A Measure Theoretical Approach to the Mean-field Maximum Principle for Training NeurODEs

Benoît Bonnet, Cristina Cipriani, Massimo Fornasier, Hui Huang

▶ **To cite this version:**

Benoît Bonnet, Cristina Cipriani, Massimo Fornasier, Hui Huang. A Measure Theoretical Approach to the Mean-field Maximum Principle for Training NeurODEs. 2021. hal-03289521v1

HAL Id: hal-03289521

https://hal.science/hal-03289521v1

Preprint submitted on 17 Jul 2021 (v1), last revised 17 Oct 2023 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Measure Theoretical Approach to the Mean-field Maximum Principle for Training NeurODEs

Benoît Bonnet*[1], Cristina Cipriani†[2], Massimo Fornasier‡ [3] and Hui Huang§[4]

[1]Inria Paris and Laboratoire Jacques-Louis Lions, Sorbonne Université, Université Paris-Diderot SPC, CNRS, Inria, 75005 Paris, France
[2,3]Technical University Munich, Department of Mathematics, Munich, Germany
[2,3]Munich Data Science Institute, Munich, Germany
[4]University of Calgary, Department of Mathematics and Statistics, Calgary, Canada

Abstract

In this paper we consider a measure-theoretical formulation of the training of NeurODEs in the form of a mean-field optimal control with L^2-regularization of the control. We derive first order optimality conditions for the NeurODE training problem in the form of a mean-field maximum principle, and show that it admits a unique control solution, which is Lipschitz continuous in time. As a consequence of this uniqueness property, the mean-field maximum principle also provides a strong quantitative generalization error for finite sample approximations. Our derivation of the mean-field maximum principle is much simpler than the ones currently available in the literature for mean-field optimal control problems, and is based on a generalized Lagrange multiplier theorem on convex sets of spaces of measures. The latter is also new, and can be considered as a result of independent interest.

Keywords: NeurODEs, Mean-Field Optimal Control, Mean-Field Maximum Principle, Lagrange Multiplier Theorem

Contents

*Email: benoit.a.bonnet@inria.fr
†Email: cristina.cipriani@ma.tum.de
‡Email: massimo.fornasier@ma.tum.de
§Email: hui.huang1@ucalgary.ca

1 Introduction

1.1 Deep learning

Deep learning is an established approach, which performs state-of-the-art on various relevant real-life applications such as speech [43] and image [44, 47] recognition, language translation [67], and also serves as a novel method for scientific computing [13, 33]. In unsupervised machine learning, deep neural networks have shown great success as well, for instance in image and speech generation [57, 58], and in reinforcement learning for solving control problems, such as mastering Atari games [56] or beating human champions in playing Go [64]. Deep learning is about realizing complex tasks as the ones mentioned above, by means of highly parametrized functions, called deep artificial neural networks $\mathcal{N} : \mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$. A classical architecture is the one of feed-forward artificial neural networks of the type

$$\mathcal{N}(x) = \rho(W_L^\top \rho(W_{L-1}^\top \dots \rho(W_1^\top x + \tau_1) \dots) + \tau_L), \tag{1.1}$$

where the function ρ is a scalar activation function acting component-wisely on vectors, the matrices $W_\ell \in \mathbb{R}^{d_{\ell-1} \times d_\ell}$ represent collections of weights, and the vectors $\tau_\ell \in \mathbb{R}^{d_\ell}$ are shifts/biases for each layer $\ell = 1, \dots, L$. Below, we shall denote by $\mathcal{F}(X) = \rho(W^\top X + \tau)$ a generic layer of the network. In practical applications, the number $L \geq 1$ of layers – determining the depth of the network – and the dimensions $d_{\ell-1} \times d_\ell$ of the weight matrices W_ℓ are typically determined by means of heuristic considerations, whereas the weight matrices and the shifts are free parameters which are learned from the training data.

Practical evidences towards certified benchmarks confirm that deep-learning algorithms are able to outperform many previously existing methods. Also, recent mathematical investigations [13, 28–30, 33, 40, 54, 55, 59, 63] have proven that deep artificial networks can approximate high dimensional functions without incurring in the curse of dimensionality, i.e. without needing

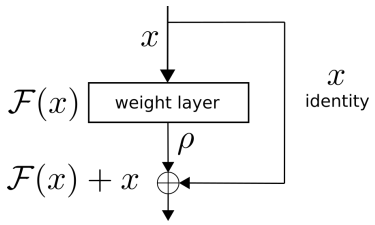

Figure 1: The layer update reads: $X^{n+1} = X^n + \mathcal{F}(X^n)$, see [44].

a number of parameters (here the weights and shifts of the network) that is exponential with respect to the input dimension in order to approximate high-dimensional functions.
While the approximation properties – also called the *expressivity* – of neural networks are becoming more and more understood and transparent [41], the training phase itself, based on suitable optimization processes, remains a (black-)box with some levels of opacity. Recent results are shedding light on this important phase of the employment of neural networks, at least in some simple cases, e.g., of linear neural networks or shallow neural networks, [7, 8, 10, 53, 71].

1.2 Training of deep nets and residual blocks

The method that is most frequently used to train deep neural networks is the so-called *back-propagation of error* [49, 62, 68], which is justified by its tremendous empirical success. All the practical advances recalled above are due to the efficacy of this method. The term backpropagation usually refers to employing stochastic gradient descent or some of its variants [65] to minimize a given loss function (e.g., mean-squared distance, Kullback-Leibler divergence, or Wasserstein distances) over the parameters of the network (weights and biases)[1], usually measuring the misfit of input-output information over a finite number of labeled training samples. On the one hand, the practical efficiency of deep learning is currently ensured in the so-called overparametrized regime by fitting a large amount of data with a larger amount of parameters. On the other hand, solving learning problems with very large numbers of layers gets increasingly harder with the total depth of the network, as the resulting non-convex optimization problems become very high-dimensional.

In the groundbreaking work [44], He et al. showed that the training error of the 56-layer CNN network remains worse than the one of a 20-layer network for the same problem, highlighting an issue which could be blamed either on the optimization function, on initialization of the network, or on the vanishing/exploding gradient phenomenon. The problem of training very deep networks has been alleviated with the introduction of a new neural network layer: the "Residual Block", see Figure 1. According to the analysis conveyed in [45], the use of identity mappings as skip connections and after-addition activations

$$X^{n+1} = X^n + \mathcal{F}(X^n) \tag{1.2}$$

turns out to be beneficial to promote the smoothness of the information propagation. Therein,

[1] In fact, "backpropagation" refers more precisely to a recursive way of applying the chain rule needed to compute the gradient of the loss with respect to weights, but it is often used also to describe any algorithmic optimization which uses such gradients. In many cases, these derivatives are computed using symbolic calculus.

the authors present several 1000-layer deep networks that can be easily trained and achieve improved accuracy. The use of such skip connections with identity mappings presupposes a rectangular shape of the network for which the depths $d_{\ell+1} = d_\ell$ of the layers are all identical.

1.3 NeurODEs and stochastic optimal control

While the arguments in [45] which support the use of residual blocks are based on empirical considerations, a recent line of research has been devoted to a more mathematical (and perhaps more rigorous) formulation of deep neural networks with residual blocks in terms of dynamical systems. In this context, the training of the network can be interpreted as a large optimal control problem, an insight that was proposed independently by Weinan E [31] and Haber-Ruthotto [42]. Later on, this dynamical approach has been greatly popularized in the machine learning community under the name of *NeurODE* by Chen et al. [27], see also [52]. The formulation starts by re-interpreting the iteration (1.2) as a discrete-time Euler approximation [9] of the following dynamical system

$$\dot{X}_t = \mathcal{F}(t, X_t, \theta_t)\,, \tag{1.3}$$

with initial condition $X_0 \in \mathbb{R}^d$. Here, the map $\mathcal{F} : \mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ represents the feed-forwarding dynamics, the parameter $\theta_t \in \mathbb{R}^m$ is a general control variable, which encodes the weights and shifts of the network, i.e. $\theta_t := (W_t, \tau_t)$. A prototypical example is given by

$$\mathcal{F}(t, X_t, \theta_t) = \rho(W_t X_t + \tau_t), \tag{1.4}$$

for instance with an activation function $\rho(\cdot) := \tanh(\cdot)$ acting componentwisely on its entries. In [31, 32], the authors proposed a *stochastic control formulation* of the training of this non-linear process, with a detailed analysis of the related optimality conditions. Therein, both the the Hamilton-Jacobi-Bellman equations [24] – based on the well-known dynamic programming principle – and the Pontryagin Maximum Principle [61] were studied in great generality. From another perspective, several recent works [1, 2, 66] in geometric control theory have aimed at explaining the efficiency of NeurODE in approximating large classes of mappings in terms of controllability properties of such systems in the group of diffeomorphisms.

In this paper, we focus on a particular instance of the more general approach by Weinan E et al. [32], which allows us to derive more specific properties of the control problem, such as the uniqueness and smoothness of solutions to the Pontryagin Maximum Principle, and a strong form of the generalization error estimates. Most importantly, our approach encompasses the prototypical model (1.4) as a possible application. Consider two random variables X_0 and Y_0 which are jointly distributed according to a law $\mu_0(x, y)$, and let us fix the depth $T > 0$ of the time-continuous neural network (1.3). Training this network then amounts to learning the control signals $(\theta_t)_{0\le t\le T}$ in such a way that the terminal output X_T of (1.3) is close to Y_0, with respect to some distortion measure $\ell(\cdot,\cdot) \in \mathcal{C}^2$. A typical choice is $\ell(x, y) =: |x - y|^2$, which is often called the *squared loss function* in the machine learning literature. The stochastic optimal control problem can hence be posed as

$$\inf_{(\theta_t)_{0\le t\le T}} J(\theta) = \inf_{(\theta_t)_{0\le t\le T}} \mathbb{E}_{\mu_0}\left[\ell(X_T, Y_0) + \lambda \int_0^T |\theta_t|^2 \,\mathrm{d}t\right] \tag{1.5}$$

$$\text{subject to (1.3)}\,.$$

The use of a regularization term of the type $\lambda \int_0^T |\theta_t|^2 \, \mathrm{d}t$ helps to promote almost surely finite controls and, as we shall see in more details, it allows for unique solutions to the first order optimality conditions. We will show that the integral boundedness of the controls will result in a controlled Lipschitz continuity of the layer forward map (1.4), which encodes the idea that the corresponding networks are stable with respect to their inputs. Other and more general regularizations are possible of course, but, for the sake of simplicity and clarity in the computations, we restrict our attention to this specific one.

1.4 Measure-theoretical approach to mean-field optimal control

In this paper, we develop a new point of view that is equivalent to that of [32], but which is not based on stochastic control considerations. We start by providing a measure-theoretic reformulation of (1.5)-(1.3), which can be interpreted as a generalized optimal transport problem or mean-field optimal control problem. To this end, let us define a new stochastic process $Z_t := (X_t, Y_t)$ satisfying

$$\dot{X}_t = \mathcal{F}(t, X_t, \theta_t), \quad \dot{Y}_t = 0, \tag{1.6}$$

with initial data (X_0, Y_0) distributed according to μ_0. Let us then denote the law of (X_t, Y_t) by $\mu_t(x, y)$. It is well-known that μ_t satisfies the following partial differential equation

$$\partial_t \mu_t + \nabla_x \cdot (\mathcal{F}(t, x, \theta_t)\mu_t) = 0, \quad \mu_t|_{t=0} = \mu_0 \,, \tag{1.7}$$

understood in the sense of Definition 2.2 below. With this transport equation at hand, we can recast the cost function (1.5) as

$$J(\mu, \theta) := \int_{\mathbb{R}^{2d}} \ell(x, y) \mathrm{d}\mu_T(x, y) + \lambda \int_0^T |\theta_t|^2 \, \mathrm{d}t \,. \tag{1.8}$$

The goal is again is to find the control $(\theta_t)_{0 \le t \le T}$ for which $J(\mu, \theta)$ is minimal when μ satisfies the PDE constraint (1.7). Observe that when the probability measure μ_0 is empirical, i.e.

$$\mu_0 := \mu_0^N = \frac{1}{N} \sum_{i=1}^N \delta_{(X_0^i, Y_0^i)}$$

then the optimal control problem (1.7)-(1.8) reduces to a classical finite particle optimal control problem with ODE constraints.

Optimal control problems over spaces of probability measures of the type (1.7)-(1.8) have been recently explored, mostly in the absence of final-point constraints and in the context of multi-agent interactions. The first contributions on this topic [36, 37] were concerned with the rigorous convergence of classical finite particle optimal controls towards their mean-field counterparts, see also the more recent work [26, 35]. The derivation of first order optimality conditions, i.e., the so-called Pontryagin Maximum Principle (PMP), has been proposed for the first time in [14] based on the leader-follower model studied in [36]. In this work, the mean-field Pontryagin Maximum Principle is derived as limit of the classical finite-particle version. The first general derivation of the PMP for mean-field optimal control problems was obtained in [20], and is based on a careful adaptation of the strategy of needle-variations to the abstract geometric

structure of Wasserstein spaces. These results are further extended in [15] to problems with general final-point and running state constraints. In the latter contribution, the proof strategy combines a finite-dimensional non-smooth multipliers rule and outer-approximations of optimal trajectories by countable families of curves generated using needle-variations. Very recently, a simpler approach has been proposed in [17], by adapting to the notion of multivalued dynamics in Wasserstein space introduced in [16] a methodology originally developed in [38], which relies on suitable linearisations of set-valued maps that produce admissible perturbed trajectories. In [22] a KKT approach is developed in Wasserstein spaces for rather general mean-field optimal control problems with H^1-controls. Therein, both the first order optimality conditions and their relationships with finite particle approximations are derived, along with the corresponding rates of convergence.

We finally point out that a completely different approach to the mean-field PMP was formulated for stochastic optimal control problems in [25] inspired by the theory of mean-field games [48] (see also [3, 12]). Similar methods, based on needle-variations in the space of measures are also leveraged in [32] and [46] for the derivation of the PMP for stochastic control problems of the form (1.3)-(1.5).

1.5 Contributions and organization of the paper

Our contributions can be summarized as follows. From a general standpoint, we start by carefully deriving general first-order optimality conditions for the measure-theoretic formulation of the optimal control of NeurODEs, which include the typical forward mappings (1.4) that appear throughout the literature related to neural networks, for instance with $\rho(\cdot) := \tanh(\cdot)$. As a matter of fact, most of the previous results in the literature do not fully encompass this simple model, as they often require global Lipschitz bounds on the transport field. Let it be noted that while our results may be derived by due adaptation from other approaches developed, e.g., in [22, 32] or [15, 17, 20], we are able to obtain a few stronger properties on the solutions of the optimal control problem than those generally presented in the literature. Moreover, our approach uses a different form of calculus that does not rely on the abstract differential structure of Wasserstein spaces [6], which is of independent interest, see the discussion below. Let us now describe our main results.

In Section 3.1, we thus start by providing a heuristic derivation of the following *mean-field Pontryagin Maximum Principle* ("PMP" in the sequel)

$$\begin{aligned}
&\partial_t \mu_t + \nabla_x \cdot (\mathcal{F}(t,x,\theta_t)\mu_t) = 0, \ \forall t \in (0,T], \quad \mu_t|_{t=0} = \mu_0\,,\\
&\partial_t \psi + \nabla_x \psi \cdot \mathcal{F}(t,x,\theta_t) = 0, \ \forall t \in [0,T), \quad \psi_t|_{t=T} = \ell(x,y)\,,\\
&\theta_t^\top = -\frac{1}{2\lambda}\int_{\mathbb{R}^{2d}} \nabla_x \psi \cdot \nabla_\theta \mathcal{F}(t,x,\theta_t)\mathrm{d}\mu_t(x,y) \text{ for all } t \in [0,T]\,,
\end{aligned} \tag{1.9}$$

which characterizes optimal trajectory-control pairs (μ, θ) for (1.7)-(1.8). In Section 3.2, we proceed to show that the above optimality system is well-posed, and prove in Theorem 3.1 that it admits a unique control solution $\theta^* \in \mathrm{Lip}([0,T];\mathbb{R}^m)$. We also derive in Corollary 3.4 a quantitative *generalization error* for finite samples, which writes

$$\left| \int_{\mathbb{R}^{2d}} \ell(x,y)\,\mathrm{d}\mu_T(x,y) - \frac{1}{N}\sum_{i=1}^N \ell(X_T^i, Y_T^i) \right| \le C W_1(\mu_0^N, \mu_0). \tag{1.10}$$

In particular, (1.10) provides a rate of convergence that depends exclusively on the approximability of μ_0 by empirical measures μ_0^N.

Remark 1.1 (Comparison with the existing literature on generalization errors)**.** *We point out that while the generalization errors established in* [46] *are sharper than those of the present paper (in the sense that they express a rate of convergence in N, which is dimension-independent), this improved stability comes at the price of considering relaxed controls – which are probability measures over $\mathbb{R}^m$ –, that are forced to be non-deterministic by means of entropic regularization terms (see also* [26]*). On the contrary, the generalization errors that we obtain relate to deterministic optimal controls with values in $\mathbb{R}^m$. A similar bound, yielding* (1.10)*, also appears in* [22, Theorem 5.1]*, under the constraint that the control is in a ball of $H^1((0,T),\mathbb{R}^m)$, which is a quite restrictive a priori assumption.*

After establishing the general form of the optimality system along with some of its interesting properties and applications, we move on to the rigorous derivation of the mean-field PMP in Section 3.3. While in [15, 17, 20] the mean-field PMP is established in greater generality – but also with significant technical effort –, we propose in this paper an alternative derivation (very much inspired by the previous work [3] of the third author), which is significantly simpler and hopefully more accessible to non-specialists. The latter can be heuristically explained as follows: under the technical assumption that the optimal control is continuous in time – which is motivated by the well-posedness of (1.9) in $\mathrm{Lip}([0,T];\mathbb{R}^m)$ discussed in Theorem 3.1 –, we prove in Theorem 3.6 that the mean-field PMP (1.9) can be obtained by means of a generalized Lagrange Multiplier Theorem on the convex subset of measures with unit mass. To this end, we use a new form of calculus recently introduced in [4], which is also simpler than the calculus in Wasserstein spaces used in [22]. In contrast to this latter work, our approach is applied in a slightly simpler setting, as the forward and backward equations in (1.9) are linear and decoupled, while therein the authors consider models for which they are non-linear and coupled.
The final and main theoretical result of the paper then reads as follows.

Theorem 1.1. *For any given $T > 0$, let $\mathcal{F}$ satisfy the Assumption 2 and 3, the initial data $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$, and the terminal condition ψ_T satisfy* (3.18)*. Assume further that $\lambda > 0$ is large enough. Then, an admissible control $\theta^* \in L^2([0,T],\mathbb{R}^m)$ fulfills the mean-field PMP* (1.9) *if and only if it is optimal. In addition, such an optimal control θ^* is uniquely determined and Lipschitz continuous.*

We then conclude the paper by presenting numerical experiments to test the proposed novel mean-field Pontryagin maximum principle, where we show the training of simple classification models in $\mathbb{R}^2$. The reason for working on simple two-dimensional examples is to provide full understanding of the properties of the resulting algorithm and a relatively easy reading and visualization of the results.

The paper is organized as follows. In Section 2 we introduce notations and recall a series of preliminary results. In Section 3 we address the derivation of the mean-field maximum principle, we study its well-posedness, and we derive the generalization error estimate (1.10). We present instructive numerical experiments on the solution of the mean-field maximum principle by means

of a shooting method in Section 4. The Appendix contains proofs of auxiliary results, including the proof of a generalized Lagrange multiplier theorem, Theorem 3.5, for constrained problems defined over convex subsets of Banach spaces.

2 Preliminaries and notations

In this section we list some preliminary notations and results from [4, Section 2.1 and Appendix A.1], which will be useful throughout the paper.

