') \sim \check{\boldsymbol{\gamma}}_0(\cdot|z_0)$, and $(\vartheta_0', \alpha_0', \beta_0') = (\vartheta_0, \alpha_0, \beta_0)$. At this point we leave unspecified the joint distribution for the initializations of the variables $\tilde{z}, \omega, \varrho$, but it is understood that one such coupling has been fixed in the computations below.

An application of Lemma (20) deduces that for every $t \in [0,T]$

$$\mathbb{E}[\sup_{0\le s\le t}|\zeta'-\zeta|] \le K_{T,D}\mathbb{E}|\zeta_0'-\zeta_0| + K_{T,D}\int_0^t W_1(\pi_s(\cdot|z_0'),\pi_s(\cdot|z_0))ds.$$

By definition of the Wasserstein distance, the left hand side of the above expression can be bounded from below by $W_1(\pi_s(\cdot|z_0'),\pi_s(\cdot|z_0))$ for any $s\in[0,t]$, and thus

$$\sup_{s\in[0,t]} W_1(\pi_s(\cdot|z_0'),\pi_s(\cdot|z_0)) \le K_{T,D}\mathbb{E}|\zeta_0'-\zeta_0| + K_{T,D}\int_0^t W_1(\pi_s(\cdot|z_0'),\pi_s(\cdot|z_0))ds.$$

By using the fact that the coupling between the distributions for the variables $(\tilde z,\omega,\varrho)$ was arbitrary we can conclude that

$$\sup_{s\in[0,t]} W_1(\pi_s(\cdot|z_0'),\pi_s(\cdot|z_0)) \le K_{T,D}(W_1(\pi_0(\cdot|z_0'),\pi_0(\cdot|z_0)) + |z_0-z_0'|) + K_{T,D}\int_0^t W_1(\pi_s(\cdot|z_0'),\pi_s(\cdot|z_0))ds.$$

At this stage we can apply a Gronwall-type argument to obtain the desired result.

□

Theorem 25 (Propagation of chaos) *Let $T,D>0$, and suppose that Assumption 8 holds. Let $(\check\gamma_0,\check\sigma_0)$ be such that $\check\gamma_0(\mathcal{Z}^2\times[0,D]^2)=1$ and $\check\sigma_0(\Theta\times[0,D]^2)=1$, and suppose that $\mathcal{F}\check\sigma_0$ and $\mathcal{F}\check\gamma_0$ are probability measures.*

For $N\in\mathbb{N}$ consider the system (4.1) associated to a sequence $\{(\check\gamma_0^N,\check\sigma_0^N)\}_{N\in\mathbb{N}}$ satisfying $\check\gamma_0^N(\mathcal{Z}^2\times[0,D]^2)=1$ and $\check\sigma_0^N(\Theta\times[0,D]^2)=1$ for all large enough N, and suppose that $\mathcal{F}\check\sigma_0^N$ and $\mathcal{F}\check\gamma_0^N$ are probability measures. We also assume that the Z_0^i belong to the support of the measure $\pi_{0,z}$.

Assume further that as $N\to\infty$ we have

$$\inf_{\upsilon_z\in\Gamma_{Opt}(\pi_{0,z}^N,\pi_{0,z})}\int W_1(\check\gamma_0^N(\cdot|z_0'),\check\gamma_0(\cdot|z_0))d\upsilon_z(z_0',z_0)\to 0, \text{ and } W_1(\check\sigma_0^N,\check\sigma_0)\to 0. \tag{4.12}$$

Then

$$W_{T,1}(\check{\boldsymbol\gamma}^N,\check{\boldsymbol\gamma})\to 0, \qquad W_{T,1}(\check{\boldsymbol\sigma}^N,\check{\boldsymbol\sigma})\to 0, \quad \tilde W_{T,1}(\boldsymbol\pi^N,\boldsymbol\pi)\to 0,$$

where $(\check{\boldsymbol\gamma},\check{\boldsymbol\sigma})$ are the laws of the mean-field system (4.10) with initial conditions drawn from $(\check\gamma_0,\check\sigma_0)$.

Proof We assume without loss of generality that for every N and every $i=1,\dots,N$, the weights ω_0^i,α_0^i belong to $[0,D]$. From Gronwall's inequality and Assumption 8 we can then see that the weights ω_t^i,α_t^i belong to $[0,De^{2M\kappa t}]$. It follows that $(\check{\boldsymbol\gamma}^N,\check{\boldsymbol\sigma}^N)\in\mathcal{A}_{T,D}$.

In what follows we let ζ^i denote the path of all variables of the i-th particle in the system (4.1), and $\zeta^{mf,i}$ the corresponding particle in (4.4); we recall that these particles are assumed to be initialized at the same location. We consider

$$\check{\boldsymbol\gamma}_t^{N,mf} := \frac{1}{N}\sum_{i=1}^N \delta_{(Z_t^{mf,i},\tilde Z_t^{mf,i}),\omega_t^{mf,i},\varrho_t^{mf,i}} \quad\text{and}\quad \check{\boldsymbol\sigma}_t^{N,mf} := \frac{1}{N}\sum_{i=1}^N \delta_{\vartheta_t^{mf,i},\alpha_t^{mf,i},\beta_t^{mf,i}},$$

that is, the empirical measures of the mean-field system of particles.

From the triangle inequality we have

$$W_{t,1}(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\gamma}}^N)+W_{t,1}(\check{\boldsymbol{\sigma}},\check{\boldsymbol{\sigma}}^N)\leq\{W_{t,1}(\check{\boldsymbol{\gamma}}^{N,mf},\check{\boldsymbol{\gamma}}^N)+W_{t,1}(\check{\boldsymbol{\sigma}}^{N,mf},\check{\boldsymbol{\sigma}}^N)\}+\{W_{t,1}(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\gamma}}^{N,mf})+W_{t,1}(\check{\boldsymbol{\sigma}},\check{\boldsymbol{\sigma}}^{N,mf})\}. \tag{4.13}$$

We now claim that a similar inequality holds for the term $\tilde{W}_{t,1}(\boldsymbol{\pi},\boldsymbol{\pi}^N)$. Namely, we prove that

$$\tilde{W}_{t,1}(\boldsymbol{\pi},\boldsymbol{\pi}^N)\leq\tilde{W}_{t,1}(\boldsymbol{\pi},\boldsymbol{\pi}^{N,mf})+\tilde{W}_{t,1}(\boldsymbol{\pi}^{N,mf},\boldsymbol{\pi}^N). \tag{4.14}$$

To see this, recall that the z coordinates of all dynamics remain unchanged and that the initializations of ζ^i and $\zeta^{mf,i}$ are the same. It follows that $\pi^N_{s,z}=\pi^N_{0,z}=\pi^{N,mf}_{0,z}=\pi^{N,mf}_{s,z}$, and thus $\Gamma_{\mathrm{Opt}}(\pi^{N,mf}_{s,z},\pi^N_{s,z})$ consists exclusively of the identity coupling. Let $\upsilon\in\Gamma_{\mathrm{Opt}}(\pi_{s,z},\pi^N_{s,z})=\Gamma_{\mathrm{Opt}}(\pi_{0,z},\pi^N_{0,z})$. From the triangle inequality for W_1 we deduce

$$\begin{aligned}\int W_1(\pi_s(\cdot|z),\pi^N_s(\cdot|z'))d\upsilon(z,z')&\leq\int(W_1(\pi_s(\cdot|z),\pi^{N,mf}_s(\cdot|z'))+W_1(\pi^{N,mf}_s(\cdot|z'),\pi^N_s(\cdot|z')))d\upsilon(z,z')\\&\leq\int W_1(\pi_s(\cdot|z),\pi^{N,mf}_s(\cdot|z'))d\upsilon(z,z')+\int W_1(\pi^{N,mf}_s(\cdot|z'),\pi^N_s(\cdot|z'))d\pi^N_{s,z}(z')\\&\leq\int W_1(\pi_s(\cdot|z),\pi^{N,mf}_s(\cdot|z'))d\upsilon(z,z')+\tilde{W}_{t,1}(\boldsymbol{\pi}^{N,mf},\boldsymbol{\pi}^N),\end{aligned}$$

for every $s\in[0,t]$. Taking the inf over υ on both sides and then the sup over $s\in[0,t]$, we obtain (4.14).

We use again the fact that the particles ζ^i and $\zeta^{mf,i}$ have the same initialization to proceed as in the proof of Corollary 21 and conclude that for every $t\in[0,T]$

$$W_{t,1}(\check{\boldsymbol{\gamma}}^{N,mf},\check{\boldsymbol{\gamma}}^N)+W_{t,1}(\check{\boldsymbol{\sigma}}^{N,mf},\check{\boldsymbol{\sigma}}^N)\leq K_{T,D}\int_0^t\{W_{s,1}(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\gamma}}^N)+W_{s,1}(\check{\boldsymbol{\sigma}},\check{\boldsymbol{\sigma}}^N)+\tilde{W}_{s,1}(\boldsymbol{\pi},\boldsymbol{\pi}^N)\}ds,$$

as well as

$$\tilde{W}_{t,1}(\boldsymbol{\pi}^{N,mf},\boldsymbol{\pi}^N)\leq K_{T,D}\int_0^t\{W_{s,1}(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\gamma}}^N)+W_{s,1}(\check{\boldsymbol{\sigma}},\check{\boldsymbol{\sigma}}^N)+\tilde{W}_{s,1}(\boldsymbol{\pi},\boldsymbol{\pi}^N)\}ds.$$

We can now combine the previous two inequalities with (4.13) and (4.14) to conclude that

$$\begin{aligned}W_{t,1}(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\gamma}}^N)+W_{t,1}(\check{\boldsymbol{\sigma}},\check{\boldsymbol{\sigma}}^N)+\tilde{W}_{t,1}(\boldsymbol{\pi},\boldsymbol{\pi}^N)\leq&\\&\{W_{t,1}(\check{\boldsymbol{\gamma}}^{N,mf},\check{\boldsymbol{\gamma}})+W_{t,1}(\check{\boldsymbol{\sigma}}^{N,mf},\check{\boldsymbol{\sigma}})+\tilde{W}_{t,1}(\boldsymbol{\pi}^{N,mf},\boldsymbol{\pi})\}\\&+K_{T,D}\int_0^t\{W_{s,1}(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\gamma}}^N)+W_{s,1}(\check{\boldsymbol{\sigma}},\check{\boldsymbol{\sigma}}^N)+\tilde{W}_{s,1}(\boldsymbol{\pi},\boldsymbol{\pi}^N)\}ds.\end{aligned}$$

Combining with Gronwall's inequality, the above implies

$$W_{t,1}(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\gamma}}^N)+W^1_{t,1}(\check{\boldsymbol{\sigma}},\check{\boldsymbol{\sigma}}^N)+\tilde{W}_{t,1}(\boldsymbol{\pi},\boldsymbol{\pi}^N)\leq e^{tK_{T,D}}\{W_{t,1}(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\gamma}}^{N,mf})+W_{t,1}(\check{\boldsymbol{\sigma}},\check{\boldsymbol{\sigma}}^{N,mf})+\tilde{W}_{t,1}(\boldsymbol{\pi},\boldsymbol{\pi}^{N,mf})\}.$$

To complete the proof we must show that the right hand side of the above expression goes to zero as $N\to\infty$. For that purpose we compare the evolutions of $\zeta^{mf,i}_{\mathcal{Z}}:=$

$(Z^{mf,i},\tilde{Z}^{mf,i},\omega^{mf,i},\varrho^{mf,i})$ and $\zeta_{\mathcal{Z}}^{mf}:=(Z^{mf},\tilde{Z}^{mf},\omega^{mf},\varrho^{mf})$, and then, separately, compare the evolutions of $\zeta_{\Theta}^{mf,i}:=(\vartheta^{mf,i},\alpha^{mf,i},\beta^{mf,i})$ and $\zeta_{\Theta}^{mf}:=(\vartheta^{mf},\alpha^{mf},\beta^{mf})$. For the first pair of evolutions, we proceed as in the proof of Lemma 20 to conclude

$$\sup_{0\leq s\leq t}|\zeta_{\mathcal{Z},s}^{mf,i}-\zeta_{\mathcal{Z},s}^{mf}|\leq K_{T,D}|\zeta_{\mathcal{Z},0}^{mf,i}-\zeta_{\mathcal{Z},0}^{mf}|+K_{T,D}\int_0^t W_1(\pi_s(\cdot|Z_0^{mf}),\pi_s(\cdot|Z_0^{mf,i}))ds.$$

We can then use Lemma 24 to obtain

$$\sup_{0\leq s\leq t}|\zeta_{\mathcal{Z},s}^{mf,i}-\zeta_{\mathcal{Z},s}^{mf}|\leq K_{T,D}|\zeta_{\mathcal{Z},0}^{mf,i}-\zeta_{\mathcal{Z},0}^{mf}|+K_{T,D}W_1(\pi_0(\cdot|Z_0^{mf}),\pi_0(\cdot|Z_0^{mf,i})).$$

Combining the above pathwise estimate with the freedom to choose the coupling for the initializations, we can conclude that

$$W_{T,1}(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\gamma}}^{N,mf}),W_{T,1}(\boldsymbol{\pi},\boldsymbol{\pi}^{N,mf})\leq K_{T,D}W_1(\pi_{0,z},\pi_{0,z}^N)$$
$$+K_{T,D}\inf_{\upsilon_z\in\Gamma_{\mathrm{Opt}}(\pi_{0,z}^N,\pi_{0,z})}\int W_1(\check{\gamma}_0^N(\cdot|z_0'),\check{\gamma}_0(\cdot|z_0))d\upsilon_z(z_0',z_0).$$

By assumption (4.12), Remark 60, and Lemma 18, it follows that the right hand side of the above expression goes to zero as $N\to\infty$. For the pair of evolutions $\zeta_{\Theta}^{mf,i}$ and ζ_{Θ}^{mf} we proceed as in the proof of Lemma 20, this time noticing that we can write

$$\sup_{0\leq s\leq t}|\zeta_{\Theta,s}^{mf,i}-\zeta_{\Theta,s}^{mf}|\leq K_{T,D}|\zeta_{\Theta,0}^{mf,i}-\zeta_{\Theta,0}^{mf}|.$$

Combining the above pathwise estimate with the freedom to choose the coupling for the initializations, we can conclude that

$$W_{T,1}(\check{\sigma},\check{\sigma}^{N,mf})\leq K_{T,D}W_1(\check{\sigma}_0,\check{\sigma}_0^N)\to 0. \tag{4.15}$$

□

4.3. *Proof of Theorem 12 and other corollaries of Theorem 25*

In this section we establish some important results that are implied by Theorem 25. The first one is Theorem 12.

Proof of Theorem 12 Let (γ_0,σ_0) be such that $\mathcal{F}\gamma_0=\overline{\pi}_0$ and $\mathcal{F}\sigma_0=\overline{\nu}_0$ and such that (4.3) holds. We introduce the measures $\check{\gamma}_0:=\gamma_0\otimes\delta_1(d\varrho)$, and $\check{\sigma}_0:=\sigma_0\otimes\delta_1(d\beta)$. That is, $\check{\gamma}_0$ is the product of γ_0 and a Dirac delta at the value 1 for the ϱ coordinate; $\check{\sigma}_0$ is defined analogously. Likewise, we define $\check{\gamma}_0^N:=\gamma_0^N\otimes\delta_1$ and $\check{\sigma}_0^N:=\sigma_0^N\otimes\delta_1$. It is clear that with these definitions we have (4.12) and thus we can invoke Theorem 25 to deduce

$$W_{T,1}(\check{\boldsymbol{\gamma}}^N,\check{\boldsymbol{\gamma}})+W_{T,1}(\check{\boldsymbol{\sigma}}^N,\check{\boldsymbol{\sigma}})\to 0,$$

where $(\check{\boldsymbol{\gamma}}^N,\check{\boldsymbol{\sigma}}^N)$ is the measure in path space induced by the particle system (4.1) with initializations as described in the statement of the theorem and with $\beta_0^i=\varrho_0^i=1$ for all i; $(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}})$, on the other hand, is the law in (4.10) with initialization $(\check{\boldsymbol{\gamma}}_0,\check{\boldsymbol{\sigma}}_0)=(\check{\gamma}_0,\check{\sigma}_0)$.

Using Lemma 18, we conclude that for every $t \in [0,T]$

$$W_1(\pi_t^N, \pi_t) = W_1(\mathcal{F}\boldsymbol{\gamma}_t^N, \mathcal{F}\boldsymbol{\gamma}_t) \leq K_{T,D} W_1(\boldsymbol{\gamma}_t^N, \boldsymbol{\sigma}_t^N) \leq K_{T,D} W_1(\check{\boldsymbol{\gamma}}_t^N, \check{\boldsymbol{\gamma}}_t) \leq K_{T,D} W_{T,1}(\check{\boldsymbol{\gamma}}^N, \check{\boldsymbol{\gamma}}).$$

Likewise,

$$W_1(\nu_t^N, \nu_t) \leq K_{T,D} W_{T,1}(\check{\boldsymbol{\sigma}}^N, \check{\boldsymbol{\sigma}}).$$

Taking the sup over all $t \in [0,T]$ in the sum of the above two expressions we get

$$\sup_{t\in[0,T]} \{W_1(\pi_t^N, \pi_t) + W_1(\nu_t^N, \nu_t)\} \leq K_{T,D}(W_{T,1}(\check{\boldsymbol{\gamma}}^N, \check{\boldsymbol{\gamma}}) + W_{T,1}(\check{\boldsymbol{\sigma}}^N, \check{\boldsymbol{\sigma}})),$$

from where the desired result now follows.

□

Corollary 30 and Remark 31 below, which we will use in section 5, are the other important consequences of Theorem 25 that we discuss in this section. In section 5 we consider an evolution $\{(\hat{\nu}_t, \hat{\pi}_t)\}_t$ closely related to (2.3) that is given by

$$\partial_t \hat{\nu}_t = \eta_t \mathrm{div}(\hat{\nu}_t \nabla_\theta \mathcal{U}_\nu(\pi_t, \nu_t; \theta)), \quad \partial_t \hat{\pi}_t = -\eta_t \mathrm{div}(\hat{\pi}_t(0, \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t, \nu_t; (z, \tilde{z})))), \tag{4.16}$$

with initializations $\hat{\nu}_0, \hat{\pi}_0$ that are absolutely continuous with respect to ν_0 and π_0, respectively. It is at this stage that we use the extra coordinates β, ϱ in (4.1). Indeed, these variables have been introduced to accommodate for the changes of measure between $\hat{\nu}_0$ and ν_0 and between $\hat{\pi}_0$ and π_0. We will be able to use the general purpose Theorem 25 to prove the consistency of particle approximations for the system (4.16).

We start with a preliminary result.

Proposition 26 *Let $\nu_t^N = \frac{1}{N}\sum_{i=1}^n \alpha_i(t)\delta_{\vartheta_i(t)}$ and $\pi_t^N = \frac{1}{N}\sum_{i=1}^N \omega_i(t)\delta_{(Z_i(t),\tilde{Z}_i(t))}$ be as in (4.2).*

Let $\beta_1, \ldots, \beta_N$ and $\varrho_1, \ldots, \varrho_N$ be two collections of non-negative scalars satisfying

$$\frac{1}{N}\sum_{i=1}^N \beta_i \alpha_i(0) = 1, \quad \frac{1}{N}\sum_{i=1}^N \varrho_i \omega_i(0) = 1.$$

Let $\hat{\nu}_t^N$ and $\hat{\pi}_t^N$ be the probability measures defined as

$$\hat{\nu}_t^N := \frac{1}{N}\sum_{i=1}^n \beta_i \alpha_i(0)\delta_{\vartheta_i(t)}, \quad \hat{\pi}_t^N := \frac{1}{N}\sum_{i=1}^N \varrho_i \omega_i(0)\delta_{(Z_i(t),\tilde{Z}_i(t))} \quad t \geq 0.$$

Then

$$\partial_t \hat{\nu}_t^N = \eta_t \mathrm{div}(\hat{\nu}_t^N \nabla_\theta \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \theta)), \quad \partial_t \hat{\pi}_t^N = -\eta_t \mathrm{div}(\hat{\pi}_t^N(0, \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t^N, \nu_t^N; (z, \tilde{z})))) \tag{4.17}$$

in the weak sense.