2.1 Analysis in measure spaces and optimal transport

We denote by $\mathcal{M}(\mathbb{R}^d)$ the space of signed Borel measures in $\mathbb{R}^d$ with finite total variation. Note that the space $\mathcal{M}(\mathbb{R}^d)$ endowed with the total variation norm

$$\|\mu\|_{TV} := \sup\left\{\int_{\mathbb{R}^d} \varphi \mathrm{d}\mu \;\middle|\; \varphi \in \mathcal{C}_0(\mathbb{R}^d),\ \ \|\varphi\|_\infty \leq 1\right\}, \tag{2.1}$$

is a Banach space, where $\mathcal{C}_0(\mathbb{R}^d)$ represents the set of continuous functions on $\mathbb{R}^d$ which vanish at infinity. By the Riesz-Markov theorem, it is known that $\mathcal{M}(\mathbb{R}^d) = (\mathcal{C}_0(\mathbb{R}^d))'$ can be identified with the topological dual of $\mathcal{C}_0(\mathbb{R}^d)$ [5, Theorem 1.54]. We further denote $\mathcal{M}^+(\mathbb{R}^d)$ the space of positive measures and by $\mathcal{P}(\mathbb{R}^d) \subset \mathcal{M}^+(\mathbb{R}^d)$ the subset of probability measures. Furthermore, $\mathcal{P}_c(\mathbb{R}^d) \subset \mathcal{P}(\mathbb{R}^d)$ represents the set of probability measures with compact support, while $\mathcal{P}_c^N(\mathbb{R}^d) \subset \mathcal{P}_c(\mathbb{R}^d)$ denotes the subset of empirical or atomic probability measures. We will also use the following representation formulas for the subset of measures with zero mass

$$\mathcal{M}_0(\mathbb{R}^d) := \left\{\mu \in (\mathcal{C}_0(\mathbb{R}^d))' \;\middle|\; \mu(\mathbb{R}^d) = \int_{\mathbb{R}^d} 1 \mathrm{d}\mu = 0\right\} =: (\mathcal{C}_0(\mathbb{R}^d))'_0\,, \tag{2.2}$$

and the subset of measures with unit mass

$$\mathcal{M}_1(\mathbb{R}^d) := \left\{\mu \in (\mathcal{C}_0(\mathbb{R}^d))' \;\middle|\; \mu(\mathbb{R}^d) = \int_{\mathbb{R}^d} 1 \mathrm{d}\mu = 1\right\} =: (\mathcal{C}_0(\mathbb{R}^d))'_1\,. \tag{2.3}$$

Moreover, we shall denote by $\mathcal{M}_{0,c}(\mathbb{R}^d), \mathcal{M}_{1,c}(\mathbb{R}^d)$ the corresponding subsets of measures whose supports are compact. One can also note that given $\mu \in \mathcal{M}(\mathbb{R}^d)$, the Jordan decomposition theorem tells us that $\mu = \mu^+ - \mu^-$ and $\|\mu\|_{TV} = \mu^+(\mathbb{R}^d) + \mu^-(\mathbb{R}^d)$, where $\mu^+, \mu^- \in \mathcal{M}^+(\mathbb{R}^d)$.

For the convenience of the reader, we briefly recall the definition of the Wasserstein metrics of optimal transport in the following definition, and refer to [6, Chapter 7] for more details.

Definition 2.1. *Let $1 \leq p < \infty$ and $\mathcal{P}_p(\mathbb{R}^d)$ be the space of Borel probability measures on $\mathbb{R}^d$ with finite p-moment. In the sequel, we endow the latter with the p-Wasserstein metric*

$$W_p^p(\mu,\nu) := \inf\left\{\int_{\mathbb{R}^{2d}} |z-\widehat{z}|^p \; d\pi(z,\widehat{z}) \;\middle|\; \pi \in \Pi(\mu,\nu)\right\} \tag{2.4}$$

where $\Pi(\mu,\nu)$ denotes the set of transport plan between μ and ν, that is the collection of all Borel probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν in the first and second component respectively. The Wasserstein distance can also be expressed as

$$W_p^p(\mu,\nu) = \inf\left\{\mathbb{E}\big[|Z-\widehat{Z}|^p\big]\right\} \tag{2.5}$$

where the infimum is taken over all possible joint distributions of random variables $(Z,\widehat{Z})$ which laws are given by μ and ν respectively.

It is a well-known result in optimal transport theory that when $p = 1$, the following alternative representation holds for the Wasserstein distance

$$W_1(\mu,\nu) = \sup\left\{ \int_{\mathbb{R}^d} \varphi(x)\mathrm{d}(\mu-\nu)(x) \;\middle|\; \varphi \in \mathrm{Lip}(\mathbb{R}^d),\ \mathrm{Lip}(\varphi) \le 1 \right\}, \tag{2.6}$$

by Kantorovich's duality [6, Chapter 6]. Here, $\mathrm{Lip}(\mathbb{R}^d)$ stands for the space of real-valued Lipschitz continuous functions on $\mathbb{R}^d$, and $\mathrm{Lip}(\varphi)$ is the Lipschitz constant of a mapping $\varphi(\cdot)$. In the sequel, we shall also use the signed generalized Wasserstein distance $\mathbb{W}_1^{1,1}$ introduced in [60], which coincides with the bounded Lipschitz distance. Given $\mu, \nu \in \mathcal{M}(\mathbb{R}^d)$, we set

$$\mathbb{W}_1^{1,1}(\mu,\nu) = \sup\left\{ \int_{\mathbb{R}^d} \varphi(x)\mathrm{d}(\mu-\nu)(x) \;\middle|\; \varphi \in \mathrm{Lip}_b(\mathbb{R}^d),\ \|\varphi\|_{\mathrm{Lip}_b} \le 1 \right\}, \tag{2.7}$$

where

$$\|\varphi\|_{\mathrm{Lip}_b} := \sup_{x\in\mathbb{R}^d} |\varphi(x)| + \mathrm{Lip}(\varphi)\,. \tag{2.8}$$

In this context, we also define the bounded Lipschitz norm of a signed measure as

$$\|\mu\|_{BL} := \mathbb{W}_1^{1,1}(\mu, 0)\,. \tag{2.9}$$

2.2 Continuity equations in the space of measures

In what follows, we recollect some basic facts about continuity equations in the space of measures, following [6, Section 8.1].

Definition 2.2. *For any given $T > 0$ and $\theta \in L^2([0,T];\mathbb{R}^m)$, we say that $\mu \in \mathcal{C}([0,T];\mathcal{P}_c(\mathbb{R}^{2d}))$ is a weak solution of (1.7) on the time interval $[0,T]$ if*

$$\int_0^T \int_{\mathbb{R}^{2d}} \Big(\partial_t \psi(t,x,y) + \nabla_x \psi(t,x,y) \cdot \mathcal{F}(t,x,\theta_t) \Big) \mathrm{d}\mu_t(x,y)\,\mathrm{d}t = 0, \tag{2.10}$$

for every $\psi \in \mathcal{C}_c^1((0,T)\times\mathbb{R}^{2d})$.

Remark 2.1. *First, note that (2.10) is equivalent to*

$$\int_{\mathbb{R}^{2d}} \psi(x,y)\mathrm{d}\mu_{t_2}(x,y) - \int_{\mathbb{R}^{2d}} \psi(x,y)\mathrm{d}\mu_{t_1}(x,y) = \int_{t_1}^{t_2} \int_{\mathbb{R}^{2d}} \nabla_x \psi(x,y)\cdot \mathcal{F}(s,x,\theta_s)\mathrm{d}\mu_s(x,y)\,\mathrm{d}s \tag{2.11}$$

for all $\psi \in \mathcal{C}_b^1(\mathbb{R}^{2d})$ and every $t_1, t_2 \in [0,T]$. This follows from the fact that the linear span of functions of the form $\psi(t,x,y) := \eta(t)\xi(x,y)$ with $\eta \in \mathcal{C}_c^1((0,T))$ and $\xi \in \mathcal{C}_c^1(\mathbb{R}^{2d})$ is dense in $\mathcal{C}_c^1((0,T)\times\mathbb{R}^{2d})$ (see e.g. [6, Remark 8.1.1]). Also, observe that since μ is a curve of compactly supported probability measures, we can use the simpler testing space $\mathcal{C}_b^1(\mathbb{R}^{2d})$ instead of $\mathcal{C}_c^1(\mathbb{R}^{2d})$ or $\mathcal{C}_0^1(\mathbb{R}^{2d})$ in (2.11).

We recall below a classical well-posedness result for (1.7) in the Cauchy-Lipschitz setting.

Assumption 1. *For any given $T > 0$, the vector field $\mathcal{F}$ satisfies the following.*

1. *For any fixed $\theta \in \mathbb{R}^m$, the map $(t,x) \mapsto \mathcal{F}(t,x,\theta)$ is a Carathéodory function.*

2. *There exists a map* $h \in L^1([0,T];\mathbb{R}^+)$ *such that for any* $\theta \in \mathbb{R}^m$*, one has*

$$|\mathcal{F}(t,x,\theta)| \leq h(t)(1+|x|), \qquad \text{for a.e. } t \in [0,T] \text{ and every } x \in \mathbb{R}^d\,.$$

In the sequel, we will denote $C_{\mathcal{F},T} := \int_0^T h(t)\,\mathrm{d}t$*.*

3. *For every* $R > 0$*, there exists a constant* $g_R > 0$ *such that for any fixed* $\theta \in \mathbb{R}^m$*, it holds*

$$|\mathcal{F}(t,x_1,\theta) - \mathcal{F}(t,x_2,\theta)| \leq g_R(1+|\theta|)|x_1 - x_2|, \qquad \text{for a.e. } t \in [0,T] \text{ and every } x_1, x_2 \in B(R)\,,$$

and given $\theta \in L^2([0,T];\mathbb{R}^m)$*, we denote* $L_{\mathcal{F},T,R,\|\theta\|_1} := g_R \int_0^T (1+|\theta_t|)\,\mathrm{d}t$*.*

Under the set of assumptions listed above, we can prove the well-posedness of the constraint PDE (1.7), which is stated in the following theorem.

Theorem 2.3. *Consider the initial data* $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$ *with* $\mathrm{supp}(\mu_0) \subset B(R)$ *for some* $R > 0$*, and let* $\mathcal{F}$ *be a map satisfying Assumption 1. Then for any given* $T > 0$ *and* $\theta \in L^2([0,T];\mathbb{R}^m)$*, there exists a unique solution* $\mu \in \mathcal{C}([0,T];\mathcal{P}_c(\mathbb{R}^{2d}))$ *to* (1.7) *in the sense of Definition 2.2. Moreover, there exists a radius* $R_T > 0$ *depending only on* R *and* $C_{\mathcal{F},T}$ *such that*

$$\mathrm{supp}(\mu_t) \subset B(R_T) \qquad \text{for all } t \in [0,T]\,. \tag{2.12}$$

Additionally, for any $s,t \in [0,T]$*, it holds*

$$W_1(\mu_t,\mu_s) \leq C(R, C_{\mathcal{F},T})|t-s|\,. \tag{2.13}$$

Denoting by μ^i *for* $i = 1,2$ *two solutions of* (1.7) *with initial data* μ_0^i *satisfying the above assumptions, we also have the stability estimate*

$$W_1(\mu_t^1,\mu_t^2) \leq e^{L_{\mathcal{F},T,R,\|\theta\|_1}} W_1(\mu_0^1,\mu_0^2) \qquad \text{for all } t \in [0,T]\,, \tag{2.14}$$

where $C_{\mathcal{F},T}$ *and* $L_{\mathcal{F},T,R,\|\theta\|_1}$ *are defined as in Assumption 1.*

The proof of this result is rather standard and we postpone it to the Appendix.

2.3 Differential calculus over convex subsets of Banach spaces

We end this series of preliminaries by introducing a notion of multi-valued Fréchet differential for functions defined on convex sets. To this end, given a convex subset E of a normed vector space X, we define

$$X_E := \mathbb{R}(E-E) = \Big\{ x \in X \;\Big|\; x = \alpha(e_1 - e_2) \text{ with } \alpha \in \mathbb{R} \text{ and } e_1, e_2 \in E \Big\},$$

and given $e \in E$, we denote by $X_e := \mathbb{R}^+(E-e)$ the convex cone of directions at e.

Definition 2.4. *Let* X*,* Y *be normed vector spaces,* $E \subset X$ *be a convex set, and* $f : E \to Y$*. Then,* f *is* <u>*F-differentiable*</u> *at* $e \in E$ *if there exists* $L \in \mathcal{L}(X_E, Y)$ *such that*

$$\lim_{\substack{e' \to e \\ e' \in E}} \frac{\|f(e') - f(e) - L(e'-e)\|_Y}{\|e'-e\|_X} = 0\,, \tag{2.15}$$

where $\mathcal{L}(X_E,Y)$ *denotes the space of bounded linear operators from* X_E *into* Y*.*

Following the previous definition, we define the F-differential of f at $e \in E$ by

$$Df(e) := \left\{ L \in \mathcal{L}(X_E, Y) \;\middle|\; L \text{ satisfies (2.15)} \right\}. \tag{2.16}$$

It can be checked that if X_e is not dense in X_E, then the mapping D is set-valued (similarly to classical convex subdifferentials). However if $v \in \overline{X}_e$, then the evaluation $Df(e)(v)$ is uniquely determined, namely it does not depend on the choice of L in $Df(e)$, and in this case we will slightly abuse the notation and write $Df(e)(v)$ to mean $L(v)$ for any $L \in Df(e)$. By a density argument, each $L \in Df(e)$ can be uniquely extended to an operator $\overline{L}$ in $\mathcal{L}(\overline{X}_E, Y)$. We will then say that $f \in \mathcal{C}^1(E;Y)$ if f is F-differentiable at each $e \in E$, and there exists a selection $e \in E \mapsto L_e \in Df(e)$ such that

$$e \mapsto L_e \quad \text{is continuous from } E \text{ into } \mathcal{L}(X_E, Y)\,, \tag{2.17}$$

where $\mathcal{L}(X_E, Y)$ is endowed with the distance induced by the standard operator norm.

Definition 2.5. *Let X, Y be normed vector spaces, $E \subset X$ be a convex set, and $f : E \to Y$. Then, f is G-differentiable at $e \in E$ if the directional right derivatives*

$$df(e,v) := \lim_{h \to 0^+} \frac{f(e+hv) - f(e)}{h}, \tag{2.18}$$

exist in Y for all $v \in X_e$.

Remark 2.2. *Obviously if f is F-differentiability at some $e \in E$, then it is G-differentiability as well with $df(e,v) = Df(e)(v)$ for all $v \in X_e$.*

We shall also use the following lemma as a criterion for $\mathcal{C}^1$ regularity, see [4, Lemma A.4].

Lemma 2.1. *Let $f : E \to F$ be a continuous map and suppose that there exists a continuous application*

$$e \in E \mapsto L_e \in \mathcal{L}(X_E, Y), \tag{2.19}$$

such that $df(e,v) = L_e v$ for all $e \in E$ and any $v \in X_e$. Then $f \in \mathcal{C}^1(E;Y)$ and $e \mapsto L_e \in Df(e)$ is an admissible selection.

3 Mean-Field Maximum Principle

In this section, we investigate first-order optimality conditions for the mean-field optimal control problem (1.7)-(1.8), which take the form of a Pontryagin Maximum Principle ("PMP" in the sequel). Their derivation – which is based on a Lagrange multiplier rule for the convex calculus introduced in Section 2 – is heuristically presented in Section 3.1. After studying the well-posedness of the optimality system in Section 3.2, we proceed to rigorously establish the PMP throughout Section 3.3.

3.1 Formal derivation of a Lagrangian formulation

We start this section by providing a formal derivation of the mean-field PMP. To this end, we first introduce the Lagrangian of the mean-field optimal control problem (1.7)-(1.8), defined by

$$\begin{aligned}\mathcal{L}(\mu,\theta,\psi) = &\int_{\mathbb{R}^{2d}} \ell(x,y)\mathrm{d}\mu_T(x,y) + \lambda\int_0^T |\theta_t|^2\,\mathrm{d}t\\ &+\int_{\mathbb{R}^{2d}}\psi(0,x,y)\mathrm{d}\mu_0(x,y) - \int_{\mathbb{R}^{2d}}\psi(T,x,y)\mathrm{d}\mu_T(x,y)\\ &+\int_0^T\int_{\mathbb{R}^{2d}}\Big(\partial_t\psi(t,x,y)+\nabla_x\psi(t,x,y)\cdot\mathcal{F}(t,x,\theta_t)\Big)\mathrm{d}\mu_t(x,y)\,\mathrm{d}t\,. \end{aligned}\tag{3.1}$$

Next, we compute its functional derivatives with respect to the curves μ and θ, namely

$$\frac{\delta\mathcal{L}}{\delta\mu_t} = \begin{cases} 0, & t=0,\ \text{(the initial variance is zero)}\\ \partial_t\psi+\nabla_x\psi\cdot\mathcal{F}, & 0<t<T,\\ \ell-\psi_T, & t=T,\end{cases}$$

and

$$\frac{\delta\mathcal{L}}{\delta\theta_t} = 2\lambda\theta_t^\top + \int_{\mathbb{R}^{2d}}\nabla_x\psi\cdot\nabla_\theta\mathcal{F}(t,x,\theta_t)\mathrm{d}\mu_t(x,y)\,.$$

for almost every $t\in[0,T]$. Then given an optimal trajectory-control pair (μ^*,θ^*) for the problem (1.7)-(1.8), we will show that there exists a Lagrange multiplier ψ^* such that

$$\frac{\delta\mathcal{L}}{\delta\mu}(\mu^*,\theta^*,\psi^*)=0 \qquad\text{and}\qquad \frac{\delta\mathcal{L}}{\delta\theta}(\mu^*,\theta^*,\psi^*)=0\,.\tag{3.2}$$

This will in particular provide us with the following backward adjoint dynamics

$$\partial_t\psi^*+\nabla_x\psi^*\cdot\mathcal{F}(t,x,\theta_t^*)=0,\tag{3.3}$$

subject to the terminal condition $\psi_T^*=\ell$, along with the fixed-point equation

$$2\lambda\theta_t^{*\top}+\int_{\mathbb{R}^{2d}}\nabla_x\psi^*\cdot\nabla_\theta\mathcal{F}(t,x,\theta_t^*)\mathrm{d}\mu_t^*(x,y)=0\,,\tag{3.4}$$

characterizing the optimal controls, where the curve μ^* satisfies the native forward dynamics

$$\partial_t\mu_t^*+\nabla_x\cdot(\mathcal{F}(t,x,\theta_t^*)\mu_t^*)=0,\qquad \mu_t^*|_{t=0}=\mu_0\,.\tag{3.5}$$

We will see below that (3.3) is understood in the sense of (3.75), and that (3.4) is understood in the sense of (3.76). Throughout this section, we will use the two following sets of of assumptions, which will be applied to derive different results.

Assumption 2. *For any given $T>0$, the vector field $\mathcal{F}$ satisfies the following.*

1. *For any fixed $\theta\in\mathbb{R}^m$, the map $(t,x)\mapsto\mathcal{F}(t,x,\theta)\in\mathbb{R}^d$ is continuous.*
2. *There exists a constant $C_\mathcal{F}>0$ such that for every $\theta\in\mathbb{R}^m$, it holds*

$$|\mathcal{F}(t,x,\theta)|\le C_\mathcal{F}(1+|x|),\quad \text{for a.e. } t\in[0,T] \text{ and every } x\in\mathbb{R}^d\,.$$

3. *There exists a constant $L_{\mathcal{F}} > 0$ such that for every $\theta \in \mathbb{R}^m$, it holds*

$$|\mathcal{F}(t,x_1,\theta)-\mathcal{F}(t,x_2,\theta)| \leq L_{\mathcal{F}}(1+|\theta|)|x_1-x_2|, \quad \text{for a.e. } t \in [0,T] \text{ and every } x_1,x_2 \in \mathbb{R}^d,$$

and we denote $L_{\mathcal{F},T,\|\theta\|_1} = L_{\mathcal{F}} \int_0^T (1+|\theta_t|)\mathrm{d}t$

4. *For all $(t,x) \in [0,T] \times \mathbb{R}^d$, the map $\theta \mapsto \mathcal{F}(t,x,\theta)$ is twice differentiable. Moreover for each $R > 0$, there exists a constant $C(d,m,R,T) > 0$ such that*

$$\|\nabla_\theta \mathcal{F}\|_{\mathcal{C}([0,T]\times B(R)\times\mathbb{R}^m;\mathbb{R}^d)} + \|\nabla^2_\theta \mathcal{F}\|_{\mathcal{C}([0,T]\times B(R);L^\infty(\mathbb{R}^m;\mathbb{R}^d))} \leq C(d,m,R,T)\,.$$

Assumption 3. *For any given $T > 0$ and $R > 0$, the vector field $\mathcal{F}$ satisfies the following.*

1. *The map $x \in \mathbb{R}^d \mapsto \mathcal{F}(t,x,\theta)$ is of class $\mathcal{C}^2$ all times $t \in [0,T]$ and any $\theta \in \mathbb{R}^m$, and for each $x \in B(R)$ it holds*

$$|\nabla_x \cdot \nabla_\theta \mathcal{F}(t,x,\theta)| + |\nabla_x \mathcal{F}(t,x,\theta)| + |\nabla^2_x \mathcal{F}(t,x,\theta)| \leq C(d,m,T,R,|\theta|)\,; \tag{3.6}$$

2. *For any $\theta^1,\theta^2 \in \mathbb{R}^m$, every $s,t \in [0,T]$ and all $x \in B(R)$, it holds*

$$|\mathcal{F}(t,x,\theta^1) - \mathcal{F}(s,x,\theta^2)| \leq C(d,m,R)\big(|t-s| + |\theta^1-\theta^2|\big)\,; \tag{3.7}$$

3. *For all fixed θ and $t \in [0,T]$, it holds*

$$|\nabla_\theta \mathcal{F}(t,x,\theta) - \nabla_\theta \mathcal{F}(t,y,\theta)| \leq C(d,m,R,|\theta|)|x-y|, \tag{3.8}$$

for every $x,y \in B(R)$.