Proof Let $\phi(\theta)$ be an arbitrary test function. From (4.1) we see that

$$\begin{aligned}\frac{d}{dt}\int \phi(\theta)d\hat{\nu}_t^N(\theta) &= \frac{1}{N}\sum_{i=1}^N \beta_i\alpha_i(0)\frac{d}{dt}\phi(\vartheta_i(t)) = \frac{1}{N}\sum_{i=1}^N \beta_i\alpha_i(0)\nabla\phi(\vartheta_i(t))\cdot\dot{\vartheta}_i(t)\\ &= -\frac{\eta_t}{N}\sum_{i=1}^N \beta_i\alpha_i(0)\nabla\phi(\vartheta_i(t))\cdot\nabla_\theta\mathcal{U}_\nu(\pi_t^N,\nu_t^N;\vartheta_i(t))\\ &= \eta_t\int \nabla\phi(\theta)\cdot\nabla_\theta\mathcal{U}_\nu(\pi_t^N,\nu_t^N;\theta)d\hat{\nu}_t^N(\theta).\end{aligned}$$

This shows that $\hat{\nu}^N$ solves equation (4.17) in the weak sense. The equation for $\hat{\pi}^N$ is deduced similarly. □

We will now proceed to relate (4.17) with (4.16). We first introduce some additional mathematical tools that will help us in this aim.

Let $\check{\mathcal{F}}:\mathcal{P}(\mathcal{Z}^2\times\mathbb{R}_+^2)\to\mathcal{M}_+(\mathcal{Z}^2)$ be the map defined via the identity

$$\int \phi(\theta)d(\check{\mathcal{F}}\check{\sigma})(\theta) = \int \alpha\beta\phi(\theta)d\check{\sigma}(\theta,\alpha,\beta),$$

for all test functions ϕ. Analogously, define $\check{F}$ as a map $\check{\mathcal{F}}:\mathcal{P}(\Theta\times\mathbb{R}_+^2)\to\mathcal{M}_+(\Theta)$, substituting any appearance of $\check{\sigma},\theta,\alpha,\beta$ in the above with $\check{\gamma},(z,\tilde{z}),\omega,\varrho$. Notice that $\check{\mathcal{F}}\check{\sigma}$ is a probability measure provided that $\int\alpha\beta d\check{\sigma}(\theta,\alpha,\beta)=1$, while an analogous statement holds when $\check{\mathcal{F}}$ acts on $\check{\gamma}$.

Let us now introduce a map $\mathcal{G}:\mathcal{C}([0,T],\mathcal{Z}^2\times\mathbb{R}_+^2)\to\mathcal{C}([0,T],\mathcal{Z}^2\times\mathbb{R}_+^2)$ defined as:

$$\mathcal{G}:\{(z_t,\tilde{z}_t,\omega_t,\varrho_t)\}_{t\in[0,T]}\mapsto\{(z_t,\tilde{z}_t,\omega_0,\varrho_0)\}_{t\in[0,T]}. \tag{4.18}$$

That is, $\mathcal{G}$ is the map that freezes the coordinates ω,ϱ of a given path, setting them to be equal to their initializations. Naturally, $\mathcal{G}$ induces, via pushforward, a map from $\mathcal{P}(\mathcal{C}([0,T],\mathcal{Z}^2\times\mathbb{R}_+^2))$ into itself; we will abuse notation slightly and will also use $\mathcal{G}$ to denote this induced map. Furthermore, we will also think of $\mathcal{G}$ as a map $\mathcal{G}:\mathcal{C}([0,T],\Theta\times\mathbb{R}_+^2)\to\mathcal{C}([0,T],\Theta\times\mathbb{R}_+^2)$ that freezes the coordinates α,β of a given path, setting them to be equal to their initializations; we will also denote by $\mathcal{G}$ the map induced via pushforward from $\mathcal{P}(\mathcal{C}([0,T],\Theta\times\mathbb{R}_+^2))$ into itself. Which of the interpretations for $\mathcal{G}$ will be used in each instance should be clear from context.

Remark 27 *Notice that $\hat{\pi}_t^N$ and $\hat{\nu}_t^N$ in Proposition 26 can be written as $\check{\mathcal{F}}((\mathcal{G}\check{\boldsymbol{\gamma}}^N)_t)$ and $\check{\mathcal{F}}((\mathcal{G}\check{\boldsymbol{\sigma}}^N)_t)$, respectively.*

Lemma 28 *Let $(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\sigma}})$ be the law of the process (4.10) initialized at a pair $(\check{\gamma}_0,\check{\sigma}_0)$. Then $\{\hat{\nu}_t:=\check{\mathcal{F}}((\mathcal{G}\check{\boldsymbol{\sigma}})_t)\}_{t\in[0,T]}$ and $\{\hat{\pi}_t:=\check{\mathcal{F}}((\mathcal{G}\check{\boldsymbol{\gamma}})_t)\}_{t\in[0,T]}$ solve the PDEs (4.16), where $\pi_t=\mathcal{F}(\boldsymbol{\gamma}_t)$ and $\nu_t=\mathcal{F}(\boldsymbol{\sigma}_t)$.*

Proof Consider the mean-field ODE (4.10). For every smooth test function ϕ we have

$$\int \phi(\theta) d\hat{\nu}_t(\theta) = \int \alpha\beta\phi(\theta) d(\mathcal{G}\sigma)_t(\theta,\alpha,\beta) = \mathbb{E}[\alpha_0\beta_0\phi(\vartheta_t)].$$

In particular,

$$\begin{aligned}\frac{d}{dt}\int \phi(\theta) d\hat{\nu}_t(\theta) = \frac{d}{dt}\mathbb{E}[\alpha_0\beta_0\phi(\vartheta_t)] &= -\mathbb{E}[\eta_t\alpha_0\beta_0\nabla\phi(\vartheta_t)\cdot\nabla_\theta\mathcal{U}_\nu(\pi_t,\nu_t;\vartheta_t)] \\ &= -\eta_t\int \nabla\phi(\theta)\cdot\nabla_\theta\mathcal{U}_\nu(\pi_t,\nu_t;\theta) d\hat{\nu}_t(\theta).\end{aligned}$$

This proves that $\hat{\nu}$ satisfies the desired equation. The equation for $\hat{\pi}$ is obtained similarly. □

Remark 29 *Notice that $\hat{\nu}_t$ and $\hat{\pi}_t$ are probability measures if $\hat{\nu}_0$ and $\hat{\pi}_0$ are.*

In what follows, we use Theorem 25 to show that, under appropriate assumptions on initializations, the system in (4.17) can be recovered from suitable particle approximations.

Corollary 30 *Let $\overline{\nu}_0$ and $\overline{\pi}_0$ be arbitrary, and let $\hat{\nu}_0$ and $\hat{\pi}_0$ be probability measures such that $\hat{\nu}_0 \ll \overline{\nu}_0$, $\hat{\pi}_0 \ll \overline{\pi}_0$, with $\frac{d\hat{\nu}_0}{d\overline{\nu}_0} \in L^\infty(\overline{\nu}_0)$ and $\frac{d\hat{\pi}_0}{d\overline{\pi}_0} \in L^\infty(\overline{\pi}_0)$. Let $\check{\gamma}_0$ and $\check{\sigma}_0$ be as in Theorem 25 and additionally assume they satisfy $\mathcal{F}\check{\gamma}_0 = \overline{\pi}_0$, $\mathcal{F}\check{\sigma}_0 = \overline{\nu}_0$, and $\check{\mathcal{F}}\check{\gamma}_0 = \hat{\pi}_0$, $\check{\mathcal{F}}\check{\sigma}_0 = \hat{\nu}_0$.*

Consider approximating particle systems as in Theorem 25 with the additional assumption that $\hat{\nu}_0^N, \hat{\pi}_0^N$ are probability measures.

Then

$$\sup_{t\in[0,T]}\{W_1(\hat{\nu}_t^N,\hat{\nu}_t) + W_1(\hat{\pi}_t^N,\hat{\pi}_t)\} \to 0, \qquad \sup_{t\in[0,T]}\{W_1(\nu_t^N,\nu_t) + W_1(\pi_t^N,\pi_t)\} \to 0,$$

as $N\to\infty$. In the above, we use the same notation as in Lemma 28 and Remark 27.

Proof First of all, let us notice that the condition $\frac{d\hat{\nu}_0}{d\overline{\nu}_0} \in L^\infty(\overline{\nu}_0)$ and $\frac{d\hat{\pi}_0}{d\overline{\pi}_0} \in L^\infty(\overline{\pi}_0)$ is used to guarantee that we can indeed build $\check{\sigma}_0$ and $\check{\gamma}_0$ with bounded supports; see the first part of Remark 31 below.

It is straightforward to check that $\mathcal{G}$ is a Lipschitz map, i.e.,

$$W_{T,1}(\mathcal{G}\check{\boldsymbol{\sigma}},\mathcal{G}\check{\boldsymbol{\sigma}}') \le 2W_{T,1}(\check{\boldsymbol{\sigma}},\check{\boldsymbol{\sigma}}'), \quad W_{T,1}(\mathcal{G}\check{\boldsymbol{\gamma}},\mathcal{G}\check{\boldsymbol{\gamma}}') \le 2W_{T,1}(\check{\boldsymbol{\gamma}},\check{\boldsymbol{\gamma}}').$$

In addition, we can find a constant $C_{T,D}$ such that for every $t\in[0,T]$

$$W_1(\check{\mathcal{F}}((\mathcal{G}\check{\boldsymbol{\sigma}})_t),\check{\mathcal{F}}((\mathcal{G}\check{\boldsymbol{\sigma}}^N)_t)) \le C_{T,D}W_1((\mathcal{G}\check{\boldsymbol{\sigma}})_t),(\mathcal{G}\check{\boldsymbol{\sigma}}^N)_t) \le C_{T,D}W_{T,1}(\mathcal{G}\check{\boldsymbol{\sigma}},\mathcal{G}\check{\boldsymbol{\sigma}}^N),$$

where the first inequality follows from a very similar approach to the one in Lemma 18. Similarly,

$$W_1(\check{\mathcal{F}}((\mathcal{G}\check{\boldsymbol{\gamma}})_t),\check{\mathcal{F}}((\mathcal{G}\check{\boldsymbol{\gamma}}^N)_t)) \le C_{T,D}W_{T,1}(\mathcal{G}\check{\boldsymbol{\gamma}},\mathcal{G}\check{\boldsymbol{\gamma}}^N).$$

We may now combine the above inequalities with Theorem 25, which allows us to obtain

$$W_{T,1}(\check{\boldsymbol{\gamma}}^N,\check{\boldsymbol{\gamma}}) + W_{T,1}(\check{\boldsymbol{\sigma}}^N,\check{\boldsymbol{\sigma}}) \to 0,$$

to deduce the desired convergence.

□

Remark 31 (Constructing initializations) *Let $\overline{\pi}_0$ and $\overline{\nu}_0$ be arbitrary, and let $\rho_\nu = \frac{d\hat{\nu}_0}{d\overline{\nu}_0}$ and $\rho_\pi = \frac{d\hat{\pi}_0}{d\overline{\pi}_0}$, which we assume satisfy $\rho_\nu \in L^\infty(\overline{\nu}_0)$ and $\rho_\pi \in L^\infty(\overline{\pi}_0)$; we further assume that $\hat{\pi}_{0,z} = \pi_{0,z}$. The latter assumption implies that $\int \rho_\pi(z,\tilde{z}) d\pi_0(\tilde{z}|z) = 1$, for all z in the support of $\pi_{0,z}$.*

Let $\check{\gamma}_0$ and $\check{\sigma}_0$ be the measures $\check{\gamma}_0 := h_{\gamma\sharp}\overline{\pi}_0$, $\check{\sigma}_0 := h_{\sigma\sharp}\overline{\nu}_0$, $h_\gamma : (z,\tilde{z}) \mapsto (z,\tilde{z},1,\rho_\pi(z,\tilde{z}))$, $h_\sigma : \theta \mapsto (\theta,1,\rho_\nu(\theta))$. Notice that $\mathcal{F}\check{\gamma}_0 = \overline{\pi}_0$ and $\check{\mathcal{F}}\check{\gamma}_0 = \hat{\pi}_0$, while $\mathcal{F}\check{\sigma}_0 = \overline{\nu}_0$ and $\check{\mathcal{F}}\check{\sigma}_0 = \hat{\nu}_0$.

We use the same objects and notation as in Remark 13 and introduce the extra variables $\beta_{ij} = \rho_\nu(\vartheta_{ij})$ and $\varrho_{ij} = \rho_\pi(\tilde{Z}_{ij}, Z_{ij})$; notice that the uniform boundedness on ρ_ν and ρ_π is imposed to guarantee that the weights ϱ_{ij} and β_{ij} are uniformly bounded. Consider the measures

$$\check{\gamma}_0^{n,m} := \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{(Z_{ij},\tilde{Z}_{ij},\omega_{ij},\varrho_{ij})}, \quad \check{\sigma}_0^{n,m} := \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{(\vartheta_{ij},\alpha_{ij},\beta_{ij})}.$$

From Lemma 59 we can find a sequence $\{(n_k,m_k)\}_{k\in\mathbb{N}}$ such that, almost surely, the induced sequence of pairs $\check{\gamma}_0^{N_k} := \check{\gamma}_0^{n_k,m_k}$, $\check{\sigma}_0^{N_k} := \check{\sigma}_0^{n_k,m_k}$ satisfies conditions (4.12). Moreover, thanks to the law of large numbers and Lemma 61 in Appendix A this subsequence can be assumed to be such that

$$\lim_{k\to\infty}\frac{1}{n_k}\sum_{i=1}^{n_k}\left|\frac{1}{\frac{1}{m_k}\sum_{j=1}^{m_k}\rho_\pi(Z_{ij},\tilde{Z}_{ij})}-1\right| = 0, \quad \lim_{k\to\infty}\frac{1}{n_k}\sum_{i=1}^{n_k}\left|\frac{1}{\frac{1}{m_k}\sum_{j=1}^{m_k}\rho_\nu(\vartheta_{ij})}-1\right| = 0. \tag{4.19}$$

We make a slight modification to the weights ϱ_{ij} and β_{ij}, normalizing them so that $\frac{1}{m_k}\sum_j \varrho_{ij} = 1$ for all i, as well as $\frac{1}{n_k m_k}\sum_{ij}\beta_{ij} = 1$. From (4.19) we can directly show that condition (4.12) continues to hold after the normalization of weights. The resulting measures $\hat{\pi}_0^{N_k} = \sum_j \varrho_{ij}\delta_{(Z_{ij},\tilde{Z}_{ij})}$ and $\hat{\nu}_0^{N_k} = \sum_{ij}\beta_{ij}\delta_{\vartheta_{ij}}$ can be seen to converge, in the Wasserstein sense, respectively, toward $\hat{\nu}_0$ and $\hat{\pi}_0$, while the measures $\pi_0^{N_k} = \frac{1}{n_k m_k}\sum_{ij}\delta_{(Z_{ij},\tilde{Z}_{ij})}$ and $\nu_0^{N_k} = \frac{1}{n_k m_k}\sum_{ij}\delta_{\vartheta_{ij}}$ converge toward $\overline{\pi}_0$ and $\overline{\nu}_0$, respectively.

Moreover, another application of the law of large numbers implies that

$$\mathcal{H}(\hat{\nu}_0^{N_k}||\nu_0^{N_k}) = \Big(\frac{1}{n_k m_k}\sum_{ij}\rho_\nu(\vartheta_{ij})\Big)^{-1}\frac{1}{n_k m_k}\sum_{ij}\log(\rho_\nu(\vartheta_{ij}))\rho_\nu(\vartheta_{ij}) - \log\left(\frac{1}{n_k m_k}\sum_{ij}\rho_\nu(\vartheta_{ij})\right)$$

converges, as $k\to\infty$, toward $\int_\Theta \log(\rho_\nu(\theta))\rho_\nu(\theta)d\overline{\nu}_0(\theta)$, which is precisely $\mathcal{H}(\hat{\nu}_0||\overline{\nu}_0)$. Likewise, we can see that $\mathcal{H}(\hat{\pi}_0^{N_k}||\pi_0^{N_k}) \to \mathcal{H}(\hat{\pi}_0||\overline{\pi}_0)$, as $k\to\infty$.

The above convergence of relative entropies will be used in the next section.

5. Long term behavior of mean-field equation and approximate Nash equilibria of (1.1)

In this section we study the long time behavior of the system of equations (2.3) appropriately initialized at some measures (π_0,ν_0). Our aim is to study the ability of system (2.3) (or slight modifications thereof) to generate approximate Nash equilibria for problem (1.1).

We start by imposing additional convexity-concavity assumptions on $\mathcal{U}$, where convexity-concavity must be interpreted in the linear interpolation sense.

Assumption 32 *We assume that $\mathcal{U}$ is convex in ν and concave in π in the linear interpolation sense. That is,*

$$\mathcal{U}(\tau\pi + (1-\tau)\hat{\pi}, \nu) \geq \tau\mathcal{U}(\pi,\nu) + (1-\tau)\mathcal{U}(\hat{\pi},\nu)$$

and

$$\mathcal{U}(\pi, \tau\nu + (1-\tau)\hat{\nu}) \leq \tau\mathcal{U}(\pi,\nu) + (1-\tau)\mathcal{U}(\pi,\hat{\nu}),$$

for all $\tau \in [0,1]$ and all probability measures $\pi, \hat{\pi} \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z})$, and $\nu, \hat{\nu} \in \mathcal{P}(\Theta)$.

Assuming that $\mathcal{U}$ has the form (1.3), the above conditions are equivalent to analogous convexity-concavity assumptions on $\mathcal{R}(\pi,\nu)$, given that $\mathcal{C}$ is linear in π.

Remark 33 *The fact that $\mathcal{U}$ is convex-concave according to linear interpolation (i.e., as introduced in Assumptions 32) does not imply that $\mathcal{U}$ is geodesically convex-concave with respect to the geometry that induces the dynamics (2.3) (see section 3.1 for a discussion on the geometric interpretation of equations (2.3)), so that convergence to a global Nash equilibrium or an approximate Nash equilibrium is not immediate. Due to this, despite Assumptions 32, without any further assumptions we will think of problem (1.1) as non-convex non-concave. We contrast this setting with the one in section 5.1, which we will refer to as the non-convex concave setting.*

As expected, the long-term convergence of the mean-field PDE to an equilibrium point is associated with the convex-concave nature of $\mathcal{U}$. It is worth noting that both the ascending and descending parts of the PDE dynamics in (3.2) can be broken down into two components: a transport term and a mass-transfer term. Intuitively, the linear interpolation type of convexity-concavity aligns with the mass-transfer term but not the transport term. Consequently, convergence requires dynamics primarily dominated by the mass-transfer term, as demonstrated in Theorem 35

In contrast, the non-convex-concave setting detailed in section 5.1 introduces a form of concavity that is compatible with the transfer term. Therefore, in this scenario, convergence imposes dynamics dominated by the transport term for the adversary*, as exhibited in Theorem 42.*

Example 34 *In the context of the motivating example in subsection 1.1, we see that Assumption 32 is satisfied provided that the loss function ℓ is a convex function in its first coordinate. This is certainly the case for both the squared-loss and the logistic loss.*

We separate our discussion into two distinctive cases: 1) a rather general non-convex non-concave setting, and 2) a non-convex concave setting. Recall that by non-convex/non-concave here we mean not *geodesically* convex/concave relative to the optimal transport geometry driving the dynamics, while we do assume convexity/concavity in the linear interpolation sense as in Assumption 32.