Remark 3.1. *Here, we check that the sets of assumptions listed above include relevant NeurODE models, and in particular that they hold for the popular subclass of feed-forwarding dynamics (1.4) given by $\mathcal{F}(t,x,\theta) := \tanh(\theta x)$ with $\theta \in \mathbb{R}^m = \mathbb{R}^{d\times d}$ and $x \in \mathbb{R}^d$.*

In this case, we have $\nabla_x \mathcal{F} = \mathrm{diag}(\mathcal{F}')\theta$ where $\mathcal{F}'(x,\theta) = (1-\tanh^2(\theta x)) \in \mathbb{R}^d$, which implies that $|\nabla_x \mathcal{F}| \leq |\theta|$. Furthermore, one has $\nabla^2_x \mathcal{F} = \frac{\partial(\mathrm{diag}(\mathcal{F}'))}{\partial x}\theta$, where $\frac{\partial(\mathrm{diag}(\mathcal{F}'))}{\partial x}$ is a $d \times d \times d$ tensor with

$$\frac{\partial(\mathrm{diag}(\mathcal{F}'))}{\partial x_i} = \mathrm{diag}(\mathcal{F}'')\theta^i\,,$$

where $\mathcal{F}'' = -2\tanh(\theta x)(1-\tanh^2(\theta x)) \in \mathbb{R}^d$ and $\theta^i \in \mathbb{R}^d$ is the i-th column of θ. This yields the following estimate $|\nabla^2_x \mathcal{F}| \leq C(d)|\theta|^2$ on the second order derivative of $\mathcal{F}$ with respect to x.

Concerning (4), one can check that $\nabla_\theta \mathcal{F} = \mathrm{diag}(\mathcal{F}')\frac{\partial(\theta x)}{\partial\theta}$, where $\frac{\partial(\theta x)}{\partial\theta}$ is a $d \times d \times d$ tensor with

$$\frac{\partial(\theta x)}{\partial\theta_{ij}} = x_j e_i, \tag{3.9}$$

where e_i is the i-th element of the canonical basis of $\mathbb{R}^d$. Hence it holds that

$$\max_{0\leq t\leq T,\, x\in B(R_T),\, \theta\in\mathbb{R}^{d\times d}} |\nabla_\theta \mathcal{F}| \leq C(d)|x| \leq C(d,|U|),$$

so that (4) is fulfilled as well.

In addition, observe that $\nabla_\theta \mathcal{F} = \mathrm{diag}(\mathcal{F}')\frac{\partial(\theta x)}{\partial \theta}$*, where* $\frac{\partial(\theta x)}{\partial \theta}$ *is a* $d \times d \times d$ *tensor as in* (3.9)*, which leads to the estimates*

$$|\nabla_\theta \mathcal{F}| \leq C(d)|x| \quad \textit{and} \quad |\nabla_\theta^2 \mathcal{F}| = \left|\nabla_\theta(\mathrm{diag}(\mathcal{F}'))\frac{\partial(\theta x)}{\partial \theta}\right| \leq C(d)|x|^2.$$

Moreover, one can check that

$$\nabla_x \cdot \nabla_\theta \mathcal{F} = \mathrm{diag}(\mathcal{F}')\nabla_x \cdot \left(\frac{\partial(\theta x)}{\partial \theta}) + \frac{\partial(\theta x)}{\partial \theta}\nabla_x \cdot (\mathrm{diag}(\mathcal{F}')\right).$$

Since $\left|\mathrm{diag}(\mathcal{F}')\nabla_x \cdot (\frac{\partial(\theta x)}{\partial \theta})\right| \leq C(d)$ *and* $\left|\frac{\partial(\theta x)}{\partial \theta}\nabla_x \cdot (\mathrm{diag}(\mathcal{F}'))\right| \leq C(d)|x||\theta|$*, we obtain*

$$|\nabla_x \cdot \nabla_\theta \mathcal{F}| \leq C(d)(1 + 2|x||\theta|).$$

Similarly we have $|\nabla_x \nabla_\theta \mathcal{F}| \leq C(d, |x|, |\theta|)$*, which leads to* $|\nabla_\theta \mathcal{F}(t, x, \theta) - \nabla_\theta \mathcal{F}(t, y, \theta)| \leq C(d, |x|, |y|, |\theta|)|x - y|$*. Lastly, it easy to check that* $|\mathcal{F}(t, x, \theta_t) - \mathcal{F}(s, x, \theta_s)| \leq C(d)|x||\theta_t - \theta_s|$*, which ends the verification the Assumptions 2.*

3.2 Well-posedness of the maximum principle

This section is devoted to the proof of the existence and uniqueness of a solution $(\mu^*, \theta^*, \psi^*) \in \mathcal{C}([0,T]; \mathcal{P}_c(\mathbb{R}^{2d})) \times \mathrm{Lip}([0,T]; \mathbb{R}^m) \times \mathcal{C}^1([0,T]; \mathcal{C}_c^2(\mathbb{R}^{2d}))$ to the first order optimality system

$$\partial_t \mu_t^* + \nabla_x \cdot (\mathcal{F}(t, x, \theta_t^*)\mu_t^*) = 0, \ \forall t \in (0,T], \quad \mu_t^*|_{t=0} = \mu_0\,, \tag{3.10}$$

$$\partial_t \psi^* + \nabla_x \psi^* \cdot \mathcal{F}(t, x, \theta_t^*) = 0, \ \forall t \in [0,T), \quad \psi_t^*|_{t=T} = \ell\,, \tag{3.11}$$

$$\theta_t^{*\top} = -\frac{1}{2\lambda}\int_{\mathbb{R}^{2d}} \nabla_x \psi^* \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^*)\mathrm{d}\mu_t^*(x, y), \quad \forall t \in [0,T]\,. \tag{3.12}$$

The main result of this section can then be stated as follows.

Theorem 3.1. *For any given* $T > 0$*, let* $\mathcal{F}$ *satisfy the Assumption 2 and 3, take an initial data* $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$ *and a terminal condition* ψ_T *satisfying* (3.18)*. Then, for* $\lambda > 0$ *large enough, there exists a triple* $(\mu^*, \theta^*, \psi^*) \in \mathcal{C}([0,T]; \mathcal{P}_c(\mathbb{R}^{2d})) \times \mathrm{Lip}([0,T]; \mathbb{R}^m) \times \mathcal{C}^1([0,T]; \mathcal{C}_c^2(\mathbb{R}^{2d}))$ *solving* (3.10)-(3.12)*. Moreover, the control solution* θ^* *is unique in* $\Gamma_C \subset L^2([0,T]; \mathbb{R}^m)$ *defined as in* (3.14)*, and* $\psi^* \in \mathcal{C}^1([0,T]; \mathcal{C}_c^2(\mathbb{R}^{2d}))$ *is in characteristic form.*

Remark 3.2. *If there exists an optimal control* $\theta^* \in L^2([0,T]; \mathbb{R}^m)$ *satisfying the maximum principle* (3.10)-(3.12)*, then the uniqueness result in Theorem 3.1 ensures that* θ^* *coincides with a Lipschitz continuous function almost everywhere. This means that in such a case there exists a smooth optimal control* $\theta^* \in \mathrm{Lip}([0,T]; \mathbb{R}^m)$*.*

To prove Theorem 3.1, we consider a compact and convex subset $\Gamma_{M,C}$ of the subspace $\mathrm{Lip}([0,T]; \mathbb{R}^m) \subset \mathcal{C}([0,T]; \mathbb{R}^m)$, defined by

$$\Gamma_{M,C} := \left\{\theta \in \mathcal{C}([0,T]; \mathbb{R}^m) \ \middle| \ |\theta_t - \theta_s| \leq M|t - s|, \ \|\theta\|_\infty \leq C_\Gamma\right\}. \tag{3.13}$$

for some constants $M, C_\Gamma > 0$. We will also make use of the following ball in $L^2([0,T]; \mathbb{R}^m)$

$$\Gamma_C := \left\{\theta \in L^2([0,T]; \mathbb{R}^m) \mid \|\theta\|_2 \leq C_\Gamma T^{\frac{1}{2}}\right\}. \tag{3.14}$$

One can easily notice that $\Gamma_{M,C} \subset \Gamma_C$.

Using arguments that are similar to those of Theorem 2.3, one can show the following result.

Proposition 3.2. *Consider an initial data $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$ with $\mathrm{supp}(\mu_0) \subset B(R)$ for some $R > 0$, and let $\mathcal{F}$ satisfy Assumption 2. Then for any $T > 0$ and $\theta \in \Gamma_{M,C}$, there exists a unique solution $\mu^\theta \in \mathcal{C}([0,T];\mathcal{P}_c(\mathbb{R}^{2d}))$ to (3.10) in the sense of Definition 2.2. Moreover, there exists some $R_T > 0$ depending only on R and $C_{\mathcal{F},T}$, such that*

$$\mathrm{supp}(\mu_t^\theta) \subset B(R_T) \qquad \text{for all } t \in [0,T]\,. \tag{3.15}$$

Additionally, for any $s,t \in [0,T]$, it holds

$$W_1(\mu_t^\theta, \mu_s^\theta) \leq C(R, C_{\mathcal{F},T})|t-s|\,. \tag{3.16}$$

If $\mu^{\theta,i}$, $i = 1,2$ are two solutions with initial data μ_0^i satisfying the above assumptions, we have

$$W_1(\mu_t^{\theta,1}, \mu_t^{\theta,2}) \leq e^{L_{F,T,C_\Gamma}} W_1(\mu_0^1, \mu_0^2) \quad \text{for all } t \in [0,T]\,. \tag{3.17}$$

Here $C_{\mathcal{F},T}$ and $L_{\mathcal{F},T,C_\Gamma}$ are defined as in Assumption 2 by replacing $\|\theta\|_1$ by $C_\Gamma T$.

In what follows, we will only be interested in what is happening inside the supports of μ^θ for $\theta \in \Gamma_{M,C}$. Therefore, we shall recast the terminal condition in (3.11) as $\psi_T \in \mathcal{C}_c^2(\mathbb{R}^{2d})$ with

$$\mathrm{supp}(\psi_T) = B(R_T) \quad \text{and} \quad \psi_T(x,y) = \ell(x,y) \quad \text{for all } x,y \in B(R_T)\,. \tag{3.18}$$

In this context, we are able to derive the following norm estimate on ψ^θ.

Proposition 3.3. *Suppose that $\mathcal{F}$ satisfies Assumption 2. Then for any $T > 0$ and $\theta \in \Gamma_{M,C}$, there exists a unique characteristic solution $\psi^\theta \in \mathcal{C}^1([0,T];\mathcal{C}_c^2(\mathbb{R}^{2d}))$ to the equation (3.11) which terminal condition satisfies (3.18). Moreover it holds*

$$\left\|\psi_t^\theta\right\|_{\mathcal{C}_c^2(\mathbb{R}^{2d})} \leq C(R', T, C_\Gamma, C_{\mathcal{F},T}, L_{F,T,C_\Gamma}) \left\|\psi_T\right\|_{\mathcal{C}^2(B(R_T))}\,, \tag{3.19}$$

for all times $t \in [0,T]$. Here the supports of ψ_t^θ satisfies the inclusion $\mathrm{supp}(\psi_t^\theta) \subset B(R'_T)$ where $R' = R + (R + C_{\mathcal{F},T}T)e^{C_{\mathcal{F},T}T}$.

The results of Proposition 3.3 are rather classical, hence their proof is reported in the Appendix.

Remark 3.3. *Here, the fact that ψ^θ is a characteristic solution means that it is obtained via the characteristic method, and is of the form $\psi^\theta(t,x,y) = \psi_T(\mathbf{\Phi}^\theta_{(T,t)}(x,y))$. Here, we denoted by $(\mathbf{\Phi}^\theta_{(\tau,t)})_{\tau,t\in[0,T]}$ the flow maps defined as in (4.8) with $\mathcal{F}(t,x) := \mathcal{F}(t,x,\theta_t)$. Characteristic solutions to (3.12) are unique because of the way they depends on terminal condition and (3.19). Note here that we do not claim to have general uniqueness in $\mathcal{C}^1([0,T];\mathcal{C}_c^2(\mathbb{R}^{2d}))$ for (3.12), i.e., there may exist $\mathcal{C}^1([0,T];\mathcal{C}_c^2(\mathbb{R}^{2d}))$ solutions that are not in the characteristic form. In what follows however, we will only consider characteristic solutions.*

Proof of Theorem 3.1. The existence of optimal controls θ^* in $\Gamma_{M,C}$ is based on the Schauder fixed point theorem [39, Theorem 11.1]. Then, the uniqueness will be obtained by additionally showing that the underlying fixed-point map is in fact a contraction in Γ_C.

• *(Existence in $\Gamma_{M,C}$)* For any $\theta \in \Gamma_{M,C}$, denote by $\mu^\theta \in \mathcal{C}([0,T];\mathcal{P}_c(\mathbb{R}^{2d}))$ the corresponding solution of (3.10) and by $\psi^\theta \in \mathcal{C}^1([0,T];\mathcal{C}^2_c(\mathbb{R}^{2d}))$ the unique characteristic solution of (3.11). Then, we introduce the continuous mapping $\Lambda:\ \Gamma_{M,C} \to \mathcal{C}([0,T];\mathbb{R}^m)$, defined by

$$\Lambda(\theta)(t) = -\frac{1}{2\lambda}\int_{\mathbb{R}^{2d}} \nabla_x \psi_t^\theta \cdot \nabla_\theta \mathcal{F}(t,x,\theta_t)\mathrm{d}\mu_t^\theta(x,y), \tag{3.20}$$

for every $\theta \in \Gamma_{M,C}$ and all times $t \in [0,T]$. We start by checking that $\Lambda(\Gamma_{M,C}) \subset \Gamma_{M,C}$ for λ large enough. On the one hand, it follows from *(3)* in Assumption 2 and (3.19) that

$$\begin{aligned}
|\Lambda(\theta)(t)| &\le \frac{1}{2\lambda}\int_{B(R_T)} \big|\nabla_x \psi_t^\theta \cdot \nabla_\theta \mathcal{F}(t,x,\theta_t)\big|\mathrm{d}\mu_t^\theta(x,y) \\
&\le \frac{1}{2\lambda} C(R_T,T) \sup_{t\in[0,T]} \big\|\psi_t^\theta\big\|_{\mathcal{C}^1(B(R'_T))} \\
&\le \frac{1}{2\lambda} C(R_T,T) C(R'_T,T,C_\Gamma,C_{\mathcal{F},T},L_{F,T,C_\Gamma}) \,\|\psi_T\,\|_{\mathcal{C}^1(B(R_T))}\ ,
\end{aligned}$$

for all $t \in [0,T]$, with the explicit diameter $R'_T = R + (R + C_{\mathcal{F},T}T)e^{C_{\mathcal{F},T}T}$. Hence, upon choosing a parameter $\lambda > 0$ that is large enough, it holds

$$\|\Lambda(\theta)\,\|_{L^\infty([0,T];\mathbb{R}^m)} \le C_\Gamma\,. \tag{3.21}$$

On the other hand, it holds for any $s,t \in [0,T]$ that

$$\begin{aligned}
|\Lambda(\theta)(t) - \Lambda(\theta)(s)| \le{}& \frac{1}{2\lambda}\left|\int_{B(R_T)} (\nabla_x \psi_t^\theta - \nabla_x \psi_s^\theta)\cdot \nabla_\theta \mathcal{F}(t,x,\theta_t)\mathrm{d}\mu_t^\theta(x,y)\right| \\
&+ \frac{1}{2\lambda}\left|\int_{B(R_T)} \nabla_x \psi_s^\theta \cdot (\nabla_\theta \mathcal{F}(t,x,\theta_t) - \nabla_\theta \mathcal{F}(s,x,\theta_s))\mathrm{d}\mu_t^\theta(x,y)\right| \\
&+ \frac{1}{2\lambda}\left|\int_{B(R_T)} \nabla_x \psi_s^\theta \cdot \nabla_\theta \mathcal{F}(s,x,\theta_s)(\mathrm{d}\mu_t^\theta - \mathrm{d}\mu_s^\theta)(x,y)\right| \\
=:{}& I_1 + I_2 + I_3\,.
\end{aligned}$$

Using the fact that $\psi^\theta \in \mathcal{C}^1([0,T];\mathcal{C}^2_c(\mathbb{R}^{2d}))$ along with *(3)* in Assumption 2, one can see that

$$I_1 \le \frac{1}{2\lambda} C(R_T,T)|t-s|\,, \tag{3.22}$$

for all $s,t \in [0,T]$. Furthermore, it follows from assumption (3.7) and the estimate (3.19) that

$$\begin{aligned}
I_2 &\le \frac{1}{2\lambda} C(R_T) \sup_{t\in[0,T]} \big\|\psi_t^\theta\big\|_{\mathcal{C}^1(B(R'_T))}\big(|t-s| + |\theta_t - \theta_s|\big) \\
&\le \frac{1}{2\lambda} C(R_T) C(R'_T,T,C_\Gamma,C_{\mathcal{F},T},L_{F,T,C_\Gamma}) \,\|\psi_T\,\|_{\mathcal{C}^1(B(R_T))}\, M|t-s|\,,
\end{aligned} \tag{3.23}$$

with the diameter expression $R'_T = R + (R + C_{\mathcal{F},T}T)e^{C_{\mathcal{F},T}T}$. Lastly by (2.6) one has

$$I_3 \le \frac{1}{2\lambda}\mathrm{Lip}\big(\nabla_x \psi_s^\theta \cdot \nabla_\theta \mathcal{F}(s,\cdot,\theta_s)\,;B(R_T)\big)W_1(\mu_t,\mu_s)\,, \tag{3.24}$$

and notice that

$$\begin{aligned}
\mathrm{Lip}\big(\nabla_x \psi_s^\theta \cdot \nabla_\theta \mathcal{F}(s,\cdot,\theta_s)\,;B(R_T)\big) \le{}& C(R'_T,T,C_\Gamma,C_{\mathcal{F},T},L_{F,T,C_\Gamma})\,\|\psi_T\,\|_{\mathcal{C}^2(B(R_T))} \\
&\times \Big(\|\nabla_\theta \mathcal{F}(s,\cdot,\theta_s)\|_{L^\infty(B(R_T))} + \mathrm{Lip}(\nabla_\theta \mathcal{F}(s,\cdot,\theta_s)\,;B(R_T))\Big) \\
\le{}& C(R'_T,T,C_\Gamma,C_{\mathcal{F},T},L_{F,T,C_\Gamma},R_T)\,\|\psi_T\,\|_{\mathcal{C}^2(B(R_T))}\ ,
\end{aligned}$$

where we have used (3.19) and *(3)* in Assumption 3. This combined with (3.16) thus yields

$$I_3 \leq \frac{1}{2\lambda} C(R_T', T, C_\Gamma, C_{\mathcal{F},T}, L_{F,T,C_\Gamma}, R_T) \, \|\psi_T\|_{\mathcal{C}^2(B(R_T))} \, |t-s| \, . \tag{3.25}$$

Collecting estimates (3.22), (3.23) and (3.25), we deduce that for $\lambda > 0$ large enough, it holds

$$|\Lambda(\theta)(t) - \Lambda(\theta)(s)| \leq M|t-s| \, . \tag{3.26}$$

Thus, we have proven that $\Lambda(\Gamma_{M,C}) \subset \Gamma_{M,C}$ when $\lambda > 0$ is taken to be sufficiently large. Hence by Schauder's fixed point theorem, the mapping Λ has at least a fixed point θ^*, namely

$$\theta^{*\top} = -\frac{1}{2\lambda} \int_{\mathbb{R}^{2d}} \nabla_x \psi_t^{\theta^*} \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^*) \mathrm{d}\mu_t^{\theta^*}(x, y) \, . \tag{3.27}$$

This concludes the existence part of the proof.