Let us start by stating the result in the non-convex non-concave setting.

Theorem 35 (Long-time behavior mean-field PDE) *Let $\epsilon > 0$. Suppose that Assumptions 8, 9, and 32 hold. Assume that ν_0 and π_0 are probability measures (with $\pi_{0,z} = \mu$) such that ν_0*

and $\pi_0(\cdot|z)$ are absolutely continuous with respect to Lebesgue measure (in each corresponding space) and their densities are lowered-bounded by some $k>0$: i.e., there exists $k>0$ for which $\frac{d\nu_0}{d\theta}(\theta)>k$, and $\frac{d\pi_0}{d\tilde{z}}(\tilde{z}|z)>k$ for all z in the support of μ.

Finally, assume that the learning rate η satisfies $\eta\in C^0([0,\infty))$ and is such that

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t\int_0^s \eta_\tau d\tau ds=\bar{\eta} \tag{5.1}$$

for $\bar{\eta}$ satisfying

$$4(L+M)^2\bar{\eta}<\epsilon.$$

Then there exists T^* such that for all $t>T^*$

$$\sup_{\pi^*\in\mathcal{P}(\mathcal{Z}^2)\ s.t\ \pi^*_z=\mu}\mathcal{U}(\pi^*,\bar{\nu}_t)-\inf_{\nu^*\in\mathcal{P}(\Theta)}\mathcal{U}(\bar{\pi}_t,\nu^*)\leq\epsilon,$$

where $\overline{\pi}_t:=\frac{1}{t}\int_0^t\pi_s ds$ and $\overline{\nu}_t:=\frac{1}{t}\int_0^t\nu_s ds$, and (π_t,ν_t) solve (2.3), when initialized at π_0,ν_0 as above.

As it turns out, we can prove a very similar result if ν_0 and $\pi_0(\cdot|z)$ are assumed to be empirical measures that are well spread out and have a sufficiently large number of support points.

Theorem 36 *Let $\epsilon>0$. Suppose that Assumptions 8, 9, and 32 hold. Assume that ν_0 and π_0 take the form*

$$\nu_0=\nu_0^M=\frac{1}{M}\sum_{i=1}^M\delta_{\theta_i};\quad \pi_0=\pi_0^N=\mu\otimes(\frac{1}{N}\sum_{j=1}^N\delta_{\tilde{z}_j}),$$

where $\theta_1,\dots,\theta_M$ and $\tilde{z}_1,\dots,\tilde{z}_N$ are i.i.d. samples from the uniform distributions over Θ and $\mathcal{Z}$, respectively. Assume, also, that M and N are large enough so that

$$C_\Theta\frac{\log(M)^{p_d}}{M^{1/d}}+C_{\mathcal{Z}}\frac{\log(N)^{p_{d'}}}{N^{1/d'}}\leq\epsilon$$

for suitable constants C_Θ and $C_{\mathcal{Z}}$ and a power p_d that takes the form $p_d=3/4$ if $d=2$ and $p_d=1/d$ if $d\geq 3$. Finally, assume that the learning rate η satisfies the same assumptions as in Theorem 35.

Then, with probability at least $1-\frac{1}{M^2}-\frac{1}{N^2}$ (on the samples $\theta_1,\dots,\theta_M$ and $\tilde{z}_1,\dots,\tilde{z}_N$), there exists T^* such that for all $t>T^*$

$$\sup_{\pi^*\in\mathcal{P}(\mathcal{Z}^2)\ s.t\ \pi^*_z=\mu}\mathcal{U}(\pi^*,\bar{\nu}_t)-\inf_{\nu^*\in\mathcal{P}(\Theta)}\mathcal{U}(\bar{\pi}_t,\nu^*)\leq 2\epsilon,$$

where $\overline{\pi}_t:=\frac{1}{t}\int_0^t\pi_s ds$ and $\overline{\nu}_t:=\frac{1}{t}\int_0^t\nu_s ds$, and (π_t,ν_t) solve (2.3), when initialized at π_0,ν_0 as above.

Remark 37 *The assumptions on the initializations π_0 and ν_0 in Theorems 35 and 36 effectively suggest that the particles in Algorithm 1 need to be well spread out throughout the*

domains at time zero. This is certainly a strong assumption, but it is not unlike other theoretical assumptions in the literature studying, mathematically, the training process of neural networks; see [13, 17, 51, 52, 53]. In the next section we discuss how the strong assumption on π_0 can be removed when one restricts the adversarial budget in the setting described in section 1.1.

Let us emphasize that Theorem 36 implies that the convergence toward approximate Nash equilibria also holds for dynamics induced by a finite particle system, provided that the particles are well spread out at initialization and the number of particles is sufficiently large.

Remark 38 *The assumption on η is easily satisfied and essentially imposes a decay rate. For instance, given $\lambda > 0$, we have that $a\exp(-\lambda t)$ and $a(t+1)^{-(1+\lambda)}$ satisfy (5.1).*

Let us highlight that the statement does not impose restrictions on the parameter κ. It is possible, via a change of time, to lower the requirements in the upper bound on $\bar{\eta}$ by instead adding lower bounds for the parameter κ that grow as ϵ decreases. This is analogous to treatments in other contexts as in [17]. Either way, the crucial point is that the mass transfer term should clearly dominate the dynamics. This is consistent with the intuition on the effects of the assumed linear convexity-concavity as highlighted in Remark 33. Note in passing that the situation when stronger concavity is assumed as presented in Theorems 42 and 43 is not the same (see Remark 44).

5.1. *The non-convex and strongly concave case*

In contrast to Theorem 35, the results in this subsection hold under no assumptions on the initialization π_0 but at the expense of additional assumptions on the payoff function $\mathcal{U}$ and a slight modification of the dynamics (2.3). These additional assumptions on $\mathcal{U}$ are not unnatural. For instance, in the motivating example from subsection 1.1, they are linked to the strength given to the adversarial cost function $\mathcal{C}$.

Assumption 39 *We assume the following uniform PL (Polyak-Lojasiewicz) condition on the functions $\mathcal{U}(\cdot,\nu)$: There exists $\lambda > 0$ such that for all $\nu \in \mathcal{P}(\Theta)$ and all $\pi \in \mathcal{P}(\mathcal{Z}^2)$ with $\pi_z = \mu$ we have*

$$\int |\nabla_{\tilde{z}} \mathcal{U}_\pi(\pi,\nu;z,\tilde{z})|^2 d\pi(z,\tilde{z}) \geq \lambda(m_\nu^* - \mathcal{U}(\pi,\nu)),$$

where $m_\nu^* := \sup_{\tilde{\pi}\ s.t.\ \tilde{\pi}_z = \mu} \mathcal{U}(\tilde{\pi},\nu)$.

Remark 40 *For simplicity, we will refer to the setting when Assumption 39 holds as the strongly concave setting, as it is often the case that one can deduce the PL condition from strong (geodesic) concavity; see Proposition 57 in Appendix A.2.*

Example 41 *Suppose that the payoff function $\mathcal{U}$ has the form (1.3) for $\mathcal{R}$ and $\mathcal{C}$ as in (1.4) and (1.5), respectively. As we show in Proposition 57 in Appendix A.2, if the set $\mathcal{Z}$ is convex (a reasonable assumption in applications), the activation and loss functions are twice continuously differentiable, and, importantly, the parameter c_a is large enough, then Assumption (39) is satisfied.*

To exploit the additional concavity on $\mathcal{U}(\cdot,\nu)$, it will be useful to consider a slight variation of (2.3) where we slow down time in the descent equation and where we remove the scaling factor η in the equation for π_t. Precisely, given $K \geq 1$ we consider the system

$$\begin{cases} \partial_t \nu_t &= \frac{\eta_t}{K} \mathrm{div}_\theta(\nu_t \nabla_\theta \mathcal{U}_\nu(\pi_t, \nu_t; \theta)) - \frac{\kappa}{K} \nu_t \left(\mathcal{U}_\nu(\pi_t, \nu_t; \theta)) - \int \mathcal{U}_\nu(\pi_t, \nu_t; \theta') d\nu_t(\theta') \right) \\ \partial_t \pi_t &= -\mathrm{div}_{z,\tilde{z}}(\pi_t(0, \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}))) + \kappa \pi_t \left(\mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}) - \int \mathcal{U}_\pi(\pi_t, \nu_t; z, \tilde{z}') d\pi_t(\tilde{z}'|z) \right), \end{cases} \tag{5.2}$$

initialized at an arbitrary $\pi_0 \in \mathcal{P}(\mathcal{Z}^2)$ with $\pi_{0,z} = \mu$ and at some ν_0. Well-posedness for this equation under Assumptions 8 and 9 can be established as for equation (2.3); we omit the details. To reflect the variations introduced in (5.2) in our Algorithm 1, it suffices to remove the η in the update for the variables $\tilde{z}_{ij}$ and to allow for the for loop over i, j to be repeated a number of times (quantity that can be tuned) before entering the loop over k.

We prove the following result.

Theorem 42 *Suppose Assumptions 8, 9, 32, and 39 hold. Assume further that there exists $k > 0$ such that $\frac{d\nu_0}{d\theta} > k$, and let π_0 be an arbitrary probability measure with $\pi_{0,z} = \mu$. Finally, assume that*

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t \int_0^s \eta_\tau d\tau ds = \bar{\eta} < \infty.$$

Fix $\epsilon > 0$. Then there exists $K_0, r_0, r_1, t_0 > 0$ such that, if $K \geq K_0$ and $\bar{\eta}/K \leq r_1$, then for all $t \geq \max\{t_0, K/r_0\}$, we have

$$\sup_{\tilde{\pi} \in \mathcal{P}(\mathcal{Z}^2) \ s.t. \ \tilde{\pi}_z = \mu} \mathcal{U}(\tilde{\pi}, \bar{\nu}_t) - \inf_{\tilde{\nu} \in \mathcal{P}(\Theta)} \mathcal{U}(\bar{\pi}_t, \tilde{\nu}) \leq \epsilon.$$

In the above, $\bar{\pi}_t := \frac{1}{t} \int_0^t \pi_s ds$ and $\bar{\nu}_t := \frac{1}{t} \int_0^t \nu_s ds$, and (π_t, ν_t) solve (5.2) initialized at ν_0, π_0 as above.

Just like Theorem 35, Theorem 42 has a version where ν_0 is only assumed to be an empirical measure that has a support that is well spread out.

Theorem 43 *Let $\epsilon > 0$. Suppose Assumptions 8, 9, 32, and 39 hold. Let π_0 be an arbitrary probability measure with $\pi_{0,z} = \mu$ and assume that ν_0 takes the form*

$$\nu_0 = \nu_0^M = \frac{1}{M} \sum_{i=1}^M \delta_{\theta_i},$$

where $\theta_1, \ldots, \theta_M$ are i.i.d. samples from the uniform distribution over Θ. Assume, also, that M is large enough so that

$$C_\Theta \frac{\log(M)^{p_d}}{M^{1/d}} \leq \epsilon$$

for a suitable constant C_Θ and a power p_d that takes the form $p_d = 3/4$ if $d = 2$ and $p_d = 1/d$ if $d \geq 3$. Finally, assume that the learning rate η satisfies the same assumptions as in Theorem 42.

There exists $K_0, r_0, r_1, t_0 > 0$ such that, if $K \geq K_0$ and $\overline{\eta}/K \leq r_1$, then, with probability at least $1 - \frac{1}{M^2}$ (on the samples $\theta_1, \dots, \theta_M$), for all $t \geq \max\{t_0, K/r_0\}$ we have

$$\sup_{\tilde{\pi} \in \mathcal{P}(\mathcal{Z}^2) \text{ s.t. } \tilde{\pi}_z = \mu} \mathcal{U}(\tilde{\pi}, \bar{\nu}_t) - \inf_{\tilde{\nu} \in \mathcal{P}(\Theta)} \mathcal{U}(\bar{\pi}_t, \tilde{\nu}) \leq 2\epsilon.$$

In the above, $\overline{\pi}_t := \frac{1}{t}\int_0^t \pi_s ds$ and $\overline{\nu}_t := \frac{1}{t}\int_0^t \nu_s ds$, and (π_t, ν_t) solve (5.2) initialized at ν_0, π_0 as above.

Remark 44 *As announced, the additional concavity assumptions bring important benefits to the algorithm, since there is no assumptions that we must impose on π_0 in either of the previous theorems. Moreover, the parameter K allows us to avoid a smallness condition on $\bar{\eta}$ as long as K and the search time are long enough.*

5.2. *Proofs of Theorem 35 and 36*

To begin our analysis, we first discuss the relationship between the system (2.3) and an associated "hat" process as in Lemma 28. The study of similar systems has been considered in works like [17]. However, here we present an alternative approach that allows us to fully justify our derivations; see Remark 48 below for more details. Our approach makes use of the larger structure that we studied in section 4. Indeed, we use the particle approximation in Remark 31 to understand the time evolution of the relative entropy between $\hat{\nu}$ and ν, and $\hat{\pi}$ and π, for arbitrary initializations. As a first step, we study the time evolutions of relative entropies when the measures $(\nu_t^{N_k}, \pi_t^{N_k})$ and $(\hat{\nu}_t^{N_k}, \hat{\pi}_t^{N_k})$ are initialized at empirical measures as in Remark 31.

Proposition 45 *Let $\overline{\pi}_0$ and $\overline{\nu}_0$ be arbitrary, and let $\hat{\pi}_0$ and $\hat{\nu}_0$ be as in Remark 31. For a fixed $k \in \mathbb{N}$, let $\nu_t^{N_k}, \hat{\nu}_t^{N_k}, \pi_t^{N_k}, \hat{\pi}_t^{N_k}$ be as in Proposition 26 when initialized as in Remark 31.*

Then

$$\frac{d}{dt}\mathcal{H}(\hat{\nu}_t^{N_k} \| \nu_t^{N_k}) = \kappa \int_\Theta \mathcal{U}_\nu(\pi_t^{N_k}, \nu_t^{N_k}; \theta) d(\hat{\nu}_t^{N_k} - \nu_t^{N_k}) \tag{5.3}$$

and

$$\frac{d}{dt}\mathcal{H}(\hat{\pi}_t^{N_k} \| \pi_t^{N_k}) = -\kappa \int_{\mathcal{Z}\times\mathcal{Z}} \mathcal{U}_\pi(\pi_t^{N_k}, \nu_t^{N_k}; z, \tilde{z}) d(\hat{\pi}_t^{N_k} - \pi_t^{N_k}).$$

Proof Notice that

$$\begin{aligned}
\frac{d}{dt}\mathcal{H}(\hat{\nu}_t^{N_k} \| \nu_t^{N_k}) &= \frac{d}{dt}\left(\frac{1}{n_k m_k} \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} \log\left(\frac{\beta_{ij}(0)\alpha_{ij}(0)}{\alpha_{ij}(t)}\right) \beta_{ij}(0)\alpha_{ij}(0) \right) \\
&= -\frac{1}{n_k m_k} \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} \frac{d}{dt} \log(\alpha_{ij}(t)) \beta_{ij}(0)\alpha_{ij}(0) \\
&= \frac{\kappa}{n_k m_k} \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} (\mathcal{U}_\nu(\pi_t^{N_k}, \nu_t^{N_k}; \vartheta_{ij}(t)) - \overline{\mathcal{U}}_\nu) \beta_{ij}(0)\alpha_{ij}(0),
\end{aligned}$$

where to go from the second to the third line we have used equation (4.1) for $\alpha_{ij}(t)$. We have also used the shorthand notation $\overline{\mathcal{U}}_\nu = \int_\Theta \mathcal{U}_\nu(\pi_t^{N_k}, \nu_t^{N_k}; \theta) d\nu_t^{N_k}(\theta)$. Identity (5.3) now follows.

The identity for $\frac{d}{dt}\mathcal{H}(\hat{\pi}_t^N \| \pi_t^N)$ follows from similar considerations, but now we rely on the fact that the weights ϱ_{ij} are normalized along every row:

$$\begin{aligned}
\frac{d}{dt}\mathcal{H}(\hat{\pi}_t^{N_k} \| \pi_t^{N_k}) &= \frac{d}{dt}\left(\frac{1}{n_k m_k} \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} \log\left(\frac{\varrho_{ij}(0)\omega_{ij}(0)}{\omega_{ij}(t)} \right) \varrho_{ij}(0)\omega_{ij}(0) \right) \\
&= -\frac{1}{n_k m_k} \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} \frac{d}{dt} \log(\omega_{ij}(t)) \varrho_{ij}(0)\omega_{ij}(0) \\
&= -\frac{\kappa}{n_k m_k} \sum_{i=1}^{n_k}\sum_{j=1}^{m_k} (\mathcal{U}_\pi(\pi_t^{N_k}, \nu_t^{N_k}; Z_{ij}, \tilde{Z}_{ij}) - \overline{\mathcal{U}}_{\pi,i}) \varrho_{ij}(0)\omega_{ij}(0).
\end{aligned}$$

In the above we have used the shorthand notation $\overline{\mathcal{U}}_{\pi,i} = \int_{\mathcal{Z}\times\mathcal{Z}} \mathcal{U}_\pi(\pi_t^{N_k}, \nu_t^{N_k}; Z_{ij}, \tilde{z}) d\pi_t^{N_k}(\tilde{z}|Z_{ij})$; recall that in our construction Z_{ij} does not depend on j. □

Next, we add one ingredient to the approximation result from Corollary 30 in search of a relationship similar to (45) but for general initializations.

Proposition 46 *Let $\overline{\pi}_0$ and $\overline{\nu}_0$ be arbitrary, and let $\hat{\pi}_0$ and $\hat{\nu}_0$ be as in Remark 31. Let $(\hat{\nu}, \hat{\pi})$ be the dynamics in Lemma 28 when initialized as in Remark 31. For every $k \in \mathbb{N}$, let $\nu_t^{N_k}, \hat{\nu}_t^{N_k}, \pi_t^{N_k}, \hat{\pi}_t^{N_k}$ be as in Proposition 26 when initialized as in Remark 31.*

Then

$$\lim_{k\to\infty} \int \mathcal{U}_\nu(\pi_s^{N_k}, \nu_s^{N_k}; \theta) d(\hat{\nu}_s^{N_k} - \nu_s^{N_k}) = \int \mathcal{U}_\nu(\pi_s, \nu_s; \theta) d(\hat{\nu}_s - \nu_s) \tag{5.4}$$

as well as

$$\lim_{k\to\infty} \int_{\mathcal{Z}\times\mathcal{Z}} \mathcal{U}_\pi(\pi_s^{N_k}, \nu_s^{N_k}; z, \tilde{z}) d(\hat{\pi}_s^{N_k} - \pi_s^{N_k}) = -\int_{\mathcal{Z}\times\mathcal{Z}} \mathcal{U}_\pi(\pi_s, \nu_s; z, \tilde{z}) d(\hat{\pi}_s - \pi_s). \tag{5.5}$$

Proof From Assumptions 8 and Corollary 30 we have

$$\left| \int \mathcal{U}_\nu(\pi_s^{N_k}, \nu_s^{N_k}; \theta) d(\hat{\nu}_s^{N_k} - \nu_s^{N_k}) - \int \mathcal{U}_\nu(\pi_s, \nu_s; \theta) d(\hat{\nu}_s^{N_k} - \nu_s^{N_k}) \right| \leq L(W_1(\nu_s, \nu_s^{N_k}) + W_1(\pi_s, \pi_s^{N_k})) \to 0,$$

as $k \to \infty$. On the other hand, since the function $\mathcal{U}_\nu(\pi_s, \nu_s, \cdot)$ is continuous with at most linear growth in $|\theta|$, and since $W_1(\nu_s^{N_k}, \nu_s) \to 0$, $W_1(\hat{\nu}_s^{N_k}, \hat{\nu}_s) \to 0$ as $k \to \infty$ by Corollary 30, it follows that

$$\lim_{k\to\infty} \left| \int \mathcal{U}_\nu(\pi_s, \nu_s; \theta) d(\hat{\nu}_s^{N_k} - \nu_s^{N_k}) - \int \mathcal{U}_\nu(\pi_s, \nu_s; \theta) d(\hat{\nu}_s - \nu_s) \right| = 0.$$

Equation (5.4) readily follows. (5.5) is obtained similarly.