• *(Uniqueness in Γ_C)* Our goal now is to prove that Λ is a contraction over $\Gamma_{M,C}$ with respect to the L^2-norm, so that that the fixed point $\theta^* \in \Gamma_{M,C}$ is actually unique in Γ_C. Indeed assuming that Λ had two distinct fixed points θ^1 and θ^2, it would hold

$$\|\theta_t^1 - \theta_t^2\|_2 = \|\Lambda(\theta^1)(t) - \Lambda(\theta^2)(t)\|_2 \leq \kappa \|\theta_t^1 - \theta_t^2\|_2 \, ,$$

leads to a contradiction since the contraction constant satisfies $0 \leq \kappa < 1$. In order to prove the contractivity of Λ, we start by fixing $t \in [0, T]$ and denote by $\mu^{\theta^1}, \mu^{\theta^2}$ two solutions of (3.10) driven by θ^1, θ^2 respectively, and with the same initial condition μ_0. Similarly, denote by $\psi^{\theta^1}, \psi^{\theta^2}$ the solutions of (3.11) generated by θ^1, θ^2 with the same terminal condition ψ_T. Then

$$\begin{aligned}
&|\Lambda(\theta^1)(t) - \Lambda(\theta^2)(t)| \\
&= \frac{1}{2\lambda} \left| \int_{\mathbb{R}^{2d}} \nabla_x \psi_t^{\theta^1} \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^1) \mathrm{d}\mu_t^{\theta^1}(x, y) - \int_{\mathbb{R}^{2d}} \nabla_x \psi_t^{\theta^2} \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^2) \mathrm{d}\mu_t^{\theta^2}(x, y) \right|
\end{aligned}$$

which can in turn be estimated by inserting suitable crossed terms as

$$\begin{aligned}
|\Lambda(\theta^1)(t) - \Lambda(\theta^2)(t)| &\leq \frac{1}{2\lambda} \left| \int_{\mathbb{R}^{2d}} \nabla_x \psi_t^{\theta^1} \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^1) (\mathrm{d}\mu_t^{\theta^1} - \mathrm{d}\mu_t^{\theta^2})(x, y) \right| \\
&\quad + \frac{1}{2\lambda} \left| \int_{\mathbb{R}^{2d}} \Big(\nabla_x \psi_t^{\theta^1} \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^1) - \nabla_x \psi_t^{\theta^2} \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^2) \Big) \mathrm{d}\mu_t^{\theta^2}(x, y) \right| \\
&=: \frac{1}{2\lambda} (|I_1| + |I_2|) \, .
\end{aligned}$$

We start by further simplifying the integral term I_2, which can be recast as

$$\begin{aligned}
|I_2| &= \left| \int_{\mathbb{R}^{2d}} \Big(\nabla_x \psi_t^{\theta^1} \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^1) - \nabla_x \psi_t^{\theta^2} \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^1) \right. \\
&\qquad \left. + \nabla_x \psi_t^{\theta^2} \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^1) - \nabla_x \psi_t^{\theta^2} \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^2) \Big) \mathrm{d}\mu_t^{\theta^2}(x, y) \right| \\
&\leq \left| \int_{\mathbb{R}^{2d}} (\nabla_x \psi_t^{\theta^1} - \nabla_x \psi_t^{\theta^2}) \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^1) \mathrm{d}\mu_t^{\theta^2}(x, y) \right| \\
&\quad + \left| \int_{\mathbb{R}^{2d}} \nabla_x \psi_t^{\theta^2} \cdot (\nabla_\theta \mathcal{F}(t, x, \theta_t^1) - \nabla_\theta \mathcal{F}(t, x, \theta_t^2)) \mathrm{d}\mu_t^{\theta^2}(x, y) \right| \\
&=: |I_3| + |I_4| \, .
\end{aligned}$$

Hence, the estimate in (3.2) is equivalent to

$$|\Lambda(\theta^1)(t) - \Lambda(\theta^2)(t)| \leq \frac{1}{2\lambda}(|I_1| + |I_3| + |I_4|)\,. \tag{3.28}$$

Let us focus on each term separately, starting from the integral I_1. Henceforth, we only consider the integrals over $B(R_T)$, in which the curves μ^{θ^i} are supported for $i = 1, 2$. By using the same reasoning as in (3.25), we obtain

$$\begin{aligned}
|I_1| &= \left| \int_{B(R_T)} \nabla_x \psi_t^{\theta^1} \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^1)(\mathrm{d}\mu_t^{\theta^1} - \mathrm{d}\mu_t^{\theta^2})(x, y) \right| \\
&\leq \mathrm{Lip}\left(\nabla_x \psi_t^{\theta^1} \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^1)\,; B(R_T)\right) W_1(\mu_t^{\theta^1}, \mu_t^{\theta^2}) \\
&\leq C(R_T', T, C_\Gamma, C_{\mathcal{F},T}, L_{F,T,C_\Gamma}, R_T)\, \|\psi_T\|_{\mathcal{C}^2(B(R_T'))}\, W_1(\mu_t^{\theta^1}, \mu_t^{\theta^2})\,.
\end{aligned} \tag{3.29}$$

Observe now that following the Appendix, the curves $\mu_t^{\theta^1}$ and $\mu_t^{\theta^2}$ are characteristic solutions of (3.10), in the sense that

$$\mu_t^{\theta^i} = \mathbf{\Phi}_{(0,t)}^{\theta^i} \sharp \mu_0 \qquad \text{for all } t \in [0, T], \tag{3.30}$$

where for $i = 1, 2$ the flow maps $\mathbf{\Phi}_{(0,t)}^{\theta^i}(\cdot)$ correspond to the underlying ODEs

$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = \mathcal{F}(t, X_t^i, \theta_t^i), \qquad \frac{\mathrm{d}Y_t^i}{\mathrm{d}t} = 0, \qquad (X_0^i, Y_0^i) = (x_0, y_0)\,,$$

Then, considering the Assumptions (2), it follows that

$$\begin{aligned}
\left|(X_t^1, Y_t^1) - (X_t^2, Y_t^2)\right| &= \left| \left(x_0 - x_0 + \int_0^t (\mathcal{F}(t, X_s^1, \theta_s^1) - \mathcal{F}(t, X_s^2, \theta_s^2))\mathrm{d}s, y_0 - y_0 \right) \right| \\
&\leq \int_0^t \left|(\mathcal{F}(t, X_s^1, \theta_s^1) - \mathcal{F}(t, X_s^2, \theta_s^2)\right| \mathrm{d}s \\
&\leq \int_0^t \left|(\mathcal{F}(t, X_s^1, \theta_s^1) - \mathcal{F}(t, X_s^2, \theta_s^1) + \mathcal{F}(t, X_s^2, \theta_s^2) - \mathcal{F}(t, X_s^2, \theta_s^2)\right| \mathrm{d}s \\
&\leq \int_0^t \left|\mathcal{F}(t, X_s^1, \theta_s^1) - \mathcal{F}(t, X_s^2, \theta_s^1)\right| \mathrm{d}s + \int_0^t \mathcal{F}(t, X_s^2, \theta_s^1) - \mathcal{F}(t, X_s^2, \theta_s^2)\right| \mathrm{d}s \\
&\leq C_{\mathcal{F},T} \int_0^t |X_s^1 - X_s^2| \mathrm{d}s + C(R_T, T) \int_0^t |\theta_s^1 - \theta_s^2| \mathrm{d}s\,.
\end{aligned} \tag{3.31}$$

Then by Gronwall's lemma and the definition of Wasserstein distance, we obtain

$$W_1(\mu_t^{\theta^1}, \mu_t^{\theta^2}) \leq W_1\left(\mathbf{\Phi}_{(0,t)}^{\theta^1} \sharp \mu_0, \mathbf{\Phi}_{(0,t)}^{\theta^2} \sharp \mu_0\right) \leq C(R_T, T) e^{C_{\mathcal{F},T} T} \|\theta^1 - \theta^2\|_2\,. \tag{3.32}$$

By using (3.32) in (3.29), we further obtain

$$|I_1| \leq C(R_T', T, C_\Gamma, C_{\mathcal{F},T}, L_{F,T,C_\Gamma}, R_T)\, \|\psi_T\|_{\mathcal{C}^2(B(R_T))}\, \|\theta^1 - \theta^2\|_2\,. \tag{3.33}$$

Now we focus on the term I_3. It follows from *1* in Assumption 3 that

$$\begin{aligned}
|I_3| &= \left| \int_{B(R_T)} (\psi_t^{\theta^1} - \psi_t^{\theta^2}) \nabla_x \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^1) \mathrm{d}\mu_t^{\theta^2}(x, y) \right| \\
&\leq C(R_T, T, C_\Gamma) \sup_{t \in [0,T]} \left\| \psi_t^{\theta^1} - \psi_t^{\theta^2} \right\|_{\mathcal{C}(B(R_T'))}\,.
\end{aligned} \tag{3.34}$$

Recalling that $\psi^{\theta^1}, \psi^{\theta^2}$ are characteristic solutions of (3.11) while using (4.21), one has

$$\begin{aligned}\left\|\psi_t^{\theta^1} - \psi_t^{\theta^2}\right\|_{\mathcal{C}(B(R'_T))} &= \left\|\psi_T\big(\mathbf{\Phi}^{\theta^1}_{(t,T)}(\cdot,\cdot)\big) - \psi_T\big(\mathbf{\Phi}^{\theta^2}_{(t,T)}(\cdot,\cdot)\big)\right\|_{\mathcal{C}(B(R'_T))} \\ &= \|\psi_T\|_{\mathcal{C}^1(B(R_T))} \left\|\mathbf{\Phi}^{\theta^1}_{(t,T)}(\cdot,\cdot) - \mathbf{\Phi}^{\theta^2}_{(t,T)}(\cdot,\cdot)\right\|_{C(B(R_T))} . \end{aligned} \tag{3.35}$$

Now we consider the following ODEs

$$\frac{\mathrm{d}X_s^i}{\mathrm{d}s} = \mathcal{F}(s, X_s^i, \theta_s^i), \quad \frac{\mathrm{d}Y_s^i}{\mathrm{d}s} = 0\,, \tag{3.36}$$

with initial condition $(X_s^i, Y_s^i)\big|_{s=t} = (x, y)$ for $i = 1, 2$. As mentioned in the proof of Proposition 3.3 one has

$$\mathbf{\Phi}^{\theta^i}_{(t,T)}(x,y) = (X_T^i, Y_T^i) = \left(x + \int_t^T \mathcal{F}(s, X_s^i, \theta_s^i)\,\mathrm{d}s\,,\ y\right). \tag{3.37}$$

Thus following the same arguments as in (3.31) and using again Gronwall's lemma, we obtain

$$\left|\mathbf{\Phi}^{\theta^1}_{(t,T)}(x,y) - \mathbf{\Phi}^{\theta^2}_{(t,T)}(x,y)\right| = |(X_T^1, Y_T^1) - (X_T^2, Y_T^2)| \le C(R_T, T)e^{C_{\mathcal{F},T}T}\|\theta^1 - \theta^2\|_2\,. \tag{3.38}$$

Hence, the term I_3 can be estimated as

$$|I_3| \le C(T, R_T, C_\Gamma, C_{\mathcal{F},T})\,\|\psi_T\|_{\mathcal{C}^1((B(R_T))}\,\|\theta^1 - \theta^2\|_2\,. \tag{3.39}$$

Lastly, we focus on the integral quantity I_4. Using *4* from Assumption (2), we can write

$$\begin{aligned}|I_4| &\le \int_{\mathbb{R}^{2d}} \nabla_x \psi_T^{\theta^2} \cdot \left|\nabla_\theta(\mathcal{F}(t,x,\theta_t^1) - \mathcal{F}(t,x,\theta_t^2))\right| \mathrm{d}\mu_t^{\theta^2}(x,y) \\ &\le \int_{\mathbb{R}^{2d}} \nabla_x \psi_T^{\theta^2} \cdot \left|\nabla_\theta^2(\mathcal{F}(t,x,\theta)\right| \left|\theta_t^1 - \theta_t^2\right| \mathrm{d}\mu_t^{\theta^2}(x,y) \\ &\le C(R_T,T)|\theta_t^1 - \theta_t^2| \sup_{t\in[0,T]} \left\|\psi_t^{\theta^2}\right\|_{C^1(B(R'_T))} \\ &\le C(R_T, T, R'_T, C_\Gamma, C_{\mathcal{F},T}, L_{F,T,C_\Gamma})\,\|\psi_T\|_{\mathcal{C}^1(B(R_T))}\,|\theta_t^1 - \theta_t^2|\,. \end{aligned} \tag{3.40}$$

Collecting the estimates from (3.33), (3.39) and (3.40), we can conclude

$$\begin{aligned}\|\Lambda(\theta^1) - \Lambda(\theta^2)\|_2 &\le \frac{1}{2\lambda} C(R'_T, R_T, T, C_{\mathcal{F},T}, C_\Gamma, L_{F,T,C_\Gamma})\,\|\psi_T\|_{\mathcal{C}^2(B(R_T))}\,\|\theta^1 - \theta^2\|_2 \\ &= \kappa_\lambda \|\theta^1 - \theta^2\|_2\,. \end{aligned}$$

Hence by choosing the parameter $\lambda > 0$ to be large enough, we obtain that $\kappa_\lambda < 1$, which means that the mapping $\Lambda : \Gamma_{M,C} \to \Gamma_{M,C}$ is a contraction and thus that its fixed point θ^* is unique in Γ_C. Thus we have obtained a solution $(\mu^*, \theta^*, \psi^*) \in \mathcal{C}([0,T]; \mathcal{P}_c(\mathbb{R}^{2d})) \times \Gamma_{M,C} \times \mathcal{C}^1([0,T]; \mathcal{C}_c^2(\mathbb{R}^{2d}))$ to equations (3.10)-(3.12), and it is unique in Γ_C. □

Remark 3.4. *As it was shown in the proof above, the size condition imposed on λ depends on some constant $C(|R'_T|, R_T, T, C_{\mathcal{F},T}, C_\Gamma, L_{F,T,C_\Gamma})$ and $\|\psi_T\|_{\mathcal{C}^2(R_T)}$. Especially for the case $F := \tanh(\theta_t x)$, we can simplify the constant as $C(R_T, T, C_\Gamma)$, which shows that λ depends on the size of the support of μ_0, on the final time $T > 0$ and on the constant C_Γ.*

In addition to its usefulness in characterizing and computing optimal controls, the mean-field maximum principle allows us to derive a quantitative norm rate of convergence of the latter with respect to the L^p-norms and a quantitative generalization error.

Corollary 3.4. *For any $T > 0$, suppose that $\mathcal{F}$ satisfies the Assumption 2 and 3, let $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$ be an initial datum, and ψ_T be a terminal condition satisfying* (3.18)*. Suppose that for $N \geq 1$ we are given $\mu_0^N := \frac{1}{N}\sum_{i=1}^N \delta_{(X_0^i, Y_0^i)} \in \mathcal{P}_c^N(\mathbb{R}^{2d})$ an approximating empirical measure, such that $\lim_{N\to\infty} W_1(\mu_0^N, \mu_0) = 0$. Let $\lambda > 0$ be large enough, so that $(\mu^*, \theta^*, \psi^*) \in \mathcal{C}([0,T]; \mathcal{P}_c(\mathbb{R}^{2d})) \times \mathrm{Lip}([0,T]; \mathbb{R}^m) \times \mathcal{C}^1([0,T]; \mathcal{C}_c^2(\mathbb{R}^{2d}))$ and $(\mu^N, \theta^N, \psi^N) \in \mathcal{C}([0,T]; \mathcal{P}_c^N(\mathbb{R}^{2d})) \times \mathrm{Lip}([0,T]; \mathbb{R}^m) \times \mathcal{C}^1([0,T]; \mathcal{C}_c^2(\mathbb{R}^{2d}))$ are the unique solutions of* (3.10)-(3.12) *with initial conditions μ_0 and μ_0^N respectively. Then*

$$\max\left\{\|\theta^N - \theta^*\|_p, \sup_{t\in[0,T]} W_1(\mu_t^N, \mu_t^*), \|\psi^N - \psi^*\|_{\mathcal{C}([0,T])}\right\} \leq C W_1(\mu_0^N, \mu_0), \tag{3.41}$$

for a constant $C > 0$ which depends on the parameters of the model, and $p \in [1, \infty]$. In particular, we obtain the following generalization error estimate

$$\left| \int_{\mathbb{R}^{2d}} \ell(x,y)\, \mathrm{d}\mu_T^*(x,y) - \frac{1}{N}\sum_{i=1}^N \ell\big(X_T^i, Y_T^i\big) \right| \leq C W_1(\mu_0^N, \mu_0). \tag{3.42}$$

Proof. By using similar arguments as in the proof of Theorem 3.1, see in particular (3.17) and (3.31)-(3.32), we can prove the stability estimate

$$\begin{aligned}\sup_{t\in[0,T]} W_1(\mu_t^N, \mu_t^*) &\leq \sup_{t\in[0,T]} W_1(\mu_t^N, \mu_t^{\theta^N}) + \sup_{t\in[0,T]} W_1(\mu_t^{\theta^N}, \mu_t^*) \\ &\leq C\left(W_1(\mu_0^N, \mu_0) + \int_0^T |\theta_t^N - \theta_t^*|\, \mathrm{d}t\right) \leq C\left(W_1(\mu_0^N, \mu_0) + \|\theta^N - \theta^*\|_p\right),\end{aligned} \tag{3.43}$$

where $\mu_t^{\theta^N}$ is the unique solution of (3.5) driven by θ^N with initial datum μ_0. Similarly, from (3.31), (3.35) and (3.38), we have

$$\|\psi^N - \psi^*\|_{\mathcal{C}([0,T])} \leq C \int_0^T |\theta_t^N - \theta_t^*|\, \mathrm{d}t \leq C\|\theta^N - \theta^*\|_p, \tag{3.44}$$

for any $p \in [1, +\infty]$. Finally, by using the fixed point equations

$$\theta^N = \Lambda(\theta^N) \qquad \text{and} \qquad \theta^* = \Lambda(\theta^*),$$

and following the estimates in the proof of Theorem 3.1, see in particular (3.28), (3.29), (3.34) and (3.40), we obtain

$$\begin{aligned}\|\theta^N - \theta^*\|_p = \|\Lambda(\theta^N) - \Lambda(\theta^*)\|_p &\leq \frac{C}{\lambda}\left(\sup_{t\in[0,T]} W_1(\mu_t^N, \mu_t^*) + \|\psi^N - \psi^*\|_{\mathcal{C}([0,T])} + \|\theta^N - \theta^*\|_p\right) \\ &\leq \frac{C}{\lambda}\left(W_1(\mu_0^N, \mu_0) + \|\theta^N - \theta^*\|_p\right),\end{aligned}$$

where we applied (3.43) and (3.44) in the last inequality. Hence for $\lambda > 0$ large enough, it holds

$$\|\theta^N - \theta^*\|_p \leq C W_1(\mu_0^N, \mu_0). \tag{3.45}$$

Combining now (3.43), (3.44) and (3.45) finally yields (3.41). The generalization error displayed in (3.42) follows from (3.43) and (3.45), since

$$\begin{aligned}\left| \int_{\mathbb{R}^{2d}} \ell(x,y)\, \mathrm{d}(\mu_T^*(x,y) - \mu_T^N(x,y)) \right| &\leq \mathrm{Lip}(\ell|_{\mathrm{supp}\,\mu \cup \mathrm{supp}\,\mu^N}) \sup_{t\in[0,T]} W_1(\mu_t^N, \mu_t^*) \\ &\leq C\left(W_1(\mu_0^N, \mu_0) + \|\theta^N - \theta^*\|_p\right) \\ &\leq C W_1(\mu_0^N, \mu_0).\end{aligned}$$

This completes the proof of Corollary 3.4. □

3.3 Rigorous derivation of the mean-field maximum principle

The previous section, we proved the well-posedness of the mean-field PMP (3.10)-(3.12) in the class of control that are (Lipschitz) continuous in time. Under this assumption, in the following we rigorously derive the optimality conditions by using a generalized Lagrange multiplier theorem over convex set. The requirement of continuity of the control is technical and it is due to our use of [60, Theorem 1] about existence of transport equations with continuous in time sources. Were such result generalized to the case of sources, which are merely measurable in time, then we could also remove the continuity assumption.
The method we present is to a certain extent a standard calculus of variations approach and it is bypassing the more technical ones based either on abstract linearisations as in [15, 17, 20], or on the calculus of Wasserstein spaces used in [22].