□

Proposition 47 *Let $\overline{\pi}_0$ and $\overline{\nu}_0$ be arbitrary, and let $\hat{\pi}_0$ and $\hat{\nu}_0$ be as in Remark 31. Let $(\hat{\nu}, \hat{\pi})$ be the dynamics in Lemma 28 when initialized as in Remark 31.*

Then the following inequalities hold:

$$\mathcal{H}(\hat{\nu}_t||\nu_t) - \mathcal{H}(\hat{\nu}_0||\overline{\nu}_0) \leq \kappa \int_0^t \left(\int \mathcal{U}_\nu(\pi_s, \nu_s; \theta) d(\hat{\nu}_s - \nu_s)(\theta) \right) ds, \quad \forall t \geq 0, \tag{5.6}$$

and

$$\mathcal{H}(\hat{\pi}_t||\pi_t) - \mathcal{H}(\hat{\pi}_0||\overline{\pi}_0) \leq -\kappa \int_0^t \left(\int \mathcal{U}_\pi(\pi_s, \nu_s; z, \tilde{z}) d(\hat{\pi}_s - \pi_s)(z, \tilde{z}) \right) ds, \quad \forall t \geq 0. \tag{5.7}$$

Proof For every $k \in \mathbb{N}$, consider $\nu_t^{N_k}, \hat{\nu}_t^{N_k}, \pi_t^{N_k}, \hat{\pi}_t^{N_k}$ be as in Proposition 26 when initialized as in Remark 31. Notice that thanks to Corollary 30 we have $W_1(\nu_s^{N_k}, \nu_s) \to 0$, $W_1(\hat{\nu}_s^{N_k}, \hat{\nu}_s) \to 0$, as $k \to \infty$.

From Proposition (45) we have

$$\mathcal{H}(\hat{\nu}_t^{N_k} \| \nu_t^{N_k}) = \mathcal{H}(\hat{\nu}_0^{N_k} \| \nu_0^{N_k}) + \int_0^t \int_\Theta \mathcal{U}_\nu(\pi_s^{N_k}, \nu_s^{N_k}; \theta) d(\hat{\nu}_s^{N_k} - \nu_s^{N_k}) ds.$$

We may now use the joint lower semi-continuity of the relative entropy w.r.t weak convergence to obtain:

$$\begin{aligned} \mathcal{H}(\hat{\nu}_t||\nu_t) \leq \liminf_{k\to\infty} \mathcal{H}(\hat{\nu}_t^{N_k}||\nu_t^{N_k}) &= \liminf_{k\to\infty} \kappa \int_0^t \left(\int \mathcal{U}_\nu(\pi_s^{N_k}, \nu_s^{N_k}; \theta) d(\hat{\nu}_s^{N_k} - \nu_s^{N_k})(\theta) \right) ds \\ &\quad + \lim_{k\to\infty} \mathcal{H}(\hat{\nu}_0^{N_k}||\nu_0^{N_k}). \\ &= \liminf_{k\to\infty} \kappa \int_0^t \left(\int \mathcal{U}_\nu(\pi_s^{N_k}, \nu_s^{N_k}; \theta) d(\hat{\nu}_s^{N_k} - \nu_s^{N_k})(\theta) \right) ds \\ &\quad + \mathcal{H}(\hat{\nu}_0||\overline{\nu}_0). \end{aligned} \tag{5.8}$$

Using Proposition (46) and the approximation properties discussed in Remark 31 we obtain (5.6). Inequality (5.7) is obtained similarly. □

Remark 48 *In contrast to the analysis presented in [17], here we have used our mean-field limit results from section 4 and the lower semi continuity properties of the relative entropy to fully justify the one-sided identities (5.6) and (5.7). As we will see below, these one-sided identities are sufficient for our analysis. Following our approach, we can sidestep the strategy considered in [17] for analyzing a similar problem. Their strategy relies on the assumption of existence and regularity of solutions to a certain PDE describing the evolution of the change of measure between processes similar to the ν and $\hat{\nu}$ considered here. Unfortunately, such PDE is not even well-defined in general, as it becomes apparent when one considers flows initialized at empirical measures. While this technical difficulty is acknowledged in [17], no solution for it is provided; see Page 29 in [17].*

With Proposition (47) in hand, and following similar steps as in [17], we can now derive results controlling exploitability under Assumptions 8 and 32.

Lemma 49 *Let π,ν be the solution of equation (2.3) initialized at probability measures π_0,ν_0 with $\pi_{0,z}=\mu$. Let π^*,ν^* be arbitrary probability measures over $\mathcal{Z}\times\mathcal{Z}$ and Θ, respectively, and suppose that $\pi^*_z=\mu$. Let*

$$\mathcal{Q}_\pi(\pi_0,\pi^*;\tau):=\inf_{\hat\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z}),\,\hat\pi_z=\mu}\{\|\pi^*-\hat\pi\|^*_{BL}+\frac{1}{\tau}\mathcal{H}(\hat\pi||\pi_0)\},$$

where $\|\cdot\|^*_{BL}$ denotes the dual of the BL (Bounded Lipschitz) norm $\|\cdot\|_{BL}:=\|\cdot\|_\infty+\mathrm{Lip}(\cdot)$. Consider also $\mathcal{Q}_\nu(\nu_0,\nu^*;\tau)$ defined as

$$\mathcal{Q}_\nu(\nu_0,\nu^*;\tau):=\inf_{\hat\nu\in\mathcal{P}(\Theta)}\{\|\nu^*-\hat\nu\|^*_{BL}+\frac{1}{\tau}\mathcal{H}(\hat\nu||\nu_0)\}.$$

Suppose that Assumptions 8 and 32 hold. Then

$$\mathcal{U}(\pi^*,\bar\nu(t))-\mathcal{U}(\bar\pi(t),\nu^*)\le B(\mathcal{Q}_\pi(\pi_0,\pi^*;\kappa Bt)+\mathcal{Q}_\nu(\nu_0,\nu^*;\kappa Bt))+\frac{2B^2}{t}\int_0^t\int_0^s\eta_\tau d\tau ds,$$

where $B:=M+L$ (see Assumption 8 for the meaning of L and M). In the above, $\overline{\pi}_t:=\frac{1}{t}\int_0^t\pi_s ds$ and $\overline{\nu}_t:=\frac{1}{t}\int_0^t\nu_s ds$.

Proof Consider two arbitrary probability measures $\hat\pi_0$ and $\hat\nu_0$ with $\hat\pi_0\ll\pi_0,\hat\nu_0\ll\nu_0$, $\hat\pi_{0,z}=\mu$, $\frac{d\hat\nu_0}{d\nu_0}\in L^\infty(\nu_0)$, and $\frac{d\hat\pi_0}{d\pi_0}\in L^\infty(\pi_0)$. We consider the dynamics $(\hat\pi_t,\hat\nu_t)$ and (π_t,ν_t) as in Proposition 47.

1. Step 1: From the concavity of $\mathcal{U}$ in its first coordinate (with respect to linear interpolation) it follows that

$$\begin{aligned}\mathcal{U}(\pi^*,\nu_t)\le&\mathcal{U}(\pi_t,\nu_t)+\int\mathcal{U}_\pi(\pi_t,\nu_t;z,\tilde z)d(\pi^*-\pi_t)\\=&\mathcal{U}(\pi_t,\nu_t)+\int\mathcal{U}_\pi(\pi_t,\nu_t;z,\tilde z)d(\pi^*-\hat\pi_t)\\&+\int\mathcal{U}_\pi(\pi_t,\nu_t;z,\tilde z)d(\hat\pi_t-\pi_t).\end{aligned}$$

Using the BL (bounded Lipschitz) norm, we get from Proposition 47 that

$$\begin{aligned}\int_0^t\mathcal{U}(\pi^*,\nu_s)ds-\int_0^t\mathcal{U}(\pi_s,\nu_s)ds\le&\int_0^t(\|\mathcal{U}_\pi(\pi_s,\nu_s;\cdot)\|_{BL}\|\pi^*-\hat\pi_s\|^*_{BL})ds\\&+\int_0^t\int\mathcal{U}_\pi(\pi_s,\nu_s;z,\tilde z)d(\hat\pi_s-\pi_s)ds\\\le&\int_0^t(\|\mathcal{U}_\pi(\pi_s,\nu_s;\cdot)\|_{BL}\|\pi^*-\hat\pi_s\|^*_{BL})ds\\&-\frac{1}{\kappa}(\mathcal{H}(\hat\pi_t||\pi_t)-\mathcal{H}(\hat\pi_0||\pi_0)).\end{aligned}\tag{5.9}$$

A similar argument using the convexity of $\mathcal{U}$ in its second coordinate deduces

$$\int_0^t \mathcal{U}(\pi_s,\nu^*)ds - \int_0^t \mathcal{U}(\pi_s,\nu_s)ds \geq -\int_0^t (|\mathcal{U}_\nu(\pi_s,\nu_s;\cdot)\|_{BL}\|\nu^* - \hat{\nu}_s\|^*_{BL})ds + \frac{1}{\kappa}(\mathcal{H}(\hat{\nu}_t||\nu_t) - \mathcal{H}(\hat{\nu}_0||\nu_0)). \tag{5.10}$$

Using again the concavity and convexity of $\mathcal{U}$, we get:

$$\mathcal{U}(\bar{\pi}_t,\nu^*) \geq \frac{1}{t}\int_0^t \mathcal{U}(\pi_s,\nu^*)ds, \qquad \mathcal{U}(\pi^*,\bar{\nu}_t) \leq \frac{1}{t}\int_0^t \mathcal{U}(\pi^*,\nu_s)ds.$$

Combining the above with (5.9), (5.10), and the fact that $\mathcal{H}(\hat{\nu}_t||\nu_t), \mathcal{H}(\hat{\pi}_t||\pi_t) \geq 0$ we conclude that

$$\begin{aligned}\mathcal{U}(\pi^*,\bar{\nu}_t) - \mathcal{U}(\bar{\pi}_t,\nu^*) \leq & \frac{1}{t}\int_0^t (\|\mathcal{U}_\nu(\pi_s,\nu_s;\cdot)\|_{BL}\|\nu^* - \hat{\nu}_s\|^*_{BL} + \|\mathcal{U}_\pi(\pi_s,\nu_s;\cdot)\|_{BL}\|\pi^* - \hat{\pi}_s\|^*_{BL})ds \\ & + \frac{1}{\kappa t}(\mathcal{H}(\hat{\nu}_0||\nu_0) + \mathcal{H}(\hat{\pi}_0||\pi_0)) \\ \leq & \frac{B}{t}\int_0^t (\|\nu^* - \hat{\nu}_s\|^*_{BL} + \|\pi^* - \hat{\pi}_s\|^*_{BL})ds + \frac{1}{\kappa t}(\mathcal{H}(\hat{\nu}_0||\nu_0) + \mathcal{H}(\hat{\pi}_0||\pi_0)).\end{aligned} \tag{5.11}$$

2. Step 2: Observe that both $\mathcal{U}_\pi$ and $\mathcal{U}_\nu$ have their BL norm bounded by $B = M + L$. To conclude, it remains to remark that

$$\begin{aligned}\frac{1}{t}\int_0^t \|\pi^* - \hat{\pi}_s\|^*_{BL}ds \leq & \|\pi^* - \hat{\pi}_0\|^*_{BL} + \frac{1}{t}\int_0^t \|\hat{\pi}_0 - \hat{\pi}_s\|^*_{BL}ds \\ = & \|\pi^* - \hat{\pi}_0\|^*_{BL} + \frac{1}{t}\int_0^t \left\{ \sup_{\|f\|_{BL}\leq 1; f\in C^1} \int f d(\hat{\pi}_s - \hat{\pi}_0) \right\} ds \\ \leq & \|\pi^* - \hat{\pi}_0\|^*_{BL} + \frac{B}{t}\int_0^t \int_0^s \eta_\tau d\tau ds,\end{aligned}$$

and similarly,

$$\frac{1}{t}\int_0^t \|\nu^* - \hat{\nu}_s\|^*_{BL}ds \leq \|\nu^* - \hat{\nu}_0\|^*_{BL} + \frac{B}{t}\int_0^t \int_0^s \eta_\tau d\tau ds.$$

Replacing in (5.11), it follows that

$$\begin{aligned}\mathcal{U}(\pi^*,\bar{\nu}_t) - \mathcal{U}(\bar{\pi}_t,\nu^*) \leq & B(\|\nu^* - \hat{\nu}_0\|^*_{BL} + \|\pi^* - \hat{\pi}_0\|^*_{BL}) \\ & + \frac{1}{\kappa t}(\mathcal{H}(\hat{\nu}_0||\nu_0) + \mathcal{H}(\hat{\pi}_0||\pi_0)) + \frac{2B^2}{t}\int_0^t \int_0^s \eta_\tau d\tau ds.\end{aligned} \tag{5.12}$$

Recall that $\hat{\pi}_0$ and $\hat{\nu}_0$ were arbitrary measures with densities with respect to π_0 and ν_0 belonging to L^∞. From a simple density argument we may now conclude the desired estimate.

□

The following Lemma is taken from [17] which in turn follows the arguments in [13].

Lemma 50 *Suppose that Assumptions 8 and 32 hold. Assume further that there exists $k>0$ such that $\frac{d\nu_0}{d\theta}(\theta)>k$, and suppose that $|\mathcal{B}_{\theta,\epsilon}\cap\Theta|\geq k'\epsilon^d$ uniformly in $\theta\in\Theta$, where $\mathcal{B}_{\theta,\epsilon}$ is the Euclidean ball of radius ϵ centered at θ. Then,*

$$\mathcal{Q}_\nu(\nu_0,\nu^*;\tau)\leq\frac{d}{\tau}\left\{1-\log\left(\frac{d}{\tau}\right)\right\}+\frac{1}{\tau}\{-\log(k)-\log(k')\}.$$

Proof We obtain a bound for the min in the definition of $\mathcal{Q}_\nu$. For a fixed $\epsilon>0$ we introduce a probability measure ν^ϵ given by

$$\nu^\epsilon(A):=\int_\Theta\frac{|\mathcal{B}_{\theta,\epsilon}\cap A|}{|\mathcal{B}_{\theta,\epsilon}\cap\Theta|}d\nu^*(\theta).$$

We now calculate $W_1(\nu^*,\nu^\epsilon)$. Consider the coupling:

$$\Upsilon(A,A'):=\int_A\frac{|\mathcal{B}_{\theta,\epsilon}\cap A'|}{|\mathcal{B}_{\theta,\epsilon}\cap\Theta|}d\nu^*(\theta).$$

Indeed, one easily verifies $\Upsilon(\Theta,A')=\nu^\epsilon(A')$ and $\Upsilon(A,\Theta)=\nu^*(A)$. Thus,

$$W_1(\nu^*,\nu^\epsilon)\leq\int_\Theta\int_\Theta|\theta-\theta'|d\zeta(\theta,\theta')=\int_\Theta\int_{\theta'\in\mathcal{B}_{\theta,\epsilon}\cap\Theta}|\mathcal{B}_{\theta,\epsilon}\cap\Theta|^{-1}|\theta-\theta'|d\theta'd\nu^*(\theta)\leq\epsilon.$$

Since for any measure $\nu\in\mathcal{P}(\Theta)$ we have that $\|\nu^*-\nu\|_{BL}^*\leq W_1(\nu^*,\nu)$, we obtain a bound of ϵ for the first term in $\mathcal{Q}_\nu$.

We now turn to the relative entropy term. Observe that the definition of ν^ϵ and Fubini's theorem gives

$$\frac{d\nu^\epsilon}{d\theta}(\theta)=\int_\Theta|\mathcal{B}_{\theta',\epsilon}\cap\Theta|^{-1}\mathbb{1}_{\mathcal{B}_{\theta',\epsilon}\cap\Theta}(\theta)d\nu^*(\theta');$$

thus, by convexity of the function $x\mapsto x\log(x)$, Jensen's inequality and Fubini's theorem, we have

$$\begin{aligned}\int_\Theta\frac{d\nu^\epsilon}{d\theta}(\theta)\log\left(\frac{d\nu^\epsilon}{d\theta}(\theta)\right)d\theta&\leq\int_\Theta\int_\Theta|\mathcal{B}_{\theta',\epsilon}\cap\Theta|^{-1}\mathbb{1}_{\mathcal{B}_{\theta',\epsilon}}(\theta)\log(|\mathcal{B}_{\theta',\epsilon}\cap\Theta|^{-1})d\nu^*(\theta')d\theta\\&\leq-\int_\Theta\log(|\mathcal{B}_{\theta',\epsilon}\cap\Theta|)d\nu^*(\theta')\leq-\log(k')-d\log(\epsilon),\end{aligned}\tag{5.13}$$

where we have used the convention $0\times-\infty=0$.

On the other hand, by assumption,

$$\int_{\Theta} \frac{d\nu^{\epsilon}}{d\theta}(\theta) \log\left(\frac{d\nu_0}{d\theta}(\theta)\right) d\theta = \int_{\Theta} \log\left(\frac{d\nu_0}{d\theta}(\theta)\right) d\nu^{\epsilon}(\theta) \geq \log(k). \tag{5.14}$$

From the above it follows

$$\mathcal{H}(\nu^{\epsilon}||\nu^0) \leq -\log(k) - \log(k') - d\log(\epsilon).$$

Hence,

$$\mathcal{Q}(\nu_0, \nu^*; \tau) \leq \epsilon - \frac{1}{\tau}(d\log(\epsilon) + \log(k') + \log(k)),$$

for every $\epsilon > 0$. Choosing $\epsilon = \frac{d}{\tau}$, the minimizer of the right hand side of the above expression, we get the desired result. □

Remark 51 *The condition $|\mathcal{B}_{\theta,\epsilon} \cap \Theta| \geq k'\epsilon^d$ uniformly over $\theta \in \Theta$ is implied by the fact that the boundary of Θ was assumed to be Lipschitz; see Assumptions 8.*

A posteriori, we can generalize Lemma 50 to allow for empirical measures that approximate in a suitable sense a measure ν_0 satisfying the assumptions in Lemma 50. This is the content of the next result.

Lemma 52 *Let $\theta_1, \ldots, \theta_M \in \Theta$ be M distinct points in Θ, and let $\nu_0^M := \frac{1}{M}\sum_{i=1}^M \delta_{\theta_i}$. Suppose that ν_0 and Θ are as in Lemma 50. Then, for every $\tau > 0$, we have*

$$\mathcal{Q}_{\nu}(\nu_0^M, \nu^*; \tau) \leq W_{\infty}(\nu_0, \nu_0^M) + \frac{d}{\tau}\left\{1 - \log\left(\frac{d}{\tau}\right)\right\} + \frac{1}{\tau}\{-\log(k) - \log(k')\},$$

where W_{∞} denotes the ∞-Wasserstein distance between ν_0 and ν_0^M.

In particular, if $\theta_1, \ldots, \theta_M$ are sampled independently from a ν_0 with a density with respect to the Lebesgue measure that is bounded and bounded away from zero, then, with probability at least $1 - 1/M^2$,

$$\mathcal{Q}_{\nu}(\nu_0^M, \nu^*; \tau) \leq C\frac{\log(M)^{p_d}}{M^{1/d}} + \frac{d}{\tau}\left\{1 - \log\left(\frac{d}{\tau}\right)\right\} + \frac{1}{\tau}\{-\log(k) - \log(k')\},$$

for a constant C that depends on ν_0, and a power p_d that takes the form $p_d = 3/4$ if $d = 2$ and $p_d = 1/d$ if $d \geq 3$.

Proof Fix $\tau > 0$ and let $\nu^* \in \mathcal{P}(\Theta)$ be an arbitrary probability measure. By Lemma 50 for any given $\varepsilon > 0$ we can find $\hat{\nu}_{\tau} \in \mathcal{P}(\Theta)$ such that

$$\|\nu^* - \hat{\nu}_{\tau}\|_{BL}^* + \frac{1}{\tau}\mathcal{H}(\hat{\nu}_{\tau}||\nu_0) \leq \varepsilon + \frac{d}{\tau}\left\{1 - \log\left(\frac{d}{\tau}\right)\right\} + \frac{1}{\tau}\{-\log(k) - \log(k')\}.$$

Let $T : \Theta \to \{\theta_1, \ldots, \theta_M\}$ be an ∞-OT map between ν_0 and ν_0^M, which exists thanks to the assumptions on ν_0 and the main result in [10]. In particular, T can be taken to satisfy

$T_\sharp \nu_0 = \nu_0^M$ and

$$\sup_{\theta\in\Theta} |\theta - T(\theta)| = W_\infty(\nu_0, \nu_0^M);$$

notice that we can indeed take an actual supremum on the left hand side of the above expression, and not just an essential supremum, thanks to the assumptions on ν_0 and the domain Θ. Having introduced the map T, we define the measure

$$\hat{\nu}_\tau^M := \sum_{i=1}^M \hat{\nu}_\tau(T^{-1}(\theta_i))\delta_{\theta_i},$$

which is an empirical version of $\hat{\nu}_\tau$. In what follows we bound $\|\nu^* - \hat{\nu}_\tau^M\|_{BL}^* + \frac{1}{\tau}\mathcal{H}(\hat{\nu}_\tau^M||\nu_0^M)$.

First,

$$\begin{aligned}
\|\nu^* - \hat{\nu}_\tau^M\|_{BL}^* &\le \|\nu^* - \hat{\nu}_\tau\|_{BL}^* + \|\hat{\nu}_\tau - \hat{\nu}_\tau^M\|_{BL}^* \\
&\le \|\nu^* - \hat{\nu}_\tau\|_{BL}^* + W_1(\hat{\nu}_\tau, \hat{\nu}_\tau^M) \\
&\le \|\nu^* - \hat{\nu}_\tau\|_{BL}^* + \int_\Theta |\theta - T(\theta)| d\hat{\nu}_\tau(\theta) \\
&\le \|\nu^* - \hat{\nu}_\tau\|_{BL}^* + \sup_{\theta\in\Theta} |\theta - T(\theta)| \\
&= \|\nu^* - \hat{\nu}_\tau\|_{BL}^* + W_\infty(\nu_0, \nu_0^M),
\end{aligned} \tag{5.15}$$

where in the second to last inequality we have used the fact that, as can be easily verified, $T_\sharp \hat{\nu}_\tau = \hat{\nu}_\tau^M$.

On the other hand, a straightforward application of Jensen's inequality reveals that

$$\mathcal{H}(\hat{\nu}_\tau^M||\nu_0^M) \le \mathcal{H}(\hat{\nu}_\tau||\nu_0).$$

Combining the above inequalities we conclude that for every $\varepsilon > 0$

$$\mathcal{Q}_\nu(\nu_0^M, \nu^*; \tau) - \varepsilon \le W_\infty(\nu_0, \nu_0^M) + \frac{d}{\tau}\left\{1 - \log\left(\frac{d}{\tau}\right)\right\} + \frac{1}{\tau}\{-\log(k) - \log(k')\},$$

which of course implies the desired bound.

When the points $\theta_1, \ldots, \theta_M$ are sampled from a distribution ν_0 satisfying the specified additional assumptions, Theorem 1.1 in [26] allows us to bound $W_\infty(\nu_0, \nu_0^M)$ by $C\frac{\log(M)^{pd}}{M^{1/d}}$ with very high probability. □

Lemma 53 *Suppose that Assumptions 8 and 32 hold. Let π_0 be such that $\pi_{0,z} = \mu$ and suppose that there exists $k > 0$ such that $\frac{d\pi_0(\tilde{z}|z)}{d\tilde{z}}(\tilde{z}) > k$ for all z in the support of μ. Suppose further that $|\mathcal{B}_{\tilde{z},\epsilon} \cap \mathcal{Z}| \ge k'\epsilon^{d'}$ uniformly in $\tilde{z} \in \mathcal{Z}$. Then, for all π^* with $\pi_z^* = \mu$, we have*

$$\mathcal{Q}_\pi(\pi_0, \pi^*; \tau) \le \frac{d'}{\tau}\left\{1 - \log\left(\frac{d'}{\tau}\right)\right\} + \frac{1}{\tau}\{-\log(k) - \log(k')\}.$$

Proof Since all measures of interest must have the same first marginal (i.e., μ) we proceed as in Lemma 50, but this time only regularizing conditional distributions. More precisely, for a given z in the support of $\mathcal{Z}$ we define the measure $\pi^\epsilon(\cdot|z)$ as follows:

$$\pi^\epsilon(A|z) := \int_{\mathcal{Z}} \frac{|\mathcal{B}_{\tilde{z},\epsilon} \cap A|}{|\mathcal{B}_{\tilde{z},\epsilon} \cap \mathcal{Z}|} d\pi^*(\tilde{z}|z).$$

We then define the measure π^ϵ as:

$$d\pi^\epsilon(z,\tilde{z}) = d\pi^\epsilon(\tilde{z}|z) d\mu(z).$$

Notice that the measure π^ϵ is such that $\pi^\epsilon_z = \mu$. Moreover, it is straightforward to show (repeating similar computations as in the proof of Lemma 50) that $W_1(\pi^\epsilon, \pi^*) \leq \epsilon$ and $\mathcal{H}(\pi^\epsilon || \pi_0) \leq -\log(k) - \log(k') - d' \log(\epsilon)$. The desired result now follows as in Lemma 50. □

Lemma 54 *Let $\tilde{z}_1, \ldots, \tilde{z}_N$ be N distinct points in $\mathcal{Z}$, and let $\pi_0^N := \mu \otimes (\frac{1}{N}\sum_{i=1}^N \delta_{\tilde{z}_i})$. Suppose that $\tilde{\mu}_0$ is a probability measure over $\mathcal{Z}$ that has a density with respect to the Lebesgue measure satisfying $\frac{d\tilde{\mu}_0}{dz} \geq k$ and $\mathcal{Z}$ is such that $|\mathcal{B}_{z,\epsilon} \cap \Theta| \geq k' \epsilon^{d'}$ uniformly in $z \in \mathcal{Z}$. Then*

$$\mathcal{Q}_\pi(\pi_0, \pi^*; \tau) \leq W_\infty(\tilde{\mu}_0, \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{z}_i}) + \frac{d'}{\tau}\left\{1 - \log\left(\frac{d'}{\tau}\right)\right\} + \frac{1}{\tau}\{-\log(k) - \log(k')\}.$$

In particular, if $\tilde{z}_1, \ldots, \tilde{z}_N$ are sampled independently from a $\tilde{\mu}_0$ with a density with respect to the Lebesgue measure that is bounded and bounded away from zero, then, with probability at least $1 - 1/N^2$,

$$\mathcal{Q}_\pi(\pi_0^N, \pi^*; \tau) \leq C\frac{\log(N)^{p_{d'}}}{N^{1/d'}} + \frac{d'}{\tau}\left\{1 - \log\left(\frac{d'}{\tau}\right)\right\} + \frac{1}{\tau}\{-\log(k) - \log(k')\},$$

for a constant C that depends on $\tilde{\mu}_0$, and a power $p_{d'}$ that takes the form $p_{d'} = 3/4$ if $d' = 2$ and $p_{d'} = 1/d'$ if $d' \geq 3$.

Proof The proof follows the same ideas as the ones in the proof of Lemma 52 and thus we skip the details. □

Proof of Theorem 35 On the one hand, by assumption, we can find T_1 such that for all $t > T_1$

$$|2B^2 \frac{1}{t}\int_0^t \int_0^s \eta_u du ds| \leq \frac{3}{4}\epsilon.$$

On the other hand, Lemmas 50 and 53 imply that there exists T_2 such that, for all $t > T_2$ and arbitrary π^* with $\pi^*_z = \mu$ and ν^*, we have

$$(\mathcal{Q}(\pi_0, \pi^*; \kappa B t) + (\mathcal{Q}(\nu_0, \nu^*; \kappa B t) \leq \frac{\epsilon}{4B}.$$

We conclude by taking $T^* = T_1 \vee T_2$ and using Lemma 49. □

Proof of Theorem 36 The proof is as the proof of Theorem 35 except that we use Lemmas 52 and 54 instead of Lemmas 50 and 53. □

5.3. *Proofs of Theorems 42 and 43*

In this section we present the proofs of Theorems 42 and 43.

Proof of Theorem 42 Throughout this proof we use m_t^* to denote the quantity

$$m_t^* := \sup_{\pi \text{ s.t. } \pi_z = \mu} \mathcal{U}(\pi, \nu_t).$$

From concavity-convexity of $\mathcal{U}$ in the linear interpolation sense we have for all arbitrary $\tilde{\pi}$ (with $\tilde{\pi}_z = \mu$) and $\tilde{\nu}$:

$$\begin{aligned}
\mathcal{U}(\tilde{\pi}, \overline{\nu}_t) - \mathcal{U}(\overline{\pi}_t, \tilde{\nu}) &\leq \frac{1}{t}\int_0^t (\mathcal{U}(\tilde{\pi}, \nu_s) - \mathcal{U}(\pi_s, \tilde{\nu}))ds \\
&= \frac{1}{t}\int_0^t (\mathcal{U}(\tilde{\pi}, \nu_s) - m_s^*)ds + \frac{1}{t}\int_0^t (m_s^* - \mathcal{U}(\pi_s, \tilde{\nu}))ds \\
&\leq \frac{1}{t}\int_0^t (m_s^* - \mathcal{U}(\pi_s, \tilde{\nu}))ds \\
&= \frac{1}{t}\int_0^t (m_s^* - \mathcal{U}(\pi_s, \nu_s))ds + \frac{1}{t}\int_0^t (\mathcal{U}(\pi_s, \nu_s) - \mathcal{U}(\pi_s, \tilde{\nu}))ds. \\
&=: \mathcal{I}_1 + \mathcal{I}_2
\end{aligned} \tag{5.16}$$

In the above, the second inequality follows from the definition of m_s^*. We will now control each of the terms $\mathcal{I}_1$ and $\mathcal{I}_2$ on the right-hand side of the above expression.

In order to control $\mathcal{I}_1$, we start by using the chain rule (e.g., see section 10.1.2 in [1]) to obtain an expression for $\frac{d}{ds}\mathcal{U}(\pi_s, \nu_s)$:

$$\begin{aligned}
\frac{d}{ds}\mathcal{U}(\pi_s, \nu_s) = &\int |\nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_s, \nu_s; z, \tilde{z})|^2 d\pi_s(z, \tilde{z}) + \kappa \int \mathcal{U}_\pi(\pi_s, \nu_s; z, \tilde{z})(\mathcal{U}_\pi(\pi_s, \nu_s; z, \tilde{z}) - \overline{\mathcal{U}}_{\pi,z}) d\pi_s(z, \tilde{z}) \\
&- \frac{\eta_t}{K}\int |\nabla_\theta \mathcal{U}_\nu(\pi_s, \nu_s; \theta)|^2 d\nu_s(\theta) - \frac{\kappa}{K}\int \mathcal{U}_\nu(\pi_s, \nu_s; \theta)(\mathcal{U}_\nu(\pi_s, \nu_s; \theta) - \mathcal{U}_\nu) d\nu_s(\theta);
\end{aligned} \tag{5.17}$$

in the above, we use the shorthand notation $\overline{\mathcal{U}}_{\pi,z}$ to denote $\int \mathcal{U}_\pi(\pi_s, \nu_s; z, \tilde{z}') d\pi_s(\tilde{z}'|z)$, and $\overline{\mathcal{U}}_\nu$ to denote $\int \mathcal{U}_\nu(\pi_s, \nu_s; \theta') d\nu_s(\theta')$. Assumption 39 implies that the first term on the right-hand side of (5.17) is bounded from below by $\lambda(m_s^* - \mathcal{U}(\pi_s, \nu_s))$. On the other hand, Jensen's inequality implies that the second term is non-negative. Finally, Assumptions 8 imply that the last terms can be bounded from below by $-\frac{\|\eta\|_\infty}{K}M^2 - \frac{2\kappa}{K}M^2$. It follows that for all $t \geq 0$

$$\begin{aligned}
\mathcal{U}(\pi_t, \nu_t) =& \mathcal{U}(\pi_0, \nu_0) + \int_0^t \frac{d}{dr}\mathcal{U}(\pi_r, \nu_r) dr \\
\geq& \mathcal{U}(\pi_0, \nu_0) - \frac{t}{K}\tilde{B} + \lambda \int_0^t (m_s^* - \mathcal{U}(\pi_s, \nu_s))ds,
\end{aligned}$$

where $\tilde{B} := (\|\eta\|_\infty + 2\kappa)M^2$. Subtracting m_t^* from both sides of the above inequality, we get, thanks to Assumptions 8,

$$\mathcal{U}(\pi_t,\nu_t) - m_t^* \geq -2M - \frac{t}{K}\tilde{B} + \lambda \int_0^t (m_s^* - \mathcal{U}(\pi_s,\nu_s))ds.$$

Equivalently,

$$m_t^* - \mathcal{U}(\pi_t,\nu_t) \leq 2M + \frac{t}{K}\tilde{B} - \lambda \int_0^t (m_s^* - \mathcal{U}(\pi_s,\nu_s))ds.$$

We thus see that the function $f(t) := m_t^* - \mathcal{U}(\nu_t,\pi_t)$ satisfies

$$f(t) \leq 2M + \frac{\tilde{B}}{K}t - \lambda \int_0^t f(s)ds,$$

and from Lemma 62 in Appendix A.4 we conclude that

$$\mathcal{I}_1 \leq \frac{\tilde{B}}{K\alpha} + \frac{A}{t},$$

for $A := \frac{1}{\lambda}|2M - \frac{\tilde{B}}{K\lambda}|$.

To estimate $\mathcal{I}_2$ in (5.18), we follow similar computations to those in the proof of Lemma 49 to conclude that

$$\begin{aligned}\int_0^t \mathcal{U}(\pi_s,\nu_s)ds - \int_0^t \mathcal{U}(\pi_s,\tilde{\nu})ds \leq \int_0^t (|\mathcal{U}_\nu(\pi_s,\nu_s;\cdot)\|_{BL}\|\tilde{\nu} - \hat{\nu}_s\|_{BL}^*)ds \\ - \int_0^t \left(\int \mathcal{U}_\nu(\pi_s,\nu_s;\theta)d(\hat{\nu}_s - \nu_s)(\theta)\right)ds,\end{aligned} \tag{5.18}$$

where now we use a modified hat process $\hat{\nu}$ satisfying

$$\partial_t \hat{\nu}_t = \frac{\eta_t}{K}\mathrm{div}_\theta(\hat{\nu}_t \nabla_\theta \mathcal{U}_\nu(\pi_t,\nu_t;\theta)),$$

initialized at an arbitrary $\hat{\nu}_0 \ll \nu_0$ with density in $L^\infty(\nu_0)$. Following a straightforward adaptation of Proposition 47, we can then see that

$$\mathcal{H}(\hat{\nu}_t||\nu_t) - \mathcal{H}(\hat{\nu}^0||\nu^0) \leq \frac{\kappa}{K}\int_0^t \left(\int \mathcal{U}_\nu(\pi_s,\nu_s;\theta)d(\hat{\nu}_s - \nu_s)(\theta)\right)ds, \quad \forall t \geq 0,$$

from where it now follows that

$$\begin{aligned}\mathcal{I}_2 &\leq \frac{1}{t}\int_0^t (\|\mathcal{U}_\nu(\pi_s,\nu_s;\cdot)\|_{BL}\|\tilde{\nu} - \hat{\nu}_s\|_{BL}^*)ds + \frac{K}{\kappa t}\mathcal{H}(\hat{\nu}_0||\nu_0) \\ &\leq B\|\tilde{\nu} - \hat{\nu}_0\|_{BL}^* + \frac{B^2}{Kt}\int_0^t \int_0^s \eta_\tau d\tau ds + \frac{K}{\kappa t}\mathcal{H}(\hat{\nu}_0||\nu_0).\end{aligned}$$

From the above we can deduce

$$\mathcal{I}_2 \leq B\mathcal{Q}_\nu(\nu_0,\tilde{\nu};\frac{\kappa}{K}Bt)) + \frac{B^2}{Kt}\int_0^t\int_0^s \eta_\tau d\tau ds.$$

Putting all our estimates together we obtain

$$\mathcal{U}(\tilde{\pi},\overline{\nu}_t) - \mathcal{U}(\overline{\pi}_t,\tilde{\nu}) \leq \frac{\tilde{B}}{K\lambda} + \frac{A}{t} + B\mathcal{Q}_\nu(\nu_0,\tilde{\nu};\frac{\kappa}{K}Bt)) + \frac{B^2}{Kt}\int_0^t\int_0^s \eta_\tau d\tau ds.$$

At this stage we can use the specific properties of ν_0 and use Lemma 50 to conclude that there are $r_0(\epsilon), K_0(\epsilon), t_0(\epsilon), r_1(\epsilon)$ such that, if $\frac{K}{t} \leq r_0(\epsilon)$, $K \geq K_0(\epsilon), t \geq t_0(\varepsilon)$, $\overline{\eta}/K \leq r_1(\epsilon)$, then

$$\sup_{\tilde{\pi}\in\mathcal{P}(\mathcal{Z}^2) \text{ s.t. } \tilde{\pi}_z=\mu} \mathcal{U}(\tilde{\pi},\overline{\nu}_t) - \inf_{\tilde{\nu}\in\mathcal{P}(\Theta)} \mathcal{U}(\overline{\pi}_t,\tilde{\nu}) \leq \epsilon.$$

□

Proof of Theorem 43 The proof is the same as the proof of Theorem 42, except that in the last step we use Lemma 52 instead of Lemma 50. □

6. Numerical examples

We illustrate our results numerically in the context of image classification on the MNIST database [30]. Our main purpose is to illustrate the effectiveness of the algorithm to obtain adversarially robust classifiers even away from the asymptotic regimes that we studied.