3.3.1 A Lagrange Multiplier Theorem over convex sets

Let X and Y be Banach spaces, $E \subset X$ be a convex set, $J : E \to \mathbb{R}$ be a continuous functional and $G : E \to Y$ be a linear mapping, both continuously F-differentiable on E in the sense of (2.17). For $x^* \in E$, denote

$$DG(x^*) := \Big\{ L \in \mathcal{L}(X_E, Y) \;\Big|\; L \text{ satisfies (2.15)} \Big\}. \tag{3.46}$$

It is known that every $L \in \mathcal{L}(X_E, Y)$ can be uniquely extended to a operator $\overline{L} \in \mathcal{L}(\overline{X}_E, Y)$ over the Banach space $\overline{X}_E$. In what follows, we will slightly abuse the notation $DG(x^*)$ to denote the set of operators obtained after extending the convex subgradients to $\overline{X}_E$.

In the following theorem, we extend the Lagrange multiplier theorem for the Banach space [70, Section 4.14] to the setting of the calculus for convex subsets introduced in Section 2. To ease the readability of the paper, the proof of this result is reported in the Appendix.

Theorem 3.5. *Let $x^* \in E$ be a solution to the following constrained optimization problem*

$$\inf_{x \in E} J(x) \quad \text{subject to} \quad G(x) = 0\,. \tag{3.47}$$

Suppose that the inclusion $x^* + X_E \subset E$ holds, and that there exists some $G'(x^*) \in DG(x^*)$ that is a surjective operator from $\overline{X}_E$ into Y. Then for any $J'(x^*) \in DJ(x^*)$, there exists a non-zero covector $p^* \in Y'$ which satisfies

$$\langle J'(x^*), z\rangle + \langle G'(x^*)z, p^*\rangle = 0 \quad \text{for all } z \in \overline{X}_E\,. \tag{3.48}$$

3.3.2 Preparation and verification of assumptions

Recall that in Theorem 2.3, we have shown that for every $\theta \in L^2([0,T];\mathbb{R}^m)$, there exists a unique solution $\mu \in \mathcal{C}([0,T];\mathcal{P}_c(\mathbb{R}^{2d}))$ to the continuity equation 1.7. In the sequel, we assume that $\theta \in \mathcal{C}([0,T];\mathbb{R}^m)$ so that the map $t \mapsto \mathcal{F}(t,x,\theta_t)$ is continuous on $[0,T]$, and that $\mathcal{F}$ satisfies Assumption 2.

Under these working assumption we can further prove that the solution μ is such that $\partial_t \mu \in \mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')$. Indeed for any $\varphi \in \mathcal{C}_b^1(\mathbb{R}^{2d})$, one has

$$\begin{aligned} \|\partial_t \mu_t\|_{(\mathcal{C}_b^1(\mathbb{R}^{2d}))'} &= \sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} |\langle \partial_t \mu_t, \varphi\rangle| \qquad (3.49)\\ &= \sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} |\langle \mathcal{F}(t,\cdot,\theta_t)\mu_t, \nabla_x \varphi\rangle| \leq \|\mathcal{F}\|_{L^\infty(\text{supp }\mu_t)} \leq C_{\mathcal{F},T}(1+|R_T|)\,. \end{aligned}$$

Additionally, it holds for any $s,t \in [0,T]$

$$\begin{aligned} \|\partial_t \mu_t - \partial_s \mu_s\|_{(\mathcal{C}_b^1(\mathbb{R}^{2d}))'} &= \sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} |\langle \partial_t \mu_t - \partial_s \mu_s, \varphi\rangle| \qquad (3.50)\\ &= \sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} |\langle \mathcal{F}(t,\cdot,\theta_t)\mu_t - \mathcal{F}(s,\cdot,\theta_s)\mu_s, \nabla_x \varphi\rangle|\\ &\leq \sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} \Big|\big\langle (\mathcal{F}(t,\cdot,\theta_t) - \mathcal{F}(s,\cdot,\theta_s))\mu_t, \nabla_x \varphi\big\rangle\Big| \qquad (3.51)\\ &\quad + \sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} |\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla_x \varphi\rangle| \qquad (3.52)\\ &\leq C|t-s| + \sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} |\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla_x \varphi\rangle|\,, \qquad (3.53) \end{aligned}$$

Observe that for every $\varphi \in \mathcal{C}_b^1(\mathbb{R}^{2d})$, there exists a sequence $(\varphi^n)_{n\in\mathbb{N}} \subset \mathcal{C}_b^2(\mathbb{R}^{2d})$ such that $\|\varphi^n - \varphi\|_{\mathcal{C}^1} \to 0$ as $n \to +\infty$. Thus, one has

$$\begin{aligned} &\sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} |\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla_x \varphi\rangle|\\ &\leq \sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} \Big|\big\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), (\nabla_x \varphi - \nabla_x \varphi^n)\big\rangle\Big| + \sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} |\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla_x \varphi^n\rangle|\\ &\leq C\,\|\varphi^n - \varphi\|_{\mathcal{C}^1} + \text{Lip}\big(\mathcal{F}(t,\cdot,\theta_t)\cdot \nabla_x \varphi^n\big) W_1(\mu_t,\mu_s) \qquad (3.54)\\ &\leq C\,\|\varphi^n - \varphi\|_{\mathcal{C}^1} + C_n|t-s|\,, \qquad (3.55) \end{aligned}$$

where we have used the Kantorovitch duality (2.6) and (4.12). This further yields that

$$\lim_{s\to t} \sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} |\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla_x \varphi\rangle| \leq \|\varphi^n - \varphi\|_{\mathcal{C}_b^1}\,, \qquad (3.56)$$

for every $n \in \mathbb{N}$. Therefore letting $n \to +\infty$ in (3.56), we can conclude

$$\sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} |\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla_x \varphi\rangle| \underset{s\to t}{\longrightarrow} 0. \qquad (3.57)$$

This combined with (3.50) and the fact that $t \mapsto \mathcal{F}(t,x,\theta_t) \in \mathbb{R}^d$ is continuous implies that $\partial_t \mu \in \mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')$. In the sequel, we will therefore consider trajectory-control pairs $(\mu^*,\theta^*) \in \mathcal{C}^1([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))') \times \mathcal{C}([0,T];\mathbb{R}^m)$ solution of the optimal control problem (1.7)-(1.8), where we have used the notation $\mu \in \mathcal{C}^1([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')$ to represent that $\mu \in \mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')$ and $\partial_t \mu \in \mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')$.

In order to apply the multiplier rule of Theorem 3.5, we need following preparations.

∘ **The setup of spaces and sets.** Let us start by defining the spaces

$$V := \widetilde{\mathcal{C}}([0,T];\mathcal{M}_{1,c}(\mathbb{R}^{2d})) \cap \mathcal{C}^1([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))'), \qquad Q := \mathcal{C}([0,T];\mathbb{R}^m)\,, \tag{3.58}$$

where

$$\widetilde{\mathcal{C}}([0,T];\mathcal{M}_{1,c}(\mathbb{R}^{2d})) := \Big\{\mu \in \mathcal{C}([0,T];\mathcal{M}_{1,c}(\mathbb{R}^{2d})) \;\Big|\; \mathrm{supp}(\mu_t) \subset S_\mu \text{ for all } t \in [0,T] \text{ where } S_\mu \subset \mathbb{R}^d \text{ is a compact set}\Big\}\,, \tag{3.59}$$

and set

$$E := V \times Q \;=\; \widetilde{\mathcal{C}}([0,T];\mathcal{M}_{1,c}(\mathbb{R}^{2d})) \cap \mathcal{C}^1([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))') \times \mathcal{C}([0,T];\mathbb{R}^m)\,. \tag{3.60}$$

Clearly, $(\mu^*,\theta^*) \in E$ since $\mathcal{P}_c(\mathbb{R}^{2d}) \subset \mathcal{M}_{1,c}(\mathbb{R}^{2d})$. We also observe that E is a convex subset of the Banach space

$$X := U \times Q = \mathcal{C}^1([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))') \times \mathcal{C}([0,T];\mathbb{R}^m)\,. \tag{3.61}$$

Due to this embedding, we shall from now on endow $\mathcal{M}_{1,c}(\mathbb{R}^{2d})$ with the weak$-^*$ topology of $(\mathcal{C}^1_b(\mathbb{R}^{2d}))'$. In what follows, we denote by $U_V := \mathbb{R}(V-V)$, and we will use the identity

$$U_V := \widetilde{\mathcal{C}}([0,T];\mathcal{M}_{0,c}(\mathbb{R}^{2d})) \cap \mathcal{C}^1([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))')\,. \tag{3.62}$$

For $\nu \in V$, we shall define U_ν as the convex cone of directions

$$U_\nu := \mathbb{R}^+(V-\nu) \subset U_V\,, \tag{3.63}$$

in keeping with the notations introduced in Section 2. In fact, one can easily check that $U_\nu = U_V$, since for any $\mu \in U_V$, one has $\mu = \mu + \nu - \nu$ with $\mu + \nu \in V$. Next we introduce

$$X_E := U_V \times Q \;=\; \widetilde{\mathcal{C}}([0,T];\mathcal{M}_{0,c}(\mathbb{R}^{2d})) \cap \mathcal{C}^1([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))') \times \mathcal{C}([0,T];\mathbb{R}^m)\,. \tag{3.64}$$

that is seen as a convex subset of X. It follows from the definitions of E and X_E that $(\mu^*,\theta^*) + X_E \subset E$, which is compatible with the assumptions of Theorem 3.5.

∘ **The setup of maps.** For any $(\mu,\theta) \in E$, let us recall that

$$J(\mu,\theta) := \int_{\mathbb{R}^{2d}} \ell(x,y)\mathrm{d}\mu_T(x,y) + \lambda \int_0^T |\theta_t|^2\,\mathrm{d}t\,, \tag{3.65}$$

which is a map from E into $\mathbb{R}^+$. We also denote by

$$G(\mu,\theta) := -\partial_t\mu - \nabla_x \cdot (\mathcal{F}(t,x,\theta)\mu)\,. \tag{3.66}$$

Seeing $G(\mu,\theta)$ as time-dependent quantity, it is easy to check that $G(\mu,\theta) \in \mathcal{C}([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))')$ for $(\mu,\theta) \in E$, and that $\langle G(\mu,\theta)_t, 1\rangle = 0$ for all $t \in [0,T]$. Indeed for any $\varphi \in \mathcal{C}^1_b(\mathbb{R}^{2d})$, it holds

$$\begin{aligned}
\|G(\mu,\theta)_t - G(\mu,\theta)_s\,\|_{(\mathcal{C}^1_b)'} &= \sup_{\|\varphi\|_{\mathcal{C}^1_b}\le 1} |\langle G(\mu,\theta)_t - G(\mu,\theta)_s, \varphi\rangle| \\
&= \|\partial_t\mu_t - \partial_s\mu_s\,\|_{(\mathcal{C}^1_b)'} + \sup_{\|\varphi\|_{\mathcal{C}^1_b}\le 1} \big|\big\langle (\mathcal{F}(t,\cdot,\theta_t) - \mathcal{F}(s,\cdot,\theta_s))\mu_t, \nabla\varphi\big\rangle\big| \\
&\quad + \sup_{\|\varphi\|_{\mathcal{C}^1_b}\le 1} \big|\big\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla\varphi\big\rangle\big|..
\end{aligned}$$

Following similar density arguments as in (3.54)-(3.57) one can deduce

$$\sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} |\langle \mathcal{F}(s,x,\theta_s)(\mu_t-\mu_s),\nabla\varphi\rangle| \leq C\|\mu_t-\mu_s\|_{(\mathcal{C}^1)'}\,. \tag{3.67}$$

This with together with the fact that $\mu\in\mathcal{C}^1([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')$ and that $t\in[0,T]\mapsto\mathcal{F}(t,\cdot,\theta_t)$ is continuous in time yields $G(\mu,\theta)\in\mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')$. Observe now that for any $\mu\in\widetilde{\mathcal{C}}([0,T];\mathcal{M}_{1,c}(\mathbb{R}^{2d}))$, there exists some compact set $S_\mu\subset\mathbb{R}^d$ such that

$$\mathrm{supp}(\mu_t)\subset S_\mu \quad \text{ for all } t\in[0,T]\,. \tag{3.68}$$

This implies that $G(\mu,\theta)$ is uniformly compactly supported in the sense of distribution, namely $G: E\to Y_0$ with

$$\begin{aligned} Y_0 :&= \widetilde{\mathcal{C}}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))'_{0,c})\\ &= \left\{ g\in\mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))') \;\middle|\; \langle g_t,1\rangle=0 \text{ and } \mathrm{supp}(g_t)\subset S_g\Subset\mathbb{R}^{2d},\ \forall t\in[0,T]\right\}. \end{aligned}$$

This allows us to define the Banach space

$$Y:=\overline{Y}_0=\overline{\widetilde{\mathcal{C}}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))'_{0,c})}\,, \tag{3.69}$$

which is a closed subspace of the Banach space $\mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')$.

Now let us verify that $G\in\mathcal{C}^1(E;Y)$ and $J\in\mathcal{C}^1(E;\mathbb{R})$. For any $t\in[0,T]$, it holds that

$$\begin{aligned} \big\|G(\mu^1,\theta^1)_t-G(\mu^2,\theta^2)_t\big\|_{(\mathcal{C}_b^1(\mathbb{R}^{2d}))'} &= \sup_{\|\varphi\|_{\mathcal{C}_b^1(\mathbb{R}^{2d})}\leq 1} \big|\langle G(\mu^1,\theta^1)_t-G(\mu^2,\theta^2)_t,\varphi\rangle\big|\\ &= \big\|\partial_t\mu_t^1-\partial_t\mu_t^2\big\|_{(\mathcal{C}_b^1)'}+\sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1}\big|\langle\mathcal{F}(t,x,\theta_t^1)(\mu_t^1-\mu_t^2),\nabla\varphi\rangle\big|\\ &\quad+\sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1}\big|\langle(\mathcal{F}(t,x,\theta_t^1)-\mathcal{F}(t,x,\theta_t^2))\mu_t^2,\nabla\varphi\rangle\big|\\ &\leq \big\|\partial_t\mu_t^1-\partial_t\mu_t^2\big\|_{(\mathcal{C}_b^1)'}+C\big\|\mu_t^1-\mu_t^2\big\|_{(\mathcal{C}_b^1)'}+C(R_T,T)|\theta_t^1-\theta_t^2|\,, \end{aligned}$$

where we have again used density arguments similar to that of (3.54)-(3.57). Thus, we have proven that

$$\begin{aligned} \big\|G(\mu^1,\theta^1)-G(\mu^2,\theta^2)\big\|_{\mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')} &\leq C\big\|\mu^1-\mu^2\big\|_{\mathcal{C}^1([0,T];\mathcal{C}_b^1(\mathbb{R}^{2d}))}\\ &\quad+C(R_T,T)\|\theta_1-\theta_2\|_{\mathcal{C}([0,T];\mathbb{R}^m)}\,, \end{aligned} \tag{3.70}$$

which implies that $G\in\mathcal{C}(E;Y)$. Similarly we have

$$\begin{aligned} &|J(\mu^1,\theta^1)-J(\mu^2,\theta^2)|\\ &\leq\left|\int_{\mathbb{R}^{2d}}\ell(x,y)\mathrm{d}(\mu_T^1-\mu_T^2)(x,y)+\int_0^T(|\theta_t^1|^2-|\theta_t^2|^2)\,\mathrm{d}t\right|\\ &\leq C\big\|\mu_T^1-\mu_T^2\big\|_{(\mathcal{C}_b^1)'}+C\Big(T,\|\theta_1\|_{\mathcal{C}([0,T];\mathbb{R}^m)},\|\theta_2\|_{\mathcal{C}([0,T];\mathbb{R}^m)}\Big)\|\theta_1-\theta_2\|_{\mathcal{C}([0,T];\mathbb{R}^m)}\,, \end{aligned}$$

where we used the fact that μ_T^1 and μ_T^2 are compactly supported. This in turn implies that $J\in\mathcal{C}(E;\mathbb{R})$.

Next, we use Lemma 2.1 to prove that both mappings are in fact $\mathcal{C}^1$-smooth. It follows from the definition (2.18) of G-derivative that for all $\mu \in V$, $\nu \in U_\mu = U_V$ and $\varphi \in \mathcal{C}_b^1(\mathbb{R}^{2d})$, one has

$$\langle \mathrm{d}_\mu G(\mu,\theta)(\nu), \varphi\rangle = \left\langle \lim_{\varepsilon\to 0^+} \frac{G(\mu+\varepsilon\nu,\theta)-G(\mu,\theta)}{\varepsilon}, \varphi \right\rangle \tag{3.71}$$

$$= \lim_{\varepsilon\to 0^+} \frac{\langle G(\mu+\varepsilon\nu,\theta),\varphi\rangle - \langle G(\mu,\theta),\varphi\rangle}{\varepsilon}$$

$$= \langle -\partial_t \nu - \nabla_x \cdot (\mathcal{F}(t,x,\theta)\nu), \varphi\rangle < +\infty\,. \tag{3.72}$$

Thus we have found a continuous operator $\mu \in V \mapsto L_\theta(\mu) \in \mathcal{L}(U_V, Y)$ such that $L_\theta(\mu)(\nu) := -\partial_t\nu - \nabla_x\cdot(\mathcal{F}(t,x,\theta)\nu) = \mathrm{d}_\mu G(\mu,\theta)(\nu)$ for all $\mu \in V$ and $\nu \in U_\mu$. Applying Lemma 2.1 allows us to conclude that $L_\theta(\mu) \in D_\mu G(\mu,\theta)$ and $G(\cdot,\theta) \in \mathcal{C}^1(V;Y)$. Additionally, remark that the standard Fréchet differential $G'_\theta(\mu,\theta) : Q \to Y$ with respect to the control curve satisfies

$$\langle G'_\theta(\mu,\theta)(\alpha),\varphi\rangle = \lim_{\varepsilon\to 0^+} \frac{\langle G(\mu,\theta+\varepsilon\alpha),\varphi\rangle - \langle G(\mu,\theta),\varphi\rangle}{\varepsilon} = \langle -\nabla_x\cdot(\nabla_\theta\mathcal{F}(t,x,\theta)\alpha\mu),\varphi\rangle < +\infty\,. \tag{3.73}$$

for all $\alpha \in Q$. The continuity of $\theta \in \mathbb{R}^m \mapsto \nabla_\theta \mathcal{F}(t,x,\theta) \in \mathbb{R}^d$ implies that $G(\mu,\cdot) \in \mathcal{C}^1(Q;Y)$ for every $\mu \in V$, and thus $G \in \mathcal{C}^1(E;Y)$. Similarly, we have

$$J'_\mu(\mu,\theta)(\nu) = \int_{\mathbb{R}^{2d}} \ell(x,y)\mathrm{d}\nu_T \qquad \text{and} \qquad J'_\theta(\mu,\theta)(\alpha) = \int_0^T 2\lambda\theta_t\cdot\alpha_t\,\mathrm{d}t\,, \tag{3.74}$$

for all $\nu \in U_\mu = U_V$ and $\alpha \in Q$. It is then easy to check that $J \in \mathcal{C}^1(E;\mathbb{R})$.