In this framework, we take the particles representing the distribution ν to be the training parameters (i.e. weights and biases) for simple convolutional networks with fixed widths and depths[2]; while the particles representing the distribution π are pairs of images where the first component is an image from the original database, and the second is an adversarial image built during the training process. We consider the square loss with an adversarial cost given by the Wasserstein-2 distance, i.e.

$$\mathcal{R}(\pi,\nu) = \int_{\mathcal{Z}\times\mathcal{Z}}\int_\Theta |h_\theta(\tilde{x}) - \tilde{y}|^2 d\nu(\theta)d\pi(z,\tilde{z}); \quad \mathcal{C}(\pi) = c_a\int_{\mathcal{Z}\times\mathcal{Z}} |z-\tilde{z}|^2 d\pi(z,\tilde{z})$$

where, $h_\theta(x)$ is the outcome of the convolutional network for the input x when setting the parameters of the network to be θ.

Given the nonlinear structure of the convolutional architecture, it would be extremely memory-consuming to apply directly the time average step 19 in Algorithm 1, as it would require us to keep track of copies of all intermediate networks in the training process. A possible solution, proposed for example in [17], is to calculate the time average on the weights only, while keeping the last position of the network-parameter particles. We implement also an alternative approach based on the resampling methods used in particle filters (see [31] for a review): we keep in memory at most a maximum number of network parameters (M'). At each

[2] Two layers with a convolutional kernel of size 5 and output channel sizes of 32 and 64 respectively, ReLu activation functions, and maxpool; and two linear layers at the end with respective output sizes 1000 and 10.

update time, we use residual systematic resampling (RSR) to pick M' parameters to keep from the list of the M' already contained in memory and the new bunch of M particles. Details of the (RSR) method can be found in [31] (see for example code 4 in Table 2). The time-average calculation of adversarial images is done similarly.

To illustrate our main result, we compute a proxy for the ratios

$$r_a := \frac{\sup_{\tilde{\pi}\in\mathcal{P}(\mathcal{Z}^2) \text{ s.t. } \tilde{\pi}_z=\mu} \mathcal{U}(\tilde{\pi},\nu^\star)}{\mathcal{U}(\pi^\star,\nu^\star)}, \text{ and } r_m := \frac{\inf_{\tilde{\nu}\in\mathcal{P}(\Theta)} \mathcal{U}(\pi^\star,\tilde{\nu})}{\mathcal{U}(\pi^\star,\nu^\star)},$$

where $(\pi^\star,\nu^\star)$ are the time-averaged distributions for the networks and adversarial images obtained after training. According to our results, we should reach an approximate Nash equilibrium, so we expect both ratios to be closed to zero. The proxy is computed as follows: we approximate the supremum in r_m by fixing $\nu^\star$ while training each one of the networks representing $\pi^\star$ with stochastic gradient descent for a fixed number of epochs (weights are kept constant). We compute then the relative change in total risk after this procedure. The proxy for r_a is computed analogously. We present a summary of the parameters used for the numerical experiments [3] and the results obtained in Table 1.

Model parameters	
N	4
M	2
η_t	$0.1(t+1)^{-1}$
κ	0.25
c_a	10

Implementation parameters	
Dataset	MNIST
Batch size	64
Training epochs	4

Results		
Accuracy		
	Time avg. on weights	Resampling
Clean	96.34%	93.53%
PGD (20 steps)	62.21%	58.49 %
Relative change of loss at solution - 5 additional training epochs		
	Time avg. on weights	Resampling
r_a	0.21%	0.03 %
r_m	1.82%	3.5%

TABLE 1 *Parameters and results of numerical experiment.*

[3] The code used to run these experiments can be found at https://github.com/camgt/robust_learning

Intuitively, we expect that the classification provided by the final time-averaged distributions of networks should be both effective and robust. To test this idea, we evaluate the accuracy of this classifier with a *clean* test sample, independent of the original distribution, and with an *adversarial* version constructed via modification of the latter using projected gradient descent (PGD) with 20 steps and a step size of 0.04. PGD constructs adversarial images by repeatedly perturbing each pixel in the image by a fixed amount choosing the sign of the perturbation to be the same as the sign of the gradient of the loss function with respect to the entry. See for example [35]. Results of this test are also included in Table 1. We observe that the overall procedure has degraded a bit the clean performance of the network but significantly improved the resistance to adversarial attacks. For reference, a baseline obtained by a simple training of a network with the same characteristics obtains in the same number of epochs a clean accuracy of 98.41% but an accuracy after the PGD attack of only 0.68% (compare also with the results in [22]). Table 1 shows that in the tested case, calculating the time average on the weights only is not just simpler but also has better results than the resampling procedure. However, there may be settings, not explored here, where the latter approach may be more advantageous. Exploring this would be an interesting research direction.

7. Conclusions

In this paper we have studied minmax problems over spaces of probability measures with a payoff structure motivated by adversarial training problems in supervised learning settings; we have studied gradient ascent-descent dynamics aimed at solving these problems. The dynamics that we have studied take the form of an evolutionary system of PDEs that can be discretized using systems of finitely many interacting particles. Under some reasonable assumptions on the payoff structure of the game, we can show that the proposed particle systems are consistent and recover the solution of the original PDE as the number of particles in the system scales up. We have also discussed the behavior of our evolutionary system of PDEs as time tends to infinity, showing that in a certain sense (see below) the system can produce approximate Nash equilibria for the adversarial game. Our results are stated under suitable assumptions on initialization in two settings of interest: 1) for non-convex non-concave payoffs (convexity and concavity understood in a suitable OT-sense), and 2) non-convex strongly-concave problems (again, in a suitable OT sense). Both settings are realistic in adversarial learning games for supervised learning tasks, while in general convexity can only be enforced by introducing additional (exogenous) regularization penalties in the payoff function.

Due to the lack of convexity of the payoff in our problem (w.r.t. the metric inducing the dynamics of our ascent-descent dynamics), we can only guarantee that time averages of the measures produced by our PDE system become approximate Nash equilibria in the $t \to \infty$ limit. For our algorithms to follow more closely our theoretical results, it was thus important to discuss strategies for constructing surrogate time averages that do not incur in memory overload and that can still recover approximate Nash equilibria for the game, at least in some benchmark learning tasks.

There are several directions for research that our work motivates. Here we mention a few.

First, the theoretical analysis that we have presented in this paper presupposes that the optimization updates take into account all (perturbed) data points, but in practice a natural strategy is to use batches of data to compute the loss (and its gradient) at each iteration. We

thus believe that it is of interest to study how the use of stochastic gradient descent (SGD) affects the resulting PDE system.

Another interesting direction for future research is the exploration of broader frameworks for adversarial learning covering *multiclass classification* settings (as opposed to regression problems as considered in this paper or just binary classification problems). In principle, one could even consider situations where prior information on the similarity of classes in a learning problem is available (e.g. the class "guitar" may be considered more similar to class "violin" than to class "baseball") as in those situations it may be beneficial to use such information to construct more nuanced models for risk and admissible adversarial attacks; for example, the work [44] discusses the advantages of using similarity or hierarchical structures between classes in different learning tasks; the work [16] explicitly discusses how to build similarities between labels using their semantic content. Our framework indeed seems better suited for regression problems, since in that setting the cost function C for the adversary can be naturally defined using something like the Wasserstein distance over the feature space times the label space, where the latter space is simply the real line. When the response variable has a discrete structure, it is less obvious how one can still define a reasonable (from the modelling perspective) cost structure for the adversary in such a way that the resulting adversarial game can still be solved using an ascent-descent scheme as explored in this paper.

Data Availability Statement

No new data were generated or analysed in support of this research. Code used to run experiments in section 6 can be found at https://github.com/camgt/robust_learning.

Acknowledgments

The authors are thankful to Theodor Misiakiewicz, Katy Craig, Matt Jacobs, and Lénaïc Chizat for enligthening discussions on topics related to the content of this paper.

Funding

NGT was partially supported by NSF-DMS grants 2005797 and 2236447, and would also like to thank the IFDS at UW-Madison and NSF through TRIPODS grant 2023239 for their support.

References

1. L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.
2. P. Awasthi, N. Frank, and M. Mohri. On the existence of the adversarial Bayes classifier. *Advances in Neural Information Processing Systems*, 34:2978–2990, 2021.
3. A. N. Bhagoji, D. Cullina, and P. Mittal. Lower bounds on adversarial robustness from optimal transport. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
4. J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

5. J. Blanchet, K. Murthy, and V. A. Nguyen. Statistical analysis of wasserstein distributionally robust estimators. 2021.
6. V. I. Bogachev, A. V. Kolesnikov, and K. V. Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309, apr 2005.
7. F. Bolley. Separability and completeness for the Wasserstein distance. In C. Donati-Martin, M. Émery, A. Rouault, and C. Stricker, editors, *Séminaire de Probabilités XLI*, volume 1934, pages 371–377. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
8. L. Bungert and K. Stinson. Gamma-convergence of a nonlocal perimeter arising in adversarial machine learning. 2022.
9. L. Bungert, N. G. Trillos, and R. Murray. The geometry of adversarial training in binary classification. 2021.
10. T. Champion, L. De Pascale, and P. Juutinen. The ∞-wasserstein distance: Local solutions and existence of optimal transport maps. *SIAM Journal on Mathematical Analysis*, 40(1):1–20, 2008.
11. R. Chen and I. C. Paschalidis. Distributionally robust learning. *Foundations and Trends® in Optimization*, 4(1-2):1–243, 2020.
12. L. Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194(1-2):487–532, 2022.
13. L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
14. L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.
15. R. L. Dobrushin. Vlasov equations. *Functional Analysis and Its Applications*, 13(2):115–123, 1979.
16. Ü. Dogan, A. A. Deshmukh, M. B. Machura, and C. Igel. Label-similarity curriculum learning. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 174–190, Cham, 2020. Springer International Publishing.
17. C. Domingo-Enrich, S. Jelassi, A. Mensch, G. Rotskoff, and J. Bruna. A mean-field analysis of two-player zero-sum games. *Advances in Neural Information Processing Systems*, 33:20215–20226, 2020.
18. W. E, C. Ma, and L. Wu. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, Feb 2022.
19. C. Finlay and A. M. Oberman. Scaleable input gradient regularization for adversarial robustness. *Machine Learning with Applications*, 3:100017, 2021.
20. N. S. Frank. Existence and minimax theorems for adversarial surrogate risks in binary classification. 2022.
21. T. O. Gallouët and L. Monsaingeon. A jko splitting scheme for kantorovich–fisher–rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017.
22. C. A. García Trillos and N. García Trillos. On the regularized risk of distributionally robust learning over deep neural networks. *Research in the Mathematical Sciences*, 9(3):54, Sept. 2022.
23. N. García Trillos, M. Jacobs, and J. Kim. The multimarginal optimal transport formulation of adversarial multiclass classification. 2022.
24. N. García Trillos and R. Murray. Adversarial classification: Necessary conditions and geometric flows. *Journal of Machine Learning Research*, 23(187):1–38, 2022.
25. N. García Trillos and D. Slepčev. Continuum limit of total variation on point clouds. *Archive for Rational Mechanics and Analysis*, 220(1):193–241, 4 2016.
26. N. García Trillos and D. Slepčev. On the rate of convergence of empirical measures in ∞-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358–1383, 2015.
27. J. K. Hale. *Ordinary differential equations*. Second edition. Malabar, Fla. : R. E. Krieger, 1980.

28. S. Kondratyev, L. Monsaingeon, and D. Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Advances in Differential Equations*, 21(11/12):1117 – 1164, 2016.
29. D. Kuhn, P. Esfahani, V. Nguyen, and S. Shafieezadeh-Abadeh. *Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning*, pages 130–166. 10 2019.
30. D. Y. LeCun, C. Cortes, and C. J. Burges. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.
31. T. Li, M. Bolic, and P. M. Djuric. Resampling Methods for Particle Filtering: Classification, implementation, and strategies. *IEEE Signal Processing Magazine*, 32(3):70–86, May 2015. Conference Name: IEEE Signal Processing Magazine.
32. M. Liero, A. Mielke, and G. Savaré. Optimal transport in competition with reaction: The hellinger–kantorovich distance and geodesic curves. *SIAM Journal on Mathematical Analysis*, 48(4):2869–2911, 2016.
33. Y. Lu. Two-scale gradient descent ascent dynamics finds mixed nash equilibria of continuous games: A mean-field perspective. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22790–22811. PMLR, 2023.
34. C. Lyu, K. Huang, and H.-N. Liang. A unified gradient regularization family for adversarial examples. In *2015 IEEE International Conference on Data Mining*, pages 301–309, 2015.
35. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.
36. L. Meunier, M. Scetbon, R. B. Pinot, J. Atif, and Y. Chevaleyre. Mixed nash equilibria in the adversarial examples game. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7677–7687. PMLR, 18–24 Jul 2021.
37. S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato, and P. Frossard. Robustness via curvature regularization, and vice versa. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9070–9078, 2019.
38. J. Peszek and D. Poyato. Heterogeneous gradient flows in the topology of fibered optimal transport. *Calculus of Variations and Partial Differential Equations*, 62(9), Oct. 2023.
39. M. S. Pydi and V. Jog. Adversarial risk via optimal transport and optimal couplings. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7814–7823. PMLR, 13–18 Jul 2020.
40. M. S. Pydi and V. Jog. The many faces of adversarial risk. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10000–10012. Curran Associates, Inc., 2021.
41. A. S. Ross and F. Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. 2018.
42. K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann. Adversarially robust training through structured gradient regularization. 2018.
43. F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.
44. C. N. Silla and A. A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, Apr. 2010.
45. A. Sinha, H. Namkoong, and J. C. Duchi. Certifying some distributional robustness with principled adversarial training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net,

2018.
46. T. Séjourné, G. Peyré, and F.-X. Vialard. Unbalanced optimal transport, from theory to numerics. 2022.
47. D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
48. Z. Tu, J. Zhang, and D. Tao. Theoretical analysis of adversarial learning: A minimax approach. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
49. C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.
50. C. Villani. *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.
51. G. Wang and L. Chizat. An Exponentially Converging Particle Method for the Mixed Nash Equilibrium of Continuous Games, Nov. 2022. arXiv:2211.01280 [cs, math].
52. S. Wojtowytsch. On the convergence of gradient descent training for two-layer relu-networks in the mean field regime. *CoRR*, abs/2005.13530, 2020.
53. S. Wojtowytsch and W. E. Can shallow neural networks beat the curse of dimensionality? a mean field training perspective. *IEEE Transactions on Artificial Intelligence*, 1(2):121–129, 2020.
54. E. C. Yeats, Y. Chen, and H. Li. Improving gradient regularization using complex-valued neural networks. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11953–11963. PMLR, 18–24 Jul 2021.
55. H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 09–15 Jun 2019.

A. Auxiliary results and computations

A.1. *Equivalence between (1.1) and DRO problems*

In this section we assume that the payoff $\mathcal{U}$ has the form (1.3) and that

$$\mathcal{R}(\pi,\nu) = R(\pi_{\tilde{z}},\nu), \quad C(\mu,\tilde{\mu}) = \inf_{\pi\in\Gamma(\mu,\tilde{\mu})} \mathcal{C}(\pi), \quad \mathcal{C}(\pi) = \int c(z,\tilde{z})d\pi(z,\tilde{z}). \tag{A.1}$$

In other words, we assume that $\mathcal{R}$'s dependence on π is only through π's second marginal, $\mathcal{C}$ is an average cost, and $C(\mu,\tilde{\mu})$ is the optimal transport problem between measures μ and $\tilde{\mu}$ for the cost c in the definition of $\mathcal{C}$. We will show that under these assumptions problems (1.1) and (1.2) are equivalent. By this we mean that it is possible to easily construct approximate Nash equilibria for one of the problems from approximate Nash equilibria for the other.

Definition 55 *We say that $(\tilde{\mu}^*,\nu^*)$ is an ε-Nash equilibrium for (1.2) if*

$$\sup_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \{R(\tilde{\mu},\nu^*) - C(\mu,\tilde{\mu})\} - \inf_{\nu\in\mathcal{P}(\Theta)} \{R(\tilde{\mu}^*,\nu) - C(\mu,\tilde{\mu}^*)\} \leq \varepsilon.$$

Proposition 56 *Suppose $\mathcal{U}$ has the form (1.3) and $\mathcal{R},\mathcal{C},C$ are as in (A.1). If (π^*,ν^*) is an ε-Nash equilibrium for problem (1.1) (see Definition 3), then $(\pi^*_{\tilde{z}},\nu^*)$ is an ε-Nash equilibrium*

for (1.2)*. Conversely, if* $(\tilde{\mu}^*,\nu^*)$ *is an* ε*-Nash equilibrium for* (1.2) *and* $\pi^* \in \Gamma(\mu,\tilde{\mu})$ *realizes the cost* $C(\mu,\tilde{\mu}^*)$*, then* (π^*,ν^*) *is an* ε*-Nash equilibrium for* (1.1)*.*

Proof Let (π^*,ν^*) be an almost Nash equilibrium for (1.1) and let $\tilde{\mu}^* = \pi^*_{\tilde{z}}$. For arbitrary $\tilde{\mu} \in \mathcal{P}(\mathcal{Z})$, assume for simplicity that there is $\hat{\pi} \in \Gamma(\mu,\tilde{\mu})$ that achieves the cost $C(\mu,\tilde{\mu})$, i.e., $\mathcal{C}(\hat{\pi}) = C(\mu,\tilde{\mu})$. Also, let $\tilde{\nu} \in \mathcal{P}(\Theta)$ be arbitrary. We see that:

$$\begin{aligned} R(\tilde{\mu}^*,\tilde{\nu}) - C(\mu,\tilde{\mu}^*) + \varepsilon &\geq R(\pi^*_{\tilde{z}},\tilde{\nu}) - \mathcal{C}(\pi^*) + \varepsilon = \mathcal{U}(\pi^*,\tilde{\nu}) + \varepsilon \\ &\geq \mathcal{U}(\hat{\pi},\nu^*) \\ &= R(\tilde{\mu},\nu^*) - \mathcal{C}(\hat{\pi}) \\ &= R(\tilde{\mu},\nu^*) - C(\mu,\tilde{\mu}). \end{aligned}$$

Given that $\tilde{\mu}$ and $\tilde{\nu}$ were arbitrary, we conclude that

$$\sup_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \{R(\tilde{\mu},\nu^*) - C(\mu,\tilde{\mu})\} - \inf_{\nu\in\mathcal{P}(\Theta)} \{R(\tilde{\mu}^*,\nu) - C(\mu,\tilde{\mu}^*)\} \leq \varepsilon,$$

which is what we wanted to prove.