3.3.3 The mean-field PMP for continuous controls: a Lagrangian approach

We are now ready to present the derivation of the first order optimality condition (3.10)-(3.12) in the class of continuous controls, by means of the Lagrange multiplier rule.

Theorem 3.6. *Let $(\mu^*,\theta^*) \in E \subset X = U \times Q$ be a solution to the optimal control problem (1.7)-(1.8). Then there exists $p^* \in Y'$ such that*

$$\langle G'_\mu(\mu^*,\theta^*)(\nu), p^*\rangle + J'_\mu(\mu^*,\theta^*)(\nu) = 0, \quad \textit{for all } \nu \in \overline{U}_V, \tag{3.75}$$

$$\langle G'_\theta(\mu^*,\theta^*)(\alpha), p^*\rangle + J'_\theta(\mu^*,\theta^*)(\alpha) = 0, \quad \textit{for all } \alpha \in Q\,. \tag{3.76}$$

Remark 3.5. *The solution $\psi^* = p^* \in \mathcal{C}^1([0,T];\mathcal{C}_c^2(\mathbb{R}^{2d}))$ constructed in Proposition 3.3 is in Y'. This is because for any $\eta \in Y \subset \mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')$, one has $\langle p^*,\eta\rangle < +\infty$.*

Proof. In order to prove our set of optimality conditions, we will use Theorem 3.5 which application has already been prepared above. Indeed we have shown that both the cost and constraint functionals are continuously F-differentiable, and it follows directly from the definitions (3.60) and (3.64) that $(\mu^*,\theta^*) + X_E \subset E$. Thus, there remains to prove that the linear operator $G'(\mu^*,\theta^*) : \overline{X}_E = \overline{U_V} \times Q \to Y$ is surjective. We split the proof of the surjectivity into two steps below.

• **Surjectivity of the partial derivative $G'_\mu(\mu^*,\theta^*):\overline{U_V}\to Y$.** We first want to show that for any given

$$\eta\in Y:=\overline{\widetilde{\mathcal{C}}([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))'_{0,c})},$$

there exists a $\nu\in\overline{U}_V$ such that

$$G'_\mu(\mu^*,\theta^*)(\nu)=\eta\,, \tag{3.77}$$

which is understood in the sense of

$$\langle G'_\mu(\mu^*_t,\theta^*_t)(\nu_t),\varphi\rangle=\langle\eta_t,\varphi\rangle\qquad\text{for all }\varphi\in\mathcal{C}^1_b(\mathbb{R}^{2d})\,. \tag{3.78}$$

To this end, it suffices to show that for a given $(\mu^*,\theta^*,\eta)\in V\times Q\times Y$, there exists some $\nu\in\overline{U_V}$ satisfying the following transport equation

$$\partial_t\nu_t+\nabla_x\cdot(\mathcal{F}(t,x,\theta^*_t)\nu_t)=-\eta_t\,, \tag{3.79}$$

with the source term $(-\eta)$ and the initial data $\nu_0\in U_{\mu_0}$. Notice that $(\mathcal{C}_b(\mathbb{R}^{2d}))'_{0,c}$ is dense in $(\mathcal{C}^1_b(\mathbb{R}^{2d}))'_{0,c}$, namely for any $\eta\in Y=\overline{\widetilde{\mathcal{C}}([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))'_{0,c})}$, there exists a sequence $(\eta^n)_{n\in\mathbb{N}}\subset\widetilde{\mathcal{C}}([0,T];(\mathcal{C}_b(\mathbb{R}^{2d}))'_{0,c})$ such that for all $\varphi\in\mathcal{C}^1_b(\mathbb{R}^{2d})$, it holds

$$\sup_{t\in[0,T]}|\langle\eta^n_t-\eta_t,\varphi\rangle|\underset{n\to+\infty}{\longrightarrow}0. \tag{3.80}$$

In particular, observe that $\sup\limits_{t\in[0,T],n\in\mathbb{N}}\|\eta^n_t\|_{(\mathcal{C}^1_b)'}<+\infty$ is uniformly bounded.

Since $\eta^n_t\in(\mathcal{C}_b(\mathbb{R}^{2d}))'_{0,c}\subset(\mathcal{C}_0(\mathbb{R}^{2d}))'_{0,c}=\mathcal{M}_{0,c}(\mathbb{R}^{2d})$, it then follows from [60, Theorem 1] that there exists a unique measure solution $\mu^{1,n}\in V$ to the following transport equation with $(-\eta^n_t)$ as a source term

$$\partial_t\mu^{1,n}_t+\nabla_x\cdot(\mathcal{F}(t,x,\theta^*_t)\mu^{1,n}_t)=-\eta^n_t,\qquad\mu^{1,n}_t|_{t=0}=\mu^1_0\in\mathcal{P}_c(\mathbb{R}^{2d})\,, \tag{3.81}$$

in a distributional sense analogous to 0(2.11), namely

$$\begin{aligned}&\int_{\mathbb{R}^{2d}}\varphi(x,y)\mathrm{d}\mu^{1,n}_{t_2}(x,y)-\int_{\mathbb{R}^{2d}}\varphi(x,y)\mathrm{d}\mu^{1,n}_{t_1}(x,y)\\&=\int_{t_1}^{t_2}\int_{\mathbb{R}^{2d}}\nabla_x\varphi(x,y)\cdot\mathcal{F}(s,x,\theta^*_s)\,\mathrm{d}\mu^{1,n}_s(x,y)\,\mathrm{d}s-\int_{t_1}^{t_2}\int_{\mathbb{R}^{2d}}\varphi(x,y)\,\mathrm{d}\eta^n_s(x,y)\,\mathrm{d}s\end{aligned}$$

for all $\varphi\in\mathcal{C}^1_b(\mathbb{R}^{2d})$ and every $t_1,t_2\in[0,T]$. Indeed, we can build a solution to above as a limit of a sequence of approximated solutions satisfying the following Euler-explicit-type splitting scheme. Fix $k\in\mathbb{N}$, and define $\Delta t=\frac{T}{2^k}$ and set $\mu^{1,n,(k)}_0=\mu_0$. Given $\mu^{1,n,(k)}_{i\Delta_t}$ for $i\in\{0,1,\cdots,2^k-1\}$, denote by $\mathcal{F}_{i\Delta t}=\mathcal{F}(i\Delta t,x,\theta^*_{i\Delta t})$ and we set

$$\mu^{1,n,(k)}_t=\Gamma^{\mathcal{F}_{i\Delta t}}_{t-i\Delta t}\sharp\mu^{1,n,(k)}_{i\Delta t}-(t-i\Delta t)\eta^n_{i\Delta t},\quad t\in[i\Delta t,(i+1)\Delta t]\,, \tag{3.82}$$

where $\Gamma^{\mathcal{F}_{i\Delta t}}_{t-i\Delta t}\sharp\mu^{1,n,(k)}_{i\Delta t}$ is the unique solution of the linear transport equation

$$\begin{cases}\partial_t f_t+\nabla\cdot(\mathcal{F}_{i\Delta t}f_t)=0,\qquad t\in(i\Delta t,(i+1)\Delta t],\\ f_{i\Delta t}=\mu^{1,n,(k)}_{i\Delta t}\,,\end{cases} \tag{3.83}$$

which is is explicitly written as a pushforward through a characteristic flow. From (3.82), we know the sequence $(\mu_t^{1,n,(k)})_{k\in\mathbb{N}}$ has uniformly bounded support, since

$$\mathrm{supp}(\mu_t^{1,n,(k)}) \subset B(R_T) \cup S_{\eta^n} \tag{3.84}$$

where $\mathrm{supp}(\eta_t^n) \subset S_{\eta^n} \Subset \mathbb{R}^{2d}$ for all $t \in [0,T]$ and we denoted by $B(R_T)$ the support of solutions to the linear transport equation obtained in (2.12). Intuitively, the support of $\mu_t^{1,n,(k)}$ is the union of the support of the solution to the linear transport equation (3.83) and the support of the source term. Similarly, it holds for $t \in [i\Delta t, (i+1)\Delta t]$

$$\|\mu_t^{1,n,(k)}\|_{(\mathcal{C}_b^1)'} \leq \|\Gamma_{t-i\Delta t}^{\mathcal{F}_{i\Delta t}} \sharp \mu_{i\Delta t}^{1,n,(k)}\|_{(\mathcal{C}_b^1)'} + \Delta t \|\eta_{i\Delta t}^n\|_{(\mathcal{C}_b^1)'} \leq \|\mu_{i\Delta t}^{1,n,(k)}\|_{(\mathcal{C}_b^1)'} + \Delta t \|\eta_{i\Delta t}^n\|_{(\mathcal{C}_b^1)'}\,. \tag{3.85}$$

This provides us with the following upper-bound

$$\sup_{t\in[0,T]} \|\mu_t^{1,n,(k)}\|_{(\mathcal{C}_b^1)'} \leq \|\mu_0^1\|_{(\mathcal{C}_b^1)'} + T \sup_{t\in[0,T]} \|\eta_t^n\|_{(\mathcal{C}_b^1)'} < +\infty\,, \tag{3.86}$$

which is uniform with respect to $n, k \in \mathbb{N}$. By letting $k \to +\infty$, we recover the existence of a solution $\mu^{1,n}$ to (3.81) such that

$$\sup_{t\in[0,T]} \mathbb{W}_1^{1,1}(\mu^{1,n}, \mu_t^{1,n,(k)}) \underset{k\to+\infty}{\longrightarrow} 0. \tag{3.87}$$

Recall that the generalized Wasserstein metric introduced in [60] is equivalent to the bounded-Lipschitz norm $\|\cdot\|_{BL}$, so that the limit curves $(\mu^{1,n})_{n\in\mathbb{N}}$ satisfy

$$\mathrm{supp}(\mu_t^{1,n}) \subset B(R_T) \cup S_{\eta^n} \quad \text{and} \quad \|\mu_t^{1,n}\|_{(\mathcal{C}_b^1)'} < +\infty \tag{3.88}$$

for all $t \in [0,T]$. This in turn implies that the sequence $(\mu_t^{1,n})_{n\in\mathbb{N}}$ is uniformly equi-bounded in $\mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')$. According to [60, Theorem 1], it follows that each curve $t \in [0,T] \mapsto \mu^{1,n}$ is Lipschitz continuous with respect to the $\|\cdot\|_{BL}$-norm, and thus it is uniformly equi-continuous with respect to the $(\mathcal{C}_b^1)'$-norm. By a direct application of the Arzelà-Ascoli theorem, there exists a subsequence of $(\mu^{1,n})_{n\in\mathbb{N}}$ that converges uniformly in $\mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')$ to some curve μ^1, which then satisfies

$$\int_{\mathbb{R}^{2d}} \varphi(x,y)\mathrm{d}\mu_{t_2}^1(x,y) - \int_{\mathbb{R}^{2d}} \varphi(x,y)\mathrm{d}\mu_{t_1}^1(x,y) \tag{3.89}$$

$$= \int_{t_1}^{t_2} \int_{\mathbb{R}^{2d}} \nabla_x \varphi(x,y) \cdot \mathcal{F}(s,x,\theta_s^*)\mathrm{d}\mu_s^1(x,y)\,\mathrm{d}s - \int_{t_1}^{t_2} \int_{\mathbb{R}^{2d}} \varphi(x,y)\mathrm{d}\eta_s(x,y)\,\mathrm{d}s\,. \tag{3.90}$$

However, recall now that the optimal curve $\mu^* \in V$ satisfies

$$\partial_t \mu_t^* + \nabla_x \cdot (\mathcal{F}(t,x,\theta_t^*)\mu_t^*) = 0, \qquad \mu_t^*|_{t=0} = \mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})\,, \tag{3.91}$$

Then, defining the curves $(\mu^{1,n} - \mu^*)_{n\in\mathbb{N}} \subset U_V$ and letting $n \to +\infty$, we can find a solution

$$\nu := \mu^1 - \mu^* = \lim_{n\to\infty} (\mu^{1,n} - \mu^*) \in \overline{U}_V,$$

to the transport equation with source term (3.79), with the initial datum $\nu_0 = \mu_0^1 - \mu_0 \in U_{\mu_0}$. This completes the proof of the surjectivity of $G'_\mu(\mu^*, \theta^*)$.

• **Surjectivity of the full derivative** $G'(\mu^*,\theta^*):\overline{X}_E=\overline{U}_V\times Q\to Y$**.** Assume that $\nu\in\overline{U}_V$ is a curve obtained as above. Then for any $\eta\in Y$, there exists $(\nu,0)\in\overline{U}_V\times Q$ such that

$$G'(\mu^*,\theta^*)(\nu,0)=G'_\mu(\mu^*,\theta^*)(\nu)+G'_\theta(\mu^*,\theta^*)(0)=\eta\,. \tag{3.92}$$

Thus, we have proven that $G'(\mu^*,\theta^*)$ is surjective. □

3.3.4 The mean-field PMP for measurable controls: an Hamiltonian approach

The goal of this subsection is to show that solutions $(\mu^*,\theta^*)\in\mathcal{C}([0,T];\mathcal{P}_c(\mathbb{R}^d))\times L^2([0,T];\mathbb{R}^m)$ the optimality condition (3.10)-(3.12) by using the Pontryagin Maximum Principle in Wasserstein spaces studied in [15, 17, 20].

For any fixed $\theta\in L^2([0,T],\mathbb{R}^m)$, we recall that $(\Phi^\theta_{(\tau,t)}(\cdot))_{\tau,t\in[0,T]}$ is the *characteristic flows* generated by the velocity field $(t,x)\in[0,T]\times\mathbb{R}^d\mapsto\mathcal{F}(t,x,\theta_t)\in\mathbb{R}^d$, defined by

$$\begin{cases}\partial_t\Phi^\theta_{(\tau,t)}(x)=\mathcal{F}\big(t,\Phi^\theta_{(\tau,t)}(x),\theta_t\big),\\ \Phi^\theta_{(\tau,\tau)}(x)=x,\end{cases} \tag{3.93}$$

for every $x\in\mathbb{R}^d$. It is a well-known result in the theory of non-linear dynamical systems (see e.g. [21, Theorem 2.3.2] that under Assumption 2, the flow maps $\Phi^\theta_{(\tau,t)}:\mathbb{R}^d\to\mathbb{R}^d$ are continuously differentiable for every $\tau,t\in[0,T]$, and the application $t\in[0,T]\mapsto\nabla_x\Phi^\theta_{(\tau,t)}(x)\in\mathbb{R}^{d\times d}$ is the unique solution of the forward linearised system

$$\begin{cases}\partial_t w(t,x)=\nabla_x\mathcal{F}\big(t,\Phi^\theta_{(\tau,t)}(x),\theta_t\big)w(t,x)\\ w(\tau,x)=\mathrm{Id}.\end{cases} \tag{3.94}$$

In the following lemma, we provide an explicit characterization of the adjoint of the differential of a flow. While it is a folklore result in the theory of non-linear ODEs, its proof is provided in very few references, and we include it in the Appendix for the sake of completeness.

Lemma 3.1. *For every $x\in\mathbb{R}^d$, the application $t\in[0,T]\mapsto\nabla_x\Phi^\theta_{(t,T)}(\Phi^\theta_{(T,t)}(x))^\top\in\mathbb{R}^{d\times d}$ is the unique solution of the backward adjoint Cauchy problem*

$$\begin{cases}\partial_t w(t,x)=-\nabla_x\mathcal{F}\big(t,\Phi^\theta_{(T,t)}(x),\theta_t\big)^\top w(t,x),\\ w(T,x)=\mathrm{Id}\,.\end{cases}$$

In the sequel, we suppose that the optimal control problem (1.7)-(1.8) admits an optimal trajectory-control pair $(\mu^*,\theta^*)\in\mathrm{Lip}([0,T];\mathcal{P}(K))\times L^2([0,T];\mathbb{R}^m)$ where $K:=B(R_T)\subset\mathbb{R}^{2d}$ is a closed ball defined for some uniform $R_T>0$. The *Hamiltonian* function $\mathbb{H}:[0,T]\times\mathcal{P}_c(\mathbb{R}^{4d})\times L^2([0,T];\mathbb{R}^m)\to\mathbb{R}$ associated to the optimal control problem is defined by

$$\mathbb{H}(t,\nu,\theta):=\int_{\mathbb{R}^{4d}}\langle r,\mathcal{F}(t,x,\theta)\rangle\mathrm{d}\nu(x,y,r,s)-\lambda|\theta|^2, \tag{3.95}$$

for almost every $t\in[0,T]$ and all $(\nu,\theta)\in\mathcal{P}_c(\mathbb{R}^{4d})\times\mathbb{R}^m$, and we denote by

$$\mathbb{J}_{4d}:=\begin{pmatrix}0&\mathrm{Id}\\-\mathrm{Id}&0\end{pmatrix},$$

the standard symplectic matrix of $\mathbb{R}^{4d}$. In this context, the PMP of [17] was adapted to unbounded control sets in [18], and can be written in context as follows.

Theorem 3.7 (Pontryagin Maximum Principle)**.** *There exists a radius $R'_T > 0$ and a uniquely determined state-costate curve $\nu^* \in \mathrm{Lip}([0,T], \mathcal{P}(K' \times K'))$ where $K' := B(R'_T) \subset \mathbb{R}^{2d}$ such that the following holds.*

(i) The curve ν^* solves the forward-backward Hamiltonian continuity equation

$$\begin{cases} \partial_t \nu_t^* + \mathrm{div}_{(x,y,r,s)}\big(\mathbb{J}_{4d} \nabla_\nu \mathbb{H}(t, \nu_t^*, \theta_t^*) \nu_t^*\big) = 0, \\ \pi^1_\# \nu_t^* = \mu_t^* \qquad\qquad \text{for all times } t \in [0,T], \\ \nu_T^* = (\mathrm{Id}, -\nabla_x \ell)_\# \mu_T^*, \end{cases} \tag{3.96}$$

where the Wasserstein gradient of the Hamiltonian is given explicitly by

$$\nabla_\nu \mathbb{H}(t, \nu_t^*, \theta_t^*)(x,y,r,s) = \begin{pmatrix} \nabla_x \mathcal{F}(t,x,\theta_t^*)^\top r \\ 0 \\ \mathcal{F}(t,x,\theta_t^*) \\ 0 \end{pmatrix},$$

for almost every $t \in [0,T]$ and all $(x,y,r,s) \in K' \times K'$.

(ii) The maximization condition

$$\mathbb{H}(t, \nu_t^*, \theta_t^*) = \max_{\theta \in \mathbb{R}^m} \mathbb{H}(t, \nu_t^*, \theta), \tag{3.97}$$

holds for almost every $t \in [0,T]$.

Below, we provide a representation formula for the state-costate curve ν^*, based on the disintegration theorem (see e.g. [6, Theorem 5.3.1]). The sufficient implication of this statement was used as early as [20] to build solutions to (3.96), while the necessary part has been established more recently in [19]. Following the notations of the Appendix, we denote by $(\mathbf{\Phi}^*_{(\tau,t)})_{\tau,t\in[0,T]}$ the characteristic flows such that $\mu_t^* = \mathbf{\Phi}^*_{(0,t)} \sharp \mu_0$ for all times $t \in [0,T]$. Observe that by construction, it holds

$$\mathbf{\Phi}^*_{(\tau,t)}(x,y) = (\Phi^*_{(\tau,t)}(x), y),$$

for all times $\tau, t \in [0,T]$ and every $(x,y) \in K'$, where $(\Phi^*_{(\tau,t)}(\cdot))_{\tau,t\in[0,T]}$ is the characteristic flow defined via (3.93) with $\theta_t := \theta_t^*$ being the optimal control.