Conversely, suppose that $(\tilde{\mu}^*,\nu^*)$ is an ε-Nash equilibrium for (1.2) and suppose $\pi^* \in \Gamma(\mu,\tilde{\mu}^*)$ realizes $C(\mu,\tilde{\mu}^*)$. Consider arbitrary π,ν with $\pi_z = \mu$ and let $\tilde{\mu} = \pi_{\tilde{z}}$. Then

$$\begin{aligned} \mathcal{U}(\pi^*,\nu) + \varepsilon &= R(\tilde{\mu}^*,\nu) - \mathcal{C}(\pi^*) + \varepsilon \\ &= R(\tilde{\mu}^*,\nu) - C(\mu,\tilde{\mu}^*) + \varepsilon \\ &\geq R(\tilde{\mu},\nu^*) - C(\mu,\tilde{\mu}) \\ &\geq R(\tilde{\mu},\nu^*) - \mathcal{C}(\pi) \\ &= \mathcal{U}(\pi,\nu^*). \end{aligned}$$

Since π (with $\pi_z = \mu$) and ν were arbitrary, we conclude that (π^*,ν^*) is an ε-Nash equilibrium for (1.1), as we wanted to prove. □

A.2. *On the PL condition of Assumption 39*

Proposition 57 *Suppose that* $\mathcal{Z}$ *is a convex set and that we select an activation function and a loss function in the setting in 1.1 that are twice continuously differentiable. Then the function* $\mathcal{U}$ *in* (1.3)*, with* $\mathcal{R}$ *and* $\mathcal{C}$ *as in* (1.4) *and* (1.5)*, satisfies the condition in Assumption 39 for all large enough* c_a*.*

Proof A straightforward computation reveals that in this case

$$\mathcal{U}_\pi(\pi,\nu;z,\tilde{z}) = \ell(h_\nu(\tilde{x}),\tilde{y}) - c_a|z-\tilde{z}|^2 =: \mathcal{U}(\nu;z,\tilde{z}).$$

Assuming that the loss function ℓ and the activation function f are at least twice continuously differentiable, we can conclude that the function $\tilde{z} \in \mathcal{Z} \mapsto \mathcal{U}(\nu;z,\tilde{z})$ (for fixed z and ν) is strongly concave (for all z and ν), provided that c_a is large enough. Indeed, this is simply because we

can bound, uniformly over z,ν, the second derivatives of the first term in $\mathcal{U}(\nu;z,\tilde{z})$. Thanks to this and Theorem 5.15 ii) in [49], we deduce that there is $\lambda>0$ such that for every $z\in\mathcal{Z}$ and $\Upsilon\in\mathcal{P}(\mathcal{Z})$ we have

$$\int_{\mathcal{Z}}|\nabla_{\tilde{z}}\mathcal{U}(\nu;z,\tilde{z})|^2 d\Upsilon(\tilde{z})\geq\lambda(\sup_{\hat{\Upsilon}\in\mathcal{P}(\mathcal{Z})}\int_{\mathcal{Z}}\mathcal{U}(\nu;z,\tilde{z})d\hat{\Upsilon}(\tilde{z})-\int_{\mathcal{Z}}\mathcal{U}(\nu;z,\tilde{z})d\Upsilon(\tilde{z})).$$

In particular, for a given $\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z})$ with $\pi_{0,z}=\mu$, we have

$$\int_{\mathcal{Z}}|\nabla_{\tilde{z}}\mathcal{U}(\nu;z,\tilde{z})|^2 d\pi(\tilde{z}|z)\geq\lambda(\sup_{\hat{\Upsilon}\in\mathcal{P}(\mathcal{Z})}\int_{\mathcal{Z}}\mathcal{U}(\nu;z,\tilde{z})d\hat{\Upsilon}(\tilde{z})-\int_{\mathcal{Z}}\mathcal{U}(\nu;z,\tilde{z})d\pi(\tilde{z}|z)),$$

for all $z\in\mathcal{Z}$ and all $\nu\in\mathcal{P}(\Theta)$. Integrating over z with respect to μ on both sides of the above inequality, we get

$$\begin{aligned}\int_{\mathcal{Z}\times\mathcal{Z}}|\nabla_{\tilde{z}}\mathcal{U}_\pi(\pi,\nu;z,\tilde{z})|^2 d\pi(z,\tilde{z}) &= \int_{\mathcal{Z}\times\mathcal{Z}}|\nabla_{\tilde{z}}\mathcal{U}(\nu;z,\tilde{z})|^2 d\pi(z,\tilde{z})\\ &\geq\lambda\left(\int_{\mathcal{Z}}\left(\sup_{\hat{\Upsilon}\in\mathcal{P}(\mathcal{Z})}\int_{\mathcal{Z}}\mathcal{U}(\nu;z,\tilde{z})d\hat{\Upsilon}(\tilde{z})\right)d\mu(z)-\mathcal{U}(\pi,\nu)\right)\\ &\geq\lambda\left(\sup_{\hat{\pi}\in\mathcal{P}(\mathcal{Z}^2)\text{ s.t. }\hat{\pi}_z=\mu}\mathcal{U}(\hat{\pi},\nu)-\mathcal{U}(\pi,\nu)\right).\end{aligned}$$

□

A.3. *Auxiliary lemmas for the construction of approximate initializations in Theorems 12 and 25*

Proposition 58 *Let A,B be two bounded Borel subsets of $\mathbb{R}^d$ and $\mathbb{R}^{d'}$, respectively. Let $\mu\in\mathcal{P}(A)$, and let $u\in A\mapsto\mu_u(\cdot)\in\mathcal{P}(B)$ be a measurable map.*

For every sequence $\{\Upsilon_n\}_{n\in\mathbb{N}}\subseteq\Gamma(\mu,\mu)$ satisfying

$$\lim_{n\to\infty}\int_{A\times A}|u-u'|d\Upsilon_n(u,u')=0,$$

we have

$$\lim_{n\to\infty}\int_{A\times A}W_1(\mu_u,\mu_{u'})d\Upsilon_n(u,u')=0.$$

Proof Sequences $\{\Upsilon_n\}_{n\in\mathbb{N}}$ satisfying the hypothesis in the proposition are called *stagnating sequences* of transport plans; see [25].

Let $\varepsilon>0$. For such $\varepsilon>0$ we can build a finite partition $\{B_l\}_{l=1,\dots,L}$ of the set B in such a way that each set B_l has diameter less than $\varepsilon/3$; this partition can be constructed by simply intersecting a grid of boxes in $\mathbb{R}^{k'}$ of diameter less than $\varepsilon/3$ with the set B. Select now a point v_l in each of the B_l. Associated to each $l=1,\dots,L$, we define a function $h_l\in L^1(\mu)$ as follows:

for every u in the support of μ, we define $h_l(u) := \mu_u(B_l)$. We now consider the measures $\hat{\mu}_u := \sum_{l=1}^{L} h_l(u)\delta_{v_l}$. Notice that these are probability measures satisfying $W_1(\hat{\mu}_u, \mu_u) \leq \varepsilon/3$. In particular, using the triangle inequality for W_1 we deduce

$$\begin{aligned}\int_{A\times A} W_1(\mu_u, \mu_{u'}) d\Upsilon_n(u,u') &\leq \int_{A\times A} W_1(\mu_u, \hat{\mu}_u) d\Upsilon_n(u,u') + \int_{A\times A} W_1(\hat{\mu}_u, \hat{\mu}_{u'}) d\Upsilon_n(u,u') \\ &\quad + \int_{A\times A} W_1(\hat{\mu}_{u'}, \mu_{u'}) d\Upsilon_n(u,u'). \\ &\leq \frac{2}{3}\varepsilon + \int_{A\times A} W_1(\hat{\mu}_u, \hat{\mu}_{u'}) d\Upsilon_n(u,u').\end{aligned}$$

Let us now find an upper bound for the term $\int_{A\times A} W_1(\hat{\mu}_u, \hat{\mu}_{u'}) d\Upsilon_n(u,u')$. By the Kantorovich duality for the W_1 distance, we have

$$W_1(\hat{\mu}_u, \hat{\mu}_{u'}) = \sup_{\mathrm{Lip}(f)\leq 1} \{\int f(v) d\hat{\mu}_u(v) - \int f(v) d\hat{\mu}_{u'}(v)\}.$$

Since the set B is bounded, and the argument inside the sup is invariant under addition of a constant to a given f, we can further assume that the sup is taken over functions f whose supremum norm is bounded by a fixed constant C. For such a function f we have

$$\int f(v) d\hat{\mu}_u(v) - \int f(v) d\hat{\mu}_{u'}(v) = \sum_{l=1}^{L} (h_l(u) - h_l(u')) f(v_l) \leq C \sum_{l=1}^{L} |h_l(u) - h_l(u')|.$$

From the above it follows

$$\int_{A\times A} W_1(\hat{\mu}_u, \hat{\mu}_{u'}) d\Upsilon_n(u,u') \leq C \sum_{l=1}^{L} \int_{A\times A} |h_l(u) - h_l(u')| d\Upsilon_n(u,u').$$

We now invoke Lemma 3.10 in [25] to conclude that the right-hand side of the above expression converges to zero as $n \to \infty$. In particular, there exists N large enough such that for all $n \geq N$ we have $C\sum_{l=1}^{L} \int_{A\times A} |h_l(u) - h_l(u')| d\Upsilon_n(u,u') \leq \frac{\varepsilon}{3}$. In turn, we conclude that if $n \geq N$, then

$$\int_{A\times A} W_1(\mu_u, \mu_{u'}) d\Upsilon_n(u,u') \leq \varepsilon.$$

This establishes the desired result. □

Lemma 59 *Let A, B be two bounded Borel subsets of $\mathbb{R}^d$ and $\mathbb{R}^{d'}$, respectively. Let $\mu \in \mathcal{P}(A)$, and let $u \in A \mapsto \mu_u(\cdot) \in \mathcal{P}(B)$ be a measurable map.*

Let $u_1, \ldots, u_n, \ldots$ be a sequence of i.i.d. samples from μ, and for each $i \in \mathbb{N}$, let $v_{i1}, \ldots, v_{im}, \ldots$, be i.i.d. samples from $\mu_{u_i}(\cdot)$. For each n and m consider the (random) measures

$$\mu^{n,m} := \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{(u_i, v_{ij})}, \quad \mu^n := \frac{1}{n} \sum_{i=1}^{n} \delta_{u_i},$$

and let $\mu^{n,m}(\cdot|u)$ be the conditional distribution, according to $\mu^{n,m}$, of the variable v given the value u of the first coordinate.

Then

$$\lim_{n\to\infty}\lim_{m\to\infty}\mathbb{E}\left[\inf_{\Upsilon_n\in\Gamma_{Opt}(\mu^n,\mu)}\int W_1(\mu^{n,m}(\cdot|u),\mu_{u'})d\Upsilon_n(u,u')\right]=0.$$

In particular, there is a sequence $\{(n_k,m_k)\}_{k\in\mathbb{N}}$ *such that*

$$\lim_{k\to\infty}\mathbb{E}\left[\inf_{\Upsilon_k\in\Gamma_{Opt}(\mu^{n_k},\mu)}\int W_1(\mu^{n_k,m_k}(\cdot|u),\mu_{u'})d\Upsilon_k(u,u')\right]=0,$$

and a subsequence (not relabeled) such that

$$\lim_{k\to\infty}\inf_{\Upsilon_k\in\Gamma_{Opt}(\mu^{n_k},\mu)}\int W_1(\mu^{n_k,m_k}(\cdot|u),\mu_{u'})d\Upsilon_k(u,u')=0,$$

almost surely.

Proof Let $\Upsilon_n\in\Gamma_{\mathrm{Opt}}(\mu^n,\mu)$. By Corollary 5.22 in [50] this random measure can be chosen in a measurable way over the tacitly defined sample space giving support to the random variables in the problem.

From the triangle inequality for W_1 we have

$$\begin{aligned}\int W_1(\mu^n(\cdot|u),\mu_{u'})d\Upsilon_n(u,u')&\le\int W_1(\mu^n(\cdot|u),\mu_u)d\Upsilon_n(u,u')+\int W_1(\mu_u,\mu_{u'})d\Upsilon_n(u,u')\\&=\frac{1}{n}\sum_{i=1}^{n}W_1(\mu^{n,m}(\cdot|u_i),\mu_{u_i})+\int W_1(\mu_u,\mu_{u'})d\Upsilon_n(u,u').\end{aligned}\tag{A.2}$$

In what follows we analyze each of the terms on the right-hand side of the above expression. We start with the second term.

Let us introduce $\hat{\Upsilon}_n:=\mathbb{E}[\Upsilon_n]$, the (deterministic) measure that acts on test functions ϕ according to

$$\int\phi(u,u')d\hat{\Upsilon}_n(u,u')=\mathbb{E}[\int\phi(u,u')d\Upsilon_n(u,u')].$$

It is straightforward to see that $\hat{\Upsilon}_n\in\Gamma(\mu,\mu)$. Now, due to the boundedness of the space A and the fact that μ^n converges weakly to μ almost surely, we know that, almost surely,

$$\lim_{n\to\infty}\int|u-u'|d\Upsilon_n(u,u')=\lim_{n\to\infty}W_1(\mu^n,\mu)=0.$$

By the dominated convergence theorem it thus follows

$$\lim_{n\to\infty}\int|u-u'|d\hat{\Upsilon}_n(u,u')=\lim_{n\to\infty}\mathbb{E}[\int|u-u'|d\Upsilon_n(u,u')]=0.$$

In particular, $\{\hat{\Upsilon}_n\}_{n\in\mathbb{N}}$ is a stagnating sequence of transport plans for μ, and thus, from Lemma 58 it follows that

$$\lim_{n\to\infty}\mathbb{E}[\int W_1(\mu_u,\mu_{u'})d\Upsilon_n(u,u')] = \lim_{n\to\infty}\int W_1(\mu_u,\mu_{u'})d\hat{\Upsilon}_n(u,u') = 0.$$

We now study the first term on the right-hand side of (A.2). To avoid introducing cumbersome notation, we will assume for simplicity that all the u_i are different so that in particular $\mu^{n,m}(\cdot|u_i) = \frac{1}{m}\sum_{j=1}^m \delta_{v_{ij}}$. We then have

$$\begin{aligned}\lim_{m\to\infty}\mathbb{E}[\frac{1}{n}\sum_{i=1}^n W_1(\mu^{n,m}(\cdot|u_i),\mu_{u_i})] &= \mathbb{E}[\lim_{m\to\infty}\frac{1}{n}\sum_{i=1}^n W_1(\mu^{n,m}(\cdot|u_i),\mu_{u_i})]\\ &= \mathbb{E}[\mathbb{E}[\lim_{m\to\infty}\frac{1}{n}\sum_{i=1}^n W_1(\mu^{n,m}(\cdot|u_i),\mu_{u_i})|u_1,\dots,u_n]]\\ &= 0,\end{aligned} \tag{A.3}$$

where we have used the dominated convergence theorem in the first line, and the fact that $\frac{1}{m}\sum_{j=1}^m \delta_{v_{ij}}$ converges almost surely in the Wasserstein sense toward μ_{u_i} in the last line. □

Remark 60 *Let $\{\mu^n\}_{n\in\mathbb{N}}$ be a sequence of probability measures over $A\times B$ and let μ be a probability measure. We show that the condition*

$$\inf_{\Upsilon_n\in\Gamma_{Opt}(\mu_u^n,\mu_u)}\int W_1(\mu^n(\cdot|u),\mu(\cdot|u'))d\Upsilon_n(u,u')\to 0$$

implies

$$W_1(\mu^n,\mu)\to 0,$$

while the converse is not true in general; in the above, μ_u^n and μ_u denote the first marginals of μ^n and μ, respectively. Indeed, suppose that the first condition holds, and for each u,u' let $\Upsilon^{u,u'}$ be a coupling between $\mu^n(\cdot|u)$ and $\mu(\cdot|u')$ realizing the W_1 distance. Also, choose Υ_n in $\Gamma_{Opt}(\mu^n,\mu)$ such that $\int W_1(\mu^n(\cdot|u),\mu(\cdot|u'))d\Upsilon_n(u,u')\to 0$, and consider the measure $d\pi_n((u,v),(u',v')) := d\Upsilon^{u,u'}(v,v')d\Upsilon_n(u,u')$. It is straightforward to verify that $\pi_n\in\Gamma(\mu^n,\mu)$ and that $\int|(u,v)-(u',v')|d\pi_n\to 0$. This implies $W_1(\mu^n,\mu)\to 0$.

As we stated earlier, the converse statement is not true. For example, taking $A=[0,1]$, $B=[0,1]$, μ the uniform measure on $[0,1]^2$, and $\mu^n = \frac{1}{n}\sum_j \delta_{(u_j,v_j)}$ with $(u_1,v_1),\dots,(u_n,v_n)$ i.i.d. samples from μ, we see that $W_1(\mu^n,\mu)\to 0$, while $\inf_{\Upsilon_n\in\Gamma_{Opt}(\mu_u^n,\mu_u)}\int W_1(\mu^n(\cdot|u),\mu(\cdot|u'))d\Upsilon_n(u,u') = 1$ for all n.

Lemma 61 *Consider the same setting and notation as in Lemma 59. Let $\rho: A\times B\to[0,D]$ be a measurable function satisfying*

$$\int\rho(u,v)d\mu_u(v) = 1,$$

for all u in the support of μ. Then, with probability one,

$$\lim_{n\to\infty}\lim_{m\to\infty}\frac{1}{n}\sum_{i=1}^n\left|\frac{1}{\frac{1}{m}\sum_{j=1}^m\rho(u_i,v_{ij})}-1\right| = 0.$$

Proof This is a direct consequence of the law of large numbers. □

A.4. *Auxiliary lemmas for section 5*

The following result follows from a Gronwall-type argument.

Lemma 62 *Let $\tilde{B}, M, K, \lambda > 0$, and suppose $h : [0,\infty) \to [0,\infty)$ is a function satisfying*

$$h(t) \leq 2M + \frac{\tilde{B}}{K}t - \lambda \int_0^t h(s)ds,$$

for all t. Then, for all $T > 0$,

$$\frac{1}{T}\int_0^T h(s)ds \leq \frac{\tilde{B}}{K\lambda} + \frac{A}{T},$$

where $A := \frac{1}{\lambda}|2M - \frac{\tilde{B}}{K\lambda}|$.

Proof The condition on h can be equivalently written as

$$h(t) - \frac{\tilde{B}}{K\lambda} \leq (2M - \frac{\tilde{B}}{K\lambda}) - \lambda \int_0^t (h(s) - \frac{\tilde{B}}{K\lambda})ds.$$

Let $H(t) := \int_0^t (h(s) - \frac{\tilde{B}}{K\lambda})ds$. The above condition can thus be written as

$$H'(t) \leq (2M - \frac{\tilde{B}}{K\lambda}) - \lambda H(t).$$

From this it follows that

$$\frac{d}{dt}(H(t)e^{\lambda t}) \leq (2M - \frac{\tilde{B}}{K\lambda})e^{\lambda t}.$$

Integrating the above expression, we get:

$$H(t)e^{\lambda t} \leq (2M - \frac{\tilde{B}}{K\lambda})\frac{1}{\lambda}(e^{\lambda t} - 1),$$

or what is the same

$$H(t) \leq (2M - \frac{\tilde{B}}{K\lambda})\frac{1}{\lambda} - (2M - \frac{\tilde{B}}{K\lambda})\frac{1}{\lambda}e^{-\lambda t}.$$

Recalling the definition of H, we deduce that

$$\frac{1}{T}\int_0^T h(s)ds \leq \frac{\tilde{B}}{K\lambda} + \frac{1}{T}(2M - \frac{\tilde{B}}{K\lambda})\frac{1}{\lambda} - \frac{1}{T}(2M - \frac{\tilde{B}}{K\lambda})\frac{1}{\lambda}e^{-\lambda T} \leq \frac{\tilde{B}}{K\lambda} + \frac{A}{T}.$$

□

B. Riemannian structure for $\mathcal{P}(\Theta \times [0,\infty))$

In this section we review the Riemannian structure for the space $\mathcal{P}(\Theta \times [0,\infty))$ that motivates the PDE dynamics given in (2.3).

B.1. *A metric on the lifted space*

We start by defining a metric tensor over the space $\Theta \times (0,\infty)$ according to:

$$g_{(\theta,\alpha)}((v,s),(\tilde{v},\tilde{s})) := \frac{\alpha}{\eta}\langle v,\tilde{v}\rangle + \frac{1}{\kappa\alpha}s\tilde{s},$$

where $\langle\cdot,\cdot\rangle$ denotes the standard inner product in Euclidean space, and κ and η are two positive parameters. In what follows we use the notation $|(v,s)|^2_{(\theta,\alpha)} := g_{(\theta,\alpha)}((v,s),(v,s))$.