Proposition 3.8 (Representation formula for state-costate curves)**.** *A state-costate curve $\nu^* \in \mathrm{Lip}([0,T], \mathcal{P}(K' \times K'))$ solves the forward-backward system (3.96) if and only if it can be represented as $\nu_t^* = (\mathbf{\Phi}^*_{(T,t)} \circ \pi^1, \pi^2) \sharp \nu_t^T$, where the curve $t \in [0,T] \mapsto \nu_t^T \in \mathcal{P}(K' \times K')$ is built via the disintegration formula as*

$$\nu_t^T := \int_{\mathbb{R}^{2d}} \sigma^*_{t,x,y}(t) \mathrm{d}\mu_T^*(x,y),$$

for all times $t \in [0,T]$. Here for μ_T^*-almost every $(x,y) \in \mathbb{R}^{2d}$, the curve $t \in [0,T] \mapsto \sigma^*_{t,x,y} \in \mathcal{P}(K')$ is chosen as the unique solution of the backward adjoint dynamics

$$\begin{cases} \partial_t \sigma^*_{x,y}(t) + \mathrm{div}_{(r,s)}(\mathcal{W}_{x,y}(t,r) \sigma^*_{x,y}(t)) = 0, \\ \sigma^*_{x,y}(T) = \delta_{(-\nabla_x \ell(x,y))}, \end{cases}$$

where

$$\mathcal{W}_{x,y}(t,r,s) := \begin{pmatrix} -\nabla_x \mathcal{F}(t,\theta_t^*,\Phi^*_{(T,t)}(x))^\top r \\ 0 \end{pmatrix},$$

for almost every $t \in [0,T]$ *and all* $(r,s) \in K'$.

It is easy to see that since the second marginal of μ^* is fixed, the matching part of the costate measure is also independent of time. In the following Lemma, we provide a first-order characterization of the maximization condition (3.97).

Lemma 3.2 (Fixed-point expression for the optimal control)**.** *Let* (μ^*,θ^*) *be an optimal pair for the problem* (1.7)-(1.8)*, and* ν^* *be the corresponding state-costate curve given by Theorem 3.7. Then for* $\lambda > 0$ *large enough, it holds*

$$\theta_t^* = \frac{1}{2\lambda} \int_{\mathbb{R}^{4d}} \nabla_\theta \mathcal{F}(t,\theta_t^*,x)^\top r \,\mathrm{d}\nu_t^*(x,y,r,s), \tag{3.98}$$

for almost every $t \in [0,T]$.

Proof. As a consequence of *(4)* in Assumptions 2, the map $\theta \in \mathbb{R}^m \mapsto \mathbb{H}(t,\nu_t^*,\theta)$ is twice differentiable for almost every $t \in [0,T]$. Moreover since $\mathrm{supp}(\nu_t^*) \subset B(R'_T)$, there exists a constant $C(R'_T) > 0$ such that

$$\sup_{\theta\in\mathbb{R}^m} \left| \nabla_\theta^2 \int_{\mathbb{R}^{4d}} \big\langle r, \mathcal{F}(t,x,\theta)\big\rangle \mathrm{d}\nu_t^*(x,y,r,s) \right| \le C(R'_T).$$

Hence for $\lambda > C(R'_T)$, the Hamiltonian is a concave function of θ, and the optimal control θ^* satisfies the pointwise maximization condition (3.97) if and only if

$$\nabla_\theta \mathbb{H}(t,\nu_t^*,\theta_t^*) = 0 \qquad \text{for a.e. } t \in [0,T], \tag{3.99}$$

which is equivalent to the fixed-point equation (3.98). □

For all times $t \in [0,T]$, we shall denote by $(x,y) \in K' \mapsto \overline{\sigma}^*(t,x,y) \in \mathbb{R}^d$ the d first components of the *barycentric projection* (see e.g. [6, Definition 5.4.2]) of the measures ν_t^T onto their first marginal $\pi^1_\# \nu_t^T = \mu_T^*$, namely

$$\overline{\sigma}^*(t,x,y) := \int_{\mathbb{R}^{2d}} r \,\mathrm{d}\sigma^*_{x,y}(t)(r,s).$$

Using this notation, one can easily check by linearity of the integral that the fixed-point equation (3.98) can be rewritten as

$$\theta_t^* = \frac{1}{2\lambda} \int_{\mathbb{R}^{2d}} \nabla_\theta \mathcal{F}(t,\theta_t^*,x)^\top \overline{\sigma}^*\big(t,\Phi^*_{(t,T)}(x),y\big)\,\mathrm{d}\mu_t^*(x,y),$$

for μ_T^*-almost every $(x,y) \in \mathbb{R}^{2d}$. Our goal now is to show that $\nabla_x \psi^*(t,\Phi^*_{(T,t)}(x),y) = -\overline{\sigma}^*(t,x,y)$ for all times $t \in [0,T]$ and μ_T^*-almost every $(x,y) \in \mathbb{R}^{2d}$, so that the adjoint variable $\psi^*(\cdot,\cdot)$ stemming from the Lagrangian method described throughout Section 3 satisfies

$$\theta_t^* = -\frac{1}{2\lambda} \int_{\mathbb{R}^{2d}} \nabla_\theta \mathcal{F}(t,\theta_t^*,x)^\top \nabla_x \psi^*(t,x,y) \mathrm{d}\mu_t^*(x,y)$$

which is exactly (3.4). This is the object of the following Proposition.

Proposition 3.9 (Rigorous link between the Hamiltonian and Lagrangian adjoint states)**.** *Let $\psi^* \in Y'$ be the unique characteristic solution of the formal adjoint equation* (3.11) *associated to an optimal pair $(\mu^*, \theta^*) \in \mathcal{C}([0,T];\mathcal{P}_c(\mathbb{R}^d)) \times L^2([0,T];\mathbb{R}^m)$. Then, it holds that*

$$\int_{\mathbb{R}^{2d}} \nabla_\theta \mathcal{F}(t,\theta_t^*,x)^\top \nabla_x \psi^*(t,x,y) \mathrm{d}\mu_t^*(x,y) = -\int_{\mathbb{R}^{2d}} \nabla_\theta \mathcal{F}(t,\theta_t^*,x)^\top \overline{\sigma}^*\big(t,\Phi^*_{(t,T)}(x),y\big)\,\mathrm{d}\mu_t^*(x,y),$$

for $\mathcal{L}^1$-almost every $t \in [0,T]$. In particular, the triple $(\mu^*,\theta^*,\psi^*) \in \mathcal{C}([0,T];\mathcal{P}_c(\mathbb{R}^{2d})) \times \mathrm{Lip}([0,T];\mathbb{R}^m) \times Y'$ satisfies the mean-field PMP (3.10)-(3.12).

In the following lemma, we prove that for μ_T^*-almost every $(x,y) \in \mathbb{R}^{2d}$, the map $t \in [0,T] \mapsto \overline{\sigma}^*(t,x,y) \in \mathbb{R}^d$ solves the backward linearised adjoint dynamics associated to the controlled velocity field $\mathcal{F} : [0,T] \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$.

Lemma 3.3. *For μ_T^*-almost every $(x,y) \in \mathbb{R}^{2d}$, the map $t \in [0,T] \mapsto \overline{\sigma}^*(t,x,y) \in \mathbb{R}^d$ is the unique solution of the backward Cauchy problem*

$$\begin{cases} \partial_t \overline{\sigma}^*(t,x,y) = -\nabla_x \mathcal{F}\big(t,\theta_t^*,\Phi^*_{(T,t)}(x)\big)^\top \overline{\sigma}^*(t,x,y) \\ \overline{\sigma}^*(T,x,y) = -\nabla_x \ell(x,y). \end{cases} \tag{3.100}$$

Proof. By definition of the barycentric projection, it is clear from the fact that $\sigma^*_{x,y}(T) = \delta_{(-\nabla\ell(x,y))}$ that $\overline{\sigma}^*(T,x,y) = -\nabla_x \ell(x,y)$ for μ_T^*-almost every $(x,y) \in \mathbb{R}^{2d}$. Moreover following the construction detailed in Proposition 3.8, it holds for any $\xi \in C_c^\infty(\mathbb{R}^{2d})$ that

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^{2d}} \xi(r,s)\mathrm{d}\sigma^*_{t,x,y}(r,s) = \int_{\mathbb{R}^{2d}} \Big\langle \nabla_r \xi(r,s), -\nabla_x \mathcal{F}\big(t,\theta_t^*,\Phi^*_{(T,t)}(x)\big)^\top r \Big\rangle \mathrm{d}\sigma^*_{t,x,y}(r,s) \tag{3.101}$$

for almost every $t \in [0,T]$. We can in particular choose test functions of the form $\xi(r,s) = \zeta(r)\phi(s)$ for some $\zeta, \phi \in C_c^\infty(\mathbb{R}^d)$. Then given an arbitrary $h \in \mathbb{R}^d$, consider ζ, ϕ to be smooth functions such that

$$\zeta(r) = \begin{cases} \langle h, r\rangle & \text{if } |r| \le R_T', \\ 0 & \text{if } |r| \ge R_T' + 1, \end{cases} \qquad \text{and} \qquad \phi(s) = \begin{cases} 1 & \text{if } |s| \le R_T', \\ 0 & \text{if } |s| \ge R_T' + 1, \end{cases}$$

for all $(r,s) \in \mathbb{R}^{2d}$. It then holds that $\nabla_r \xi(r,s) = \phi(s)\nabla\zeta(r) = h$ for every $(r,s) \in K'$, which upon recalling that $\mathrm{supp}(\sigma^*_{t,x,y}) \subset K'$ for all times $t \in [0,T]$ yields together with (3.101) that

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle h, \overline{\sigma}^*(t,x,y)\rangle = \Big\langle h, -\nabla_x \mathcal{F}\big(t,\theta_t^*,\Phi^*_{(T,t)}(x)\big)^\top \overline{\sigma}^*(t,x,y)\Big\rangle,$$

for almost every $t \in [0,T]$. Since $h \in \mathbb{R}^d$ is arbitrary, we can indeed conclude that the map $t \in [0,T] \mapsto \overline{\sigma}^*(t,x,y) \in \mathbb{R}^d$ is a solution of the Cauchy problem (3.100). The uniqueness follows from Assumption 2 together with classical Grönwall estimates. □

Proof of Proposition 3.9. Following Proposition 3.3, we recall that the adjoint variable ψ^* of the Lagrangian approach is defined via the method of characteristics, namely

$$\psi^*(t,x,y) := \ell\big(\mathbf{\Phi}^*_{(t,T)}(x,y)\big) = \ell\big(\Phi^*_{(t,T)}(x),y\big),$$

for all $(t,x,y) \in [0,T] \times \mathbb{R}^{2d}$. Differentiating with respect to $x \in \mathbb{R}^d$ in the previous expression, further we obtain

$$\nabla_x \psi^*(t,x,y) = \nabla_x \Phi^*_{(t,T)}(x)^\top \nabla_x \ell\big(\Phi^*_{(t,T)}(x),y\big).$$

Evaluating this expression at $\Phi^*_{(T,t)}(x)$ for some $(x,y) \in \mathrm{supp}(\mu^*_T)$, the previous identity reads

$$\nabla_x \psi^*\big(t, \Phi^*_{(T,t)}(x), y\big) = \nabla_x \Phi^*_{(t,T)}\big(\Phi^*_{(T,t)}(x)\big)^\top \nabla_x \ell(x,y),$$

for all times $t \in [0,T]$ and μ^*_T-almost every $(x,y) \in \mathbb{R}^{2d}$. Observe now that by Lemma 3.1, the mapping $t \in [0,T] \mapsto \nabla_x \Phi^*_{(t,T)}\big(\Phi^*_{(T,t)}(x)\big)^\top \nabla_x \ell(x,y) \in \mathbb{R}^d$ is the unique solution of the backward Cauchy problem

$$\begin{cases} \partial_t w(t,x,y) = -\nabla_x \mathcal{F}\big(t, \theta^*_t, \Phi^*_{(T,t)}(x)\big)^\top w(t,x,y), \\ w(T,x,y) = \nabla_x \ell(x,y). \end{cases}$$

By standard Cauchy-Lipschitz uniqueness, this allows us to conclude that $\nabla_x \psi^*\big(t, \Phi^*_{(T,t)}(x), y\big) = -\overline{\sigma}^*(t,x,y)$ for all times $t \in [0,T]$ and μ^*_T-almost every $(x,y) \in \mathbb{R}^{2d}$, which in particular yields

$$\begin{aligned} \theta^*_t &= \int_{\mathbb{R}^{2d}} \nabla_\theta \mathcal{F}(t, \Phi^*_{(T,t)}(x), \theta^*_t)^\top \overline{\sigma}^*(t,x,y) \mathrm{d}\mu^*_T(x,y) \\ &= -\int_{\mathbb{R}^{2d}} \nabla_\theta \mathcal{F}(t, \Phi^*_{(T,t)}(x), \theta^*_t)^\top \nabla_x \psi^*\big(t, \Phi^*_{(T,t)}(x), y\big) \mathrm{d}\mu^*_T(x,y) \\ &= -\int_{\mathbb{R}^{2d}} \nabla_\theta \mathcal{F}(t, x, \theta^*_t)^\top \nabla_x \psi^*\big(t,x,y\big) \mathrm{d}\mu^*_t(x,y) \end{aligned}$$

for almost every $t \in [0,T]$, and concludes the proof of our claim. □

We can now conclude this section with the following summarizing result, Theorem 1.1.

Theorem 3.10. *For any given $T > 0$, let $\mathcal{F}$ satisfy the Assumption 2 and 3, the initial data $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$, and the terminal condition ψ_T satisfy* (3.18)*. Assume further that $\lambda > 0$ is large enough. Then, an admissible control $\theta^* \in L^2([0,T], \mathbb{R}^m)$ fulfills the mean-field PMP* (3.10)-(3.12) *if and only if it is optimal. In addition, such an optimal control θ^* is uniquely determined and Lipschitz continuous.*

Proof. The result follows by combining Theorem 3.1 and Theorem 3.6. □

4 Numerical experiments

We conclude this paper with a few instructive numerical experiments, which highlight the features of a shooting method for the mean-field maximum principle. Extensive discussions on other numerical implementations and experiments are reported in [11, 42, 50, 51]. In these works, impressive results in high dimensions have been presented and discussed, while in the present work we would like to focus more simply on understanding the mechanism of the algorithm and the interplay of its different parameters. Hence, we look at insightful examples in 1D and 2D, in order to give a simple and immediate explanation of how our method can be employed for a classification task, which is a typical application of deep learning methods. While we focus on moderate dimensions, we believe that our findings are general enough to explain the functioning of the algorithm also for higher dimensional data, such as images, and we refer to the above mentioned papers for more details.

4.1 General setting

Shooting techniques are often used to solve deterministic optimal control problems by reducing them locally to finite dimensional equations, which are solved repeatedly for different initial values that are iteratively updated. In our case, we start with an initial random guess of the control parameter $(\theta_t^0)_{t\in[0,T]}$, we solve the optimality conditions (3.10),(3.11) and (3.12) in order to update the control parameter to $(\theta_t^1)_{t\in[0,T]}$, and then use the latter as a datum for the second iteration of the shooting method. This process, more formally written as the update policy $\theta_t^{n+1} = \Lambda(\theta_t^n)$, is repeated iteratively until convergence. In the proof of Theorem 3.1, we showed that such iterations are contractive as soon as they remain bounded, and provided that the regularization parameter $\lambda > 0$ is enough. Therefore by construction, the convergence of the shooting scheme is readily guaranteed for bounded iterations. Moreover, Corollary 3.4 also ensure the convergence of the empirical solutions obtained for N particles/samples as $N \to \infty$. We illustrate the numerical method in Algorithm 1.

Algorithm 1 Shooting Method

1: Initialize the layers $\theta^0 = (\theta_t^0)_{t\in[0,T]}$

2: **for** k = 0 ... number of iterations **do**

3: Find a curve $t \in [0,T] \mapsto \mu_t^k$ which solves the forward equation

$$\partial_t \mu_t^k + \nabla_x \cdot \left(\mathcal{F}(t,x,\theta_t^k)\mu_t^k\right) = 0, \qquad \mu_t^k|_{t=0} = \mu_0\,. \tag{4.1}$$

4: Find a curve $t \in [0,T] \mapsto \psi_t^k$ which solves the backward equation

$$\partial_t \psi_t^k + \nabla_x \psi_t^k \cdot \mathcal{F}(t,x,\theta_t^k) = 0, \qquad \psi_t^k(x,y)|_{t=T} = |x-y|^2\,. \tag{4.2}$$

5: Find a new set of layers $(\theta_t^{k+1})_{t\in[0,T]}$ by solving

$$\theta_t^{k+1} + \frac{1}{2\lambda}\int_{\mathbb{R}^{2d}} \nabla_x \psi_t^k(x,y) \cdot \nabla_\theta \mathcal{F}(t,x,\theta_t^{k+1}) \mathrm{d}\mu_t^k(x,y) = 0\,. \tag{4.3}$$

6: **end for**

As already mentioned in the introduction, the dynamics (4.1) is a linear transport equation that describes the forward pass of the initial data through the network. As such, it can be solved, e.g., via a particle method: Given an initial distribution μ_0, we can sample N particles and their corresponding labels and evolve them in time for $t \in (0,T]$ according to their governing ODEs

$$\frac{\mathrm{d}X_i(t)}{\mathrm{d}t} = \mathcal{F}(t, X_i(t), \theta_t), \qquad \frac{\mathrm{d}Y_i(t)}{\mathrm{d}t} = 0$$

where $X_i(t) \in \mathbb{R}^d$ is the position of i-th sampled particle and $Y_i(t) \in \mathbb{R}^d$ is its label at time – or equivalently on the layer – $t \in [0,T]$.

The backward equation (4.2) is independent from the forward evolution (4.1) and, as such, it can be solved simultaneously. Observe that (4.2) is also a transport equation, but it is defined backward in time since a boundary condition is prescribed at the final time $t = T$. As the

terminal condition is a continuous function, finite differences in space and an explicit time-scheme can be used to solve this equation. In particular, the upwind method has been used to perform the space discretization of the backward equation. Not only is this method suitable for transport equations, but it is also ideal in this case where the flow velocity $\mathcal{F}$ depends on both x and t, i.e., it can change at every point of the domain.

Finally, given the primal-dual solutions (μ_t, ψ_t) of equations (4.1)-(4.2) respectively, we can solve (4.3). To do so, we compute for each $t \in [0, T]$ the root of the following function

$$f(\theta_t) = \theta_t + \frac{1}{2\lambda} \int_{\mathbb{R}^{2d}} \nabla_x \psi_t(x, y) \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t) \mathrm{d}\mu_t(x, y). \tag{4.4}$$

The integral with respect to μ_t is explicitly computed via the sum over the number of particles N as μ_t is an empirical/atomic distribution in the present context. Moreover, given all the discrete values of $\psi_t(x, y)$ that have been obtained by employing finite differences to solve the backward equation (4.2), the function $\psi_t(x, y)$ and its gradient $\nabla_x \psi_t(x, y)$ can be interpolated, e.g., using splines, in order to be able to evaluate these functions in whatever position $X_i(t)$ the particles may be located at in the domain. So, the fixed point equation (4.3) can be approximated by

$$f(\theta_t) \approx \theta_t + \frac{1}{2\lambda N} \sum_{i=0}^{N} \nabla_x \psi_t(X_i(t), Y_i(t)) \cdot \nabla_\theta \mathcal{F}(t, X_i(t), \theta_t), \tag{4.5}$$

whose root can be found using any root-finding algorithm such as Newton-Raphson, Bisection, or Brent's method, depending on the particular test case we work on. Here, the exclusive source of approximation is the interpolation error of ψ_t.

4.2 Results

In this section, we will show how the three optimality conditions, namely forward, backward, and parameter update ((4.1), (4.2), (4.3) respectively) are used to solve a classification task: we are given an initial distribution μ_0 of data and labels, where any point with first coordinate of positive sign is corresponding to a label in the corresponding orthant, while a negative orthant label is assigned to all those points with first coordinate of negative sign (in 1D we have one coordinate only). Then, our goal is to find the control parameter θ that moves the particles sampled from μ_0 in a way such that, at the final time T, all the particles with positive sign first coordinate are close to the positive orthant label and the particles with negative sign first coordinate are close to the negative orthant one. This task is performed through a neural network with $L\lfloor \frac{T}{\mathrm{d}t} \rfloor$ layers, where $\mathrm{d}t$ is the time discretization step used to solve both the forward (4.1) and the backward (4.2) equations. We will consider the layer forward map $\mathcal{F}(t, x, \theta_t) := \tanh(\theta_t x)$. In particular, we consider both the case in which $\theta_t \in \mathbb{R}^m := \mathbb{R}^{d \times d}$ for all $t \in [0, T]$, and the one in which the control variable encodes the weights and the shifts of the network, i.e. $\theta_t = (W_t, \tau_t)$, where $W_t \in \mathbb{R}^{d \times d}$ and $\tau_t \in \mathbb{R}^d$. This case corresponds to the layer forward map $\mathcal{F}(t, x, \theta_t) = \tanh(W_t x + \tau_t)$. The test cases for the initial distribution are the followings.