It is straightforward to verify that the gradient of a scalar function $\phi(\theta,\alpha)$ with respect to the inner product g, which we denote by $\overline{\nabla}\phi$, can be written as

$$\overline{\nabla}\phi = (\frac{\eta}{\alpha}\nabla_\theta\phi, \kappa\alpha\partial_\alpha\phi), \tag{B.1}$$

where $\nabla_\theta\phi(\theta,\alpha)$ is the usual gradient of ϕ in the θ variable and $\partial_\alpha\phi(\theta,\alpha)$ is the partial derivative of ϕ with respect to α. Notice that $\overline{\nabla}\phi$ is a vector in $\mathbb{R}^p \times \mathbb{R}$.

Relative to the base metric g in $\Theta\times(0,\infty)$, we define a Wasserstein metric, in dynamic form, over the space of probability measures $\mathcal{P}(\Theta\times[0,\infty)])$. More precisely, for $\sigma,\sigma' \in \mathcal{P}(\Theta\times[0,\infty))$ we consider

$$W^2_{2,g}(\sigma,\hat{\sigma}) = \inf_{\{(\beta_t,\sigma_t)\}_{t\in[0,1]}\in \mathrm{CE}(\sigma,\tilde{\sigma})} \int_0^1 \int |\overline{\nabla}\beta_t(\theta,\alpha)|^2_{\theta,\alpha} d\sigma_t(\theta,\alpha)dt, \tag{B.2}$$

where the set $\mathrm{CE}(\sigma,\tilde{\sigma})$ consists of all solutions $t\in[0,1]\mapsto(\beta_t,\sigma_t)$ to the (intrinsic) *continuity equation*

$$\begin{cases} \partial_t\sigma_t + \overline{\mathrm{div}}(\sigma_t\overline{\nabla}\beta_t) = 0, \\ \sigma(0) = \sigma, \quad \sigma(1) = \sigma'; \end{cases} \tag{B.3}$$

in particular, $\overline{\mathrm{div}}$ denotes the divergence in the space $\Theta\times(0,\infty)$ when endowed with the metric g. In general, equation (B.3) has to be interpreted in the weak sense, i.e., it must hold that

$$\frac{d}{dt}\int \phi(\theta,\alpha)d\sigma_t(\theta,\alpha) = \int g_{(\theta,\alpha)}(\overline{\nabla}\beta_t(\theta,\alpha),\overline{\nabla}\phi(\theta,\alpha))d\sigma(\theta,\alpha)$$

for all $t\in(0,1)$ and all ϕ regular enough test functions.

More than the metric (B.2) itself, from formula (B.2) we are interested in the implicit formal Riemannian structure that we can endow $\mathcal{P}(\Theta\times[0,\infty))$ with and that can be used to motivate, heuristically, gradient descent or projected gradient descent dynamics in the space $\mathcal{P}(\Theta\times[0,\infty)])$. As is standard when interpreting optimal transport from a Riemannian geometric perspective, one can think of the set $\mathcal{T}_\sigma := \{\overline{\nabla}\beta \text{ s.t. } \beta:\Theta\times(0,\infty)\mapsto\mathbb{R}\}$ as a formal tangent plane to the formal manifold $\mathcal{P}(\Theta\times[0,\infty))$ at the point σ, and over this formal tangent plane one can define an inner product $\langle\cdot,\cdot\rangle_\sigma$ according to

$$\langle\overline{\nabla}\beta,\overline{\nabla}\beta'\rangle_\sigma := \int g_{(\theta,\alpha)}(\overline{\nabla}\beta(\theta,\alpha),\overline{\nabla}\beta'(\theta,\alpha))d\sigma_t(\theta,\alpha).$$

Before we finish this section, we state a result that we use in the sequel and that allows us to write the continuity equation (B.3) in terms of basic Euclidean divergence and gradient operators.

Proposition 63 *The intrinsic continuity equation from* (B.3) *can be written, in terms of the Euclidean divergence* $\mathrm{div}_{\theta,\alpha}$ *in* $\mathbb{R}^p \times \mathbb{R}$*, as*

$$\partial_t \sigma_t + \mathrm{div}_{\theta,\alpha}(\sigma_t v_{\sigma_t}) = 0,$$

where v_σ *is the vector field*

$$v_\sigma(\theta,\alpha) := (\frac{\eta}{\alpha}\nabla_\theta \beta(\theta,\alpha), \kappa\alpha\partial_\alpha\beta(\theta,\alpha)).$$

Proof This is a consequence of the following simple observation. For all regular enough test functions ϕ we have

$$\begin{aligned}\frac{d}{dt}\int \phi d\sigma_t &= \int g_{(\theta,\alpha)}(\overline{\nabla}\beta, \overline{\nabla}\phi) d\sigma_t \\ &= \int \left(\frac{\eta}{\alpha}\nabla_\theta\phi\cdot\nabla_\theta\beta + \kappa\alpha\partial_\alpha\phi\partial_\alpha\beta\right) d\sigma_t \\ &= \int \langle \nabla_{\theta,\alpha}\phi, v_\sigma\rangle d\sigma_t,\end{aligned}$$

where in the above we use $\langle\cdot,\cdot\rangle$ to denote the standard Euclidean inner product in $\mathbb{R}^p\times\mathbb{R}$ and $\nabla_{\theta,\alpha}\phi$ to denote the standard gradient in $\mathbb{R}^p\times\mathbb{R}$.

□

B.2. *Vertical and horizontal vector fields in* $\mathcal{P}(\Theta\times[0,\infty))$

We now introduce and discuss some relevant subspaces of the formal tangent plane T_σ. We will use these subspaces later on.

The horizontal space T^h_σ at σ is defined as

$$T^h_\sigma := \{\overline{\nabla}\beta \text{ s.t. } \beta(\theta,\alpha) = \alpha\varphi(\theta) \text{ for some } \varphi\},$$

and the vertical space T^v_σ as

$$T^v_\sigma := \{\overline{\nabla}\beta \text{ s.t. } \langle\overline{\nabla}\beta, \overline{\nabla}\beta'\rangle_\sigma = 0, \quad \text{for all} \quad \overline{\nabla}\beta' \in T^h_\sigma\}.$$

The vertical space T^v_σ represents the directions that infinitesimally leave $\mathcal{F}\sigma$ invariant, while the horizontal space is T^v_σ's orthogonal complement.

Let us denote by $\mathcal{N} = \{\sigma \text{ s.t. } \mathcal{F}\sigma \in \mathcal{P}(\Theta)\}$. For $\sigma\in\mathcal{N}$, we consider the subspace $\mathcal{T}_\sigma\mathcal{N}$ of $\mathcal{T}_\sigma$ defined as

$$\mathcal{T}_\sigma\mathcal{N} := \{\overline{\nabla}\phi \text{ s.t. } \int \alpha\partial_\alpha\phi(\theta,\alpha)d\sigma = 0\}.$$

The subspace $\mathcal{T}_\sigma\mathcal{N}$ can be interpreted as the space of tangent vectors of all curves passing by σ that stay in $\mathcal{N}$.

Remark 64 *The space* $\mathcal{M}_+(\Theta)$ *can be endowed with a metric, the Wasserstein-Fisher-Rao metric, that makes the map* $\mathcal{F}$ *into a Riemannian submersion. Indeed, notice that for two*

potentials of the form $\alpha\varphi(\theta)$ and $\alpha\varphi'(\theta)$ (i.e., two potentials inducing horizontal vector fields at a point σ), we have the identity

$$\langle\overline{\nabla}(\alpha\varphi),\overline{\nabla}(\alpha\varphi')\rangle_\sigma = \int_{\Theta\times[0,\infty)} \alpha(\eta\nabla_\theta\varphi\cdot\nabla_\theta\varphi'+\kappa\varphi\varphi')d\sigma(\theta,\alpha) = \int_\Theta(\eta\nabla_\theta\varphi\cdot\nabla_\theta\varphi'+\kappa\varphi\varphi')d\mathcal{F}\sigma(\theta).$$

In other words, the above inner product in fact does not depend on the specific σ, but only on $\mathcal{F}\sigma$.

We refer the reader to the references [14, 21, 28, 32, 46] for details about the Wasserstein-Fisher-Rao geometry.

B.3. *Gradient flows of lifted energies*

We introduced in section 3.1.1 a projection mapping $\mathcal{F}$ characterised by equation (3.1). We are interested in describing a Riemannian-like metric for the lifted space $\mathcal{P}(\Theta\times[0,\infty))$ with respect to which we will define gradient flows of $\mathcal{J}$.

Let us start by hihlighting that to lift a functional $J:\mathcal{M}_+(\Theta)\to(-\infty,\infty]$ to a functional on $\mathcal{P}(\Theta\times[0,\infty))$, we simply consider the composition of J with the projection map $\mathcal{F}$ as follows:

$$\mathcal{J}(\sigma) := J(\mathcal{F}\sigma), \quad \sigma\in\mathcal{P}(\Theta\times[0,\infty)). \tag{B.4}$$

In particular, if J has the form

$$J(\nu) = \int j(\theta,\nu)d\nu(\theta), \quad \nu\in\mathcal{M}_+(\Theta),$$

then

$$\mathcal{J}(\sigma) = \int \alpha j(\theta,\mathcal{F}\sigma)d\sigma(\theta,\alpha).$$

Given an arbitrary energy $\mathcal{J}:\mathcal{P}(\Theta\times[0,\infty))\to(-\infty,\infty]$ (not necessarily of the form (B.4)), the gradient (descent) flow of $\mathcal{J}$ with respect to the Riemannian geometry introduced in section B.1 takes the form:

$$\partial_t\sigma_t - \overline{\mathrm{div}}(\sigma_t\overline{\nabla}\mathcal{J}_{\sigma_t}) = 0, \tag{B.5}$$

where $\mathcal{J}_\sigma$ is the first variation of $\mathcal{J}$ at the point σ, defined as we did in the beginning of section 1.2. For more details on the interpretation of (B.5) as a gradient flow see Chapter 8.2 in [49].

In case $\mathcal{J}$ has the structure of a lifted energy as in (B.4), its first variation can be computed as follows. Let σ,σ^* and let $\nu=\mathcal{F}\sigma$ and $\nu^*=\mathcal{F}\sigma^*$. Using the linearity of the map $\mathcal{F}$ (which is evident from its definition) we get:

$$\begin{aligned}
\frac{d}{d\varepsilon}|_{\varepsilon=0}\mathcal{J}(\sigma+\varepsilon(\sigma^*-\sigma)) &= \frac{d}{d\varepsilon}|_{\varepsilon=0}J(\mathcal{F}(\sigma+\varepsilon(\sigma^*-\sigma))) \\
&= \frac{d}{d\varepsilon}|_{\varepsilon=0}J(\mathcal{F}\sigma+\varepsilon(\mathcal{F}\sigma^*-\mathcal{F}\sigma)) \\
&= \int_\Theta J_\nu(\theta)d(\nu^*-\nu) \\
&= \int_{\Theta\times[0,\infty)} \alpha J_\nu(\theta)d(\sigma^*-\sigma),
\end{aligned}$$

where J_ν is the first variation of J at the point ν. In other words, the first variation of $\mathcal{J}$ at σ takes the form αJ_ν, where J_ν is the first variation of J; this specific form for $\mathcal{J}_\nu$ should not

be surprising, since the function $\mathcal{J}$ is constant along vertical vector fields and thus its gradient should be a horizontal vector field. Plugging this expression back in (B.5), we conclude that the gradient flow of a lifted energy $\mathcal{J}$ takes the form:

$$\partial\sigma_t - \overline{\mathrm{div}}(\sigma_t \overline{\nabla}(\alpha \mathcal{J}_{\nu_t})) = 0; \quad \mathcal{F}\sigma_t = \nu_t,$$

which, after using Proposition (63), can also be written as

$$\begin{cases} \partial_t \sigma_t - \mathrm{div}_{\theta,\alpha}(\sigma_t v_\sigma) = 0; \\ v_\sigma(\theta,\alpha) = (\eta \nabla_\theta J_{\nu_t}(\theta), \kappa\alpha J_{\nu_t}(\theta)); \quad \nu_t = \mathcal{F}(\sigma_t). \end{cases} \tag{B.6}$$

B.4. *Projected gradients*

In general, σ_t from (B.6) may not belong to $\mathcal{N}$ for $t > 0$, even if initialized at a $\sigma_0 \in \mathcal{N}$. If we want to guarantee that $\nu_t = \mathcal{F}\sigma_t \in \mathcal{P}(\Theta)$ for all t, we must then project the (Wasserstein) gradient of the energy $\mathcal{J}$ driving the dynamics (B.6) onto the subspace $\mathcal{T}_\sigma\mathcal{N}$.

Given σ and $\nu = \mathcal{F}\sigma$, we write the potential αJ_ν as

$$\alpha J_\nu(\theta) = \alpha(J_\nu(\theta) - \int J_\nu(\theta') d\nu(\theta')) + \alpha \int J_\nu(\theta') d\nu(\theta').$$

A direct computation shows that

$$\langle \overline{\nabla}(\alpha \int J_\nu(\theta') d\nu(\theta'))), \overline{\nabla}\phi(\theta,\alpha)\rangle_\sigma = 0,$$

for all $\overline{\nabla}\phi \in \mathcal{T}_\sigma\mathcal{N}$; this means that $\overline{\nabla}(\alpha \int J_\nu(\theta') d\nu(\theta'))) \in \mathcal{T}_\sigma\mathcal{N}^\perp$. Another direct computation shows that $\overline{\nabla}(\alpha(J_\nu(\theta) - \int J_\nu(\theta') d\nu(\theta'))) \in \mathcal{T}_\sigma\mathcal{N}$. From this we can then see that $\overline{\nabla}(\alpha(J_\nu(\theta) - \int J_\nu(\theta') d\nu(\theta')))$ is the projection of $\overline{\nabla}(\alpha \mathcal{J}_\nu)$ onto $\mathcal{T}_\sigma\mathcal{N}$.

Using Proposition (63), we can thus conclude that

$$\begin{cases} \partial_t \sigma_t - \mathrm{div}_{\theta,\alpha}(\sigma_t v_\sigma) = 0; \\ v_\sigma(\theta,\alpha) = (\eta \nabla_\theta J_{\nu_t}(\theta), \kappa\alpha(J_{\nu_t}(\theta) - \int J_{\nu_t}(\theta')\nu_t(\theta'))); \quad \nu_t = \mathcal{F}(\sigma_t), \end{cases} \tag{B.7}$$

represents projected (onto $\mathcal{N}$) gradient descent dynamics of the lifted energy $\mathcal{J}$.

B.5. *An analogous geometric structure for $\mathcal{M}_+(\mathcal{Z}\times\mathcal{Z})$*

There is a similar geometric structure to the one we discussed in the previous sections that the space $\mathcal{M}_+(\mathcal{Z}\times\mathcal{Z})$ can be endowed with. In what follows we use γ to denote elements in the lifted space $\mathcal{P}(\mathcal{Z}\times\mathcal{Z}\times[0,\infty))$ and represent elements in $\mathcal{Z}\times\mathcal{Z}\times[0,\infty)$ with triplets of the form $(z,\tilde{z},\omega)$. The space $\mathcal{P}(\mathcal{Z}\times\mathcal{Z}\times[0,\infty))$ is endowed with a Wasserstein metric just as in (B.2), obtained by changing any appearance of θ with $(z,\tilde{z})$ and any appearance of α with ω. We will use $\mathcal{F}$ (we use the same notation as in section 3.1.1 for simplicity) to denote the projection map $\mathcal{F}:\mathcal{P}(\mathcal{Z}\times\mathcal{Z}\times[0,\infty)) \to \mathcal{M}_+(\mathcal{Z}\times\mathcal{Z})$. An arbitrary functional $J:\mathcal{M}_+(\mathcal{Z}\times\mathcal{Z}) \to (-\infty,\infty]$ can be lifted to $\mathcal{P}(\mathcal{Z}\times\mathcal{Z}\times[0,\infty))$ by composition with $\mathcal{F}$ (we use $\mathcal{J}$ as in the previous sections to denote this composition). The structure of the first variation of $\mathcal{J}$ is ωJ_π, where J_π is the first variation of J at $\pi = \mathcal{F}(\gamma)$.

Since problem (1.1) forces us to restrict to measures π with first marginal equal to μ, we consider evolution equations that can be seen as suitable (projected) gradient *ascent* versions of the gradient *ascent* flow of a lifted energy $\mathcal{J}$ w.r.t. the Wasserstein metric discussed above. Such evolution equation takes the form:

$$\begin{cases} \partial_t \gamma_t + \mathrm{div}_{(z,\tilde{z}),\omega}(\gamma_t v_\gamma) = 0, \\ v_\gamma(z,\tilde{z},\omega) = (0, \eta \nabla_{\tilde{z}} J_\pi(z,\tilde{z}), \kappa\omega\,(J_\pi(z,\tilde{z}) - \int J_\pi(z,\tilde{z}')d\pi_t(\tilde{z}'|z))); \quad \pi_t = \mathcal{F}\gamma_t. \end{cases} \tag{B.8}$$

To motivate the zero in the first component of $v_\gamma(z,\tilde{z},\omega)$, suppose that $t \mapsto \pi_t$ has the form

$$\pi_t = \sum_{i,j=1}^{N} \omega_{ij,t} \delta_{(z_{ij,t},\tilde{z}_{ij,t})},$$

where π_t solves the evolution equation

$$\partial_t \pi_t + \mathrm{div}_{z,\tilde{z}}(\pi_t \vec{V}_t) = 0$$

for some vector field $\vec{V}_t(z,\tilde{z}) = (V_{1,t}(z,\tilde{z}), V_{2,t}(z,\tilde{z}))$ that changes smoothly in time. We claim that if π_t^1 is constant in time, then $V_{1,t}$ must be equal to zero at all points in the support of π_t^1 (and thus of the support of π_0^1). Indeed, it is enough to notice that if $V_{0,t}(z_{ij},\tilde{z}_{ij})$ was different from 0, then for all small enough $t > 0$ we would have that $z_{ij,t}$ is different from $z_{i'j',0}$ for all $i'j'$, implying that the support of π_t^1 is different from the support of π_0^1 for small enough $t > 0$. This would contradict the assumption that π_t^1 was constant in time.

B.6. *Dynamics in lifted and non-lifted space*

We end this appendix by proving the connection between the dynamics in the lifted and non-lifted space.

Proof of Proposition 6 Taking a test function $\phi(\theta)$ we see that

$$\begin{aligned} \frac{d}{dt}\int \phi(\theta) d\nu_t = \frac{d}{dt}\int \alpha\phi(\theta) d\sigma_t(\theta,\alpha) &= \eta_t \int \alpha \nabla_\theta \phi(\theta)\cdot \nabla_\theta \mathcal{U}(\pi_t,\nu_t;\theta) d\sigma_t(\theta,\alpha) \\ &\quad + \kappa \int \alpha(\mathcal{U}_\nu(\pi_t,\nu_t;\theta) - \int \mathcal{U}_\nu(\pi_t,\nu_t;\theta')d\nu_t(\theta'))d\sigma_t(\theta,\alpha) \\ &= \eta_t \int \nabla_\theta \phi(\theta)\cdot \nabla_\theta \mathcal{U}(\pi_t,\nu_t;\theta) d\nu_t(\theta) \\ &\quad + \kappa \int (\mathcal{U}_\nu(\pi_t,\nu_t;\theta) - \int \mathcal{U}_\nu(\pi_t,\nu_t;\theta')d\nu_t(\theta'))d\nu_t(\theta), \end{aligned}$$

which is the weak form of the second equation in (2.3). The equation for π is deduced similarly. □