- *Bimodal Gaussian in 1D and 2D:* in the monodimensional case, the initial distribution μ_0 is a bimodal Gaussian, the particles sampled from it are concentrated around the points 1 and -1 and are assigned to the label $y = 2$ if they have a positive sign, or to the label

$y = -2$ if they have a negative sign. Similarly, in the bidimensional case the particles are initially concentrated around $(-1, -1)$ and $(+1, +1)$, but now their labels are assigned according to the sign of their first coordinate, i.e. if $X_i(0) = (X_i^1(0), X_i^2(0))$ is the initial position of the i-th particle, then this will have label $(-2, -2)$ if $X_i^1(0) < 0$ and label $(+2, +2)$ if its coordinate $X_i^1(0)$ is positive.

- *Unimodal Gaussian in 1D and 2D*: since in the previous case the initial particles are already well-separated in the respective orthant, we also perform also the classification of the particles sampled from an initial unimodal Gaussian centered in the origin that have corresponding positive label $+1$ and negative label -1 in the monodimensional case. Similarly as before, in the bidimensional case, the particles with positive first coordinate are assigned to a positive label $(+1, +1)$ and to a negative label $(-1, -1)$ when their first coordinate is negative.

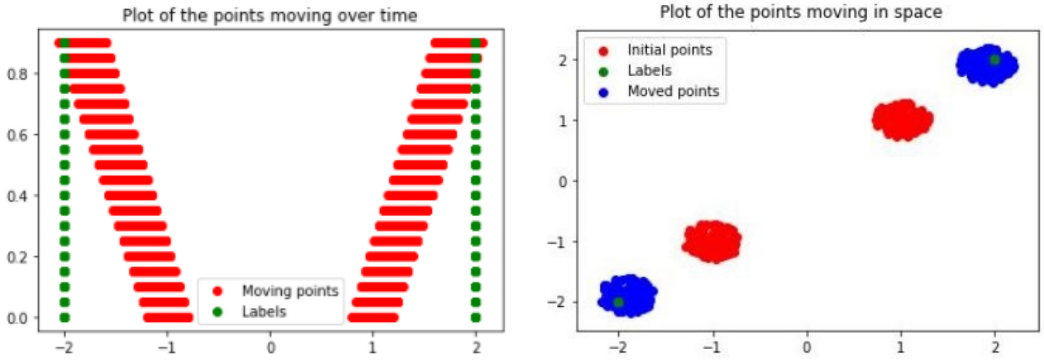

Figure 2: Left: Evolution in time of the particles from the monodimensional initial bimodal distribution μ_0 to μ_T; Right: Plot of the initial bidimensional bimodal distribution μ_0 and the final distribution μ_T.

Figure 2 shows the results obtained in the case of the bimodal distribution in 1D (on the left) and its corresponding bidimensional case (on the right). In both cases, $T = 1$ and $\mathrm{d}t = 0.05$ which corresponds to a neural network with $L = 20$ layers, and both the layer forward maps with or without biases are used. The initial guess of θ^0 is $\theta_t^0 = 0$ for all $t \in [0, T]$ and the parameter λ is set to 0.1. The forward equation (4.1) is solved using $N = 200$ particles, and the backward equation (4.2) is solved in the same domain as the forward equation, namely $x \in R_T \subset \mathbb{R}$ where R_T is defined as in (2.12). The y variable is taken in a subset of $\mathbb{R}$ as large as R_T and the same space discretization in the data dimension x and labels dimension y is used, i.e. $dx = dy = 0.1$. The same holds for the bidimensional case, where $y \in \mathbb{R}^2$ and hence the space discretization steps $dx_1 = dx_2 = dy_1 = dy_2 = 0.1$ are chosen. Finally, the root of the function in equation (4.5) is found using Brent's method and then the shooting method is applied for a total of 15 (outer) iterations.

The results obtained in the case of an initial unimodal distribution in 1D and 2D are presented in Figure 3, respectively, left and right plots. The same parameters (namely number of layers, number of particles, space and time discretization, initial guess of θ^0, and number of iterations of the shooting method) can be used in the unimodal case. The only parameter

that changes is λ which is set to 0.001 in the monodimensional case, and to 0.0001 in the bidimensional one. The case of unimodal Gaussian is more difficult than the bimodal one as the particles are really close to the splitting point, i.e., the origin, and it might happen that during an iteration of the shooting method some of the values of θ_t that are obtained move the particles to the other orthant, which will consequently lead these particles to be attracted to the wrong label. We notice that this behavior sometimes happens, but the particles generally learn to split into two groups and move to the proper labels, as depicted in Figure 3. In particular, some particles appear to be a bit isolated from the others, even if they go in the direction of the labels: these are precisely those "confused" particles that were first moved to the opposite orthant and then attracted to the wrong label.

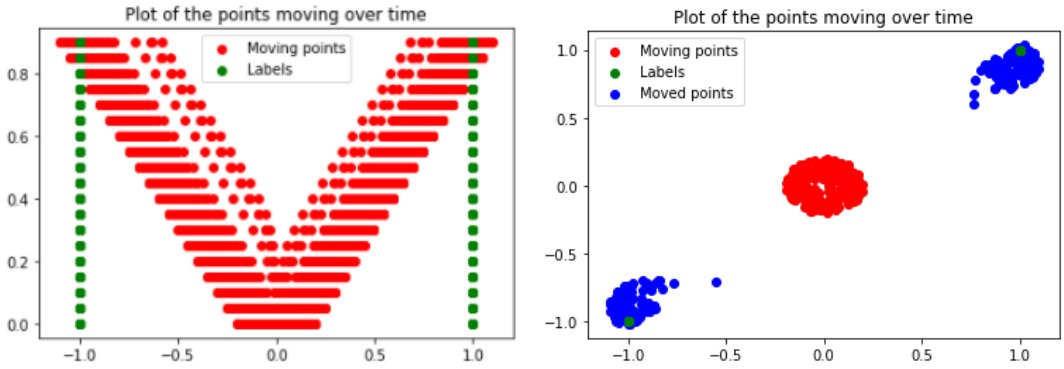

Figure 3: Left: Evolution in time of the particles from the monodimensional initial unimodal distribution μ_0 to μ_T; Right: Plot of the initial bidimensional unimodal distribution μ_0 and the final distribution μ_T.

Contribution of the number of iterations of the shooting method. We now test how many iterations of the shooting method are necessary to obtain a good result by first starting from the initial guess $\theta_t^0 \equiv 0$, and then from another initial guess $\theta^0 \equiv 1$, which is closer to the optimal solution. In the case of zero initial guesses, our experiments show that after only one iteration of the shooting method, a reasonable result for θ is obtained, meaning that the parameter is constant in t but manages to move the particles in the location of the labels. At the second iteration of the shooting method, the newly learned parameter θ decreases in time and, after the third iteration, it remains stable to the values previously found, i.e. it converges to a control parameter that correctly moves the particles to the exact location of the labels. While in the case of initial guess close to the optimal solution, i.e. θ identically equal to one, already at the first iteration the θ that is obtained decreases in time and stabilises to the proper values. Hence, from both two cases, it is clear that it is not necessary to perform many iterations of the shooting method, even while starting from an initial guess $(\theta_t^0)_{t\in[0,T]}$ that is far away from the optimal solution. On the left of Figure 4, the L^2 distance between shooting method solutions, denoted by $\epsilon(k) = \left\|\theta_t^{k+1} - \theta_t^k\right\|_2$, is plotted for each $k = 0, ...,$number of iterations, starting from different initial guesses θ_t^0. It appears that independently from the initial guess, the distance between consecutive solutions goes to zero in a few iterations (which is also shown on the right of Figure 4 where, after the second iteration, it becomes impossible to distinguish between

consecutive solutions) with different velocities depending on the initial guess. This seems to be an advantage with respect to typical deep learning methods that require many more iterations to find the minimum of the loss function.
Moreover, it is interesting to notice that θ decreasing in time means that the particles at the beginning are moving faster in the direction of the labels and then when they are close enough, they slow down to precisely stop at labels' position. The dynamics of the iterations, in the monodimensional case in which an initial bimodal Gaussian is fed to a network with layer forward map $\mathcal{F}(t, x, \theta_t) = \tanh(\theta_t x)$, is depicted in the plot on the right of Figure 4, where the initial guess is $\theta_t = 1 \ \forall t$.

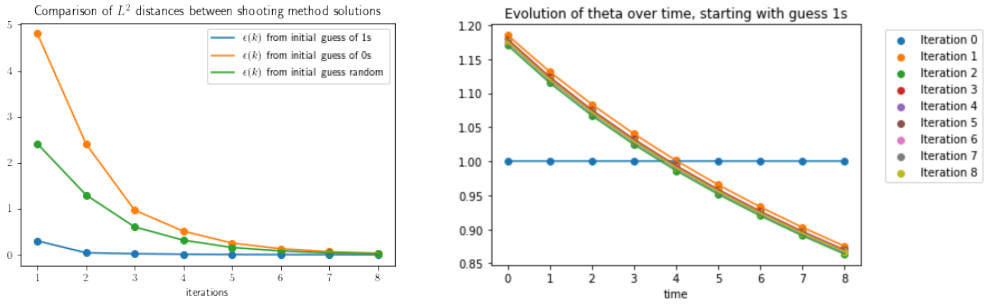

Figure 4: Left: L^2 distance between successive solutions of the shooting method over the number of iterations, and starting from different initial guess, namely $\theta_t^0 = 1$, $\theta_t^0 = 0$, and $\theta_t^0 = r_t$, $r_t \sim \mathcal{U}(0, 1)$ for all t; Right: values of θ_t over time, starting from initial guess $\theta_t^0 = 1$ for all t.

On the effect of the parameter λ. A very fundamental factor that has to be taken into consideration is that of the regularization parameter λ, appearing in the fixed-point equation (4.3) of the optimality conditions. The parameter λ is a real positive number decided a priori, that determines the influence of the regularization term in the loss function (1.5), and hence monitors how large the L^2-norm of θ is allowed to be. In particular, since the layer forward map $\mathcal{F}$ depends on θ, its norm highly influences the velocity flow of the particles in the forward equation. Hence, if the initial distribution μ_0 of the particles is far away from the labels, λ needs to be set to a small value – e.g. 0.1 –, to allow $\|\theta\|_2$ to be large enough to reach the labels, otherwise the particles will not have enough velocity to arrive to the correct location at time $T > 0$. However, always choosing a small λ is not a good technique either. Indeed, our experiments show that small values of λ cause the fixed point mapping $f(\theta_t)$ of (4.3) to have many steep picks, which makes it impossible to use derivative-based methods such as Newton's algorithm to find its root. In case of exceedingly small λ, this can even lead to a function $f(\theta_t)$ with multiple roots, which may cause the algorithm to oscillate between solutions, also reflecting the potential loss of uniqueness of optimal controls.

Let us now look at an instructive example, in which a unimodal monodimensional Gaussian centered in zero is fed to a neural network that has layer forward map without biases. In Figure 3 (left plot), λ is set to 0.01, leading to a correct solution. But in Figure 5, we notice that if λ is set to be too large, then $\|\theta\|_2$ is not large enough to move the particles to the location of the labels and thus we obtain the behavior on the left of Figure 5 where the particles are moving

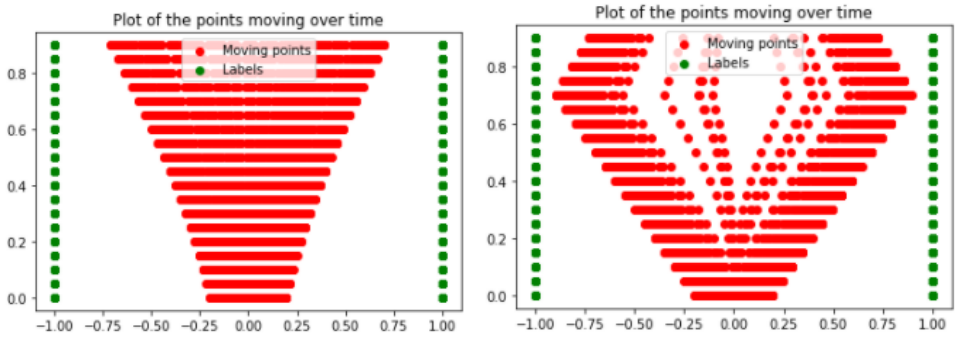

Figure 5: Left: unimodal initial distribution, case with $\lambda = 0.1$; Right: unimodal initial distribution, case with $\lambda = 0.0001$

in the correct direction but not quickly enough. On the contrary, if λ is too small, the control θ_t obtained at every iteration of the shooting method leads to an oscillating behavior between the correct result and another solution, which is shown on the right of Figure 5. In this case, the particles arrive too quickly to the labels, i.e. for $t < T$, due to the fact that small values of λ lead to large control magnitudes $\|\theta\|_2$, which influences the velocity of the particles. At this point, the method should be able to learn a θ_{t+1} which stops the particles in order to remain at the position of the labels, but again the small value of λ does not push easily θ_{t+1} to be zero and allows the norm of θ to remain large. As a result, the particles, instead of remaining in the location of the labels, start simply moving in the opposite direction. This behavior is not surprising as it is in accordance with Remark (3.4), for which λ needs to be set to a large value, but the precise quantity that is needed depends on the initial distribution of μ_0 and the domain C_Γ in which the root can be found. Indeed, in the simpler case of a bimodal Gaussian initial distribution λ does not have to be too small (recall that it was set to 0.1 to produce the plot on the left of Figure 2), but in the more challenging case of a unimodal Gaussian initial distribution, its value has to be small enough to give the necessary velocity to the particles in order to let them split and reach the labels (indeed $\lambda = 0.001$ in the case on the left of Figure 3). Moreover, these considerations still hold in case of activation function with bias: in this case, the parameter can be split in two $\lambda = [\lambda_0, \lambda_1]$, set to different values in order to control separately the norm of W and the one of τ, which is fundamental when the Gaussian is centered in zero and the optimal W should be greater than 1, while the optimal τ should be zero.

Influence of the time and space discretization. A first remark in connection with the role of λ regards the number of layers of the neural network, hence the time discretization $\mathrm{d}t$ step. Figure 6 shows an experiments in 2D: starting from the bimodal distribution and the same initial guess θ^0, the shooting method is repeated 15 times with $\lambda = 0.1$ and $\mathrm{d}x = 0.1$. The difference between the three plots in Figure 6 is that different numbers of layers are employed, i.e., $\mathrm{d}t = 0.2, 0.1, 0.05$ respectively from left to right. Clearly, the case with $\mathrm{d}t = 0.05$ is the one that works best, because if $\mathrm{d}t$ is too large, the particles do not have enough time to reach the labels (as in the case with $\mathrm{d}t = 0.2$, i.e 5 layers) or they reach them, but not completely (as in the case with $\mathrm{d}t = 0.1$, i.e. 10 layers). These issues can clearly be overcome by using a smaller λ, but considering the difficulty in tuning λ, it is more convenient to increase the number of

layers instead. This is consistent with the common technique in the deep learning community to increase the number of layers to obtain better results.

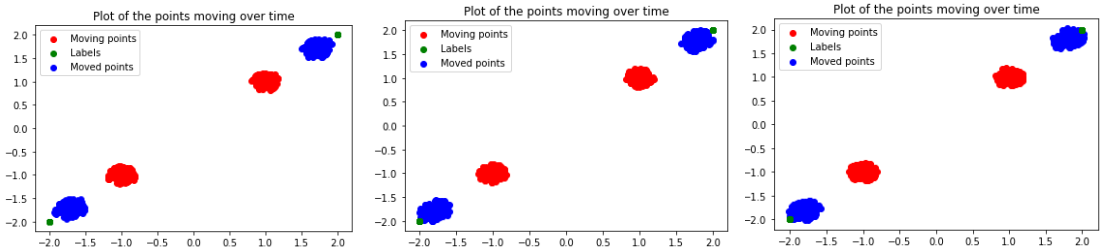

Figure 6: Left: bimodal initial distribution in 2D with dt=0.2; Center: bimodal initial distribution in 2D with dt=0.1; Right: bimodal initial distribution in 2D with dt=0.05

Moreover, we need to keep in mind that the time discretization has to be chosen in accordance with the space discretization $\mathrm{d}x$ appearing in the backward equation, as the Courant number has to be kept below 1 in order to guarantee convergence of the numerical scheme. It is interesting to notice that in the case of unimodal distribution, increasing the space discretization to $\mathrm{d}x = dy = 0.2$ is surprisingly beneficial. This is because the Courant number that needs to be set to a value between 0 and 1, but not too close to neither of them, depends on the function $\mathcal{F}(t, x, \theta_t)$ and, since all the particles $X_i(0)$ are initially close to zero, this number tends to be too small. Hence, convergence is obtained when the space discretization is increased.

An implementation in Python of our algorithm, together with videos and code to reproduce our results, can be found at the following repository *https://github.com/CristinaCipriani/MeanfieldPMP-NeurODE-training*.

Appendix

In this section, we collect the proof of auxiliary results of the paper.

Proof of Theorem 2.3. Recall the associated characteristic system of ODEs

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = \mathcal{F}(t, X_t, \theta_t), \quad \frac{\mathrm{d}Y_t}{\mathrm{d}t} = 0\,. \tag{4.6}$$

Since for any given $\theta \in L^2([0,T];\mathbb{R}^m)$, the velocity field $\mathcal{F}$ satisfies the regularity and growth conditions in Assumption 1, standard results for Carathéodory differential equations [34, Chapter 1] ensure that for any initial condition $(x_0, y_0) \in B(R)$, the above system has a unique solution $(X_t, Y_t) \in \mathrm{Lip}([0,T];\mathbb{R}^{2d})$ on $[0,T]$. Moreover following e.g. [37, Theorem A.2], it holds

$$|X_t| \leq \left(|x_0| + \int_0^t h(s)\,\mathrm{d}s\right) e^{\int_0^t h(s)\,\mathrm{d}s} \leq (R + C_{\mathcal{F},T}) e^{C_{\mathcal{F},T}} \quad \text{and} \qquad Y_t = y_0\,, \tag{4.7}$$

for all $t \in [0,T]$. We consider the underlying characteristic flow between times $\tau, t \in [0,T]$, defined by

$$\mathbf{\Phi}^{\mathcal{F}}_{(\tau,t)} : \ \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d, \quad (x_\tau, y_\tau) \mapsto \mathbf{\Phi}^{\mathcal{F}}_{(\tau,t)}(x_\tau, y_\tau) = (X_t^{x_\tau}, Y_t^{y_\tau})\,, \tag{4.8}$$

where (X^{x_0}, Y^{y_0}) is the solution to (4.6) with the initial data $(x_\tau, y_\tau) \in \mathbb{R}^{2d}$ at time $\tau \in [0,T]$. Setting $\tau = 0$ and given a datum $\mu_0 \in \mathcal{P}_c^a(\mathbb{R}^{2d})$, we can use the characteristic flow to define the following curve of measures

$$\mu_t := \mathbf{\Phi}^{\mathcal{F}}_{(0,t)} \sharp \mu_0 \,, \tag{4.9}$$

for all times $t \in [0,T]$, which equivalently means that

$$\int_{\mathbb{R}^{2d}} \psi(t,x,y) \mathrm{d}\mu_t(x,y) = \int_{\mathbb{R}^{2d}} \psi(t, X_t^{x_0}, Y_t^{y_0}) \mathrm{d}\mu_0(x_0,y_0) \,. \tag{4.10}$$

for all $\psi \in \mathcal{C}_b^1([0,T] \times \mathbb{R}^{2d})$. It is well known that μ_t is a measure solution to the equation (1.7). Indeed, using the change of variables formula for the push-forward measure, the chain rule, and once more the change of variables, one has

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^{2d}} \psi(t,x,y) \mathrm{d}\mu_t(x,y) &= \int_{\mathbb{R}^{2d}} \frac{\mathrm{d}}{\mathrm{d}t} \psi(t, X_t^{x_0}, Y_t^{y_0}) \mathrm{d}\mu_0(x_0,y_0) \\
&= \int_{\mathbb{R}^{2d}} \Big(\partial_t \psi(t, X_t^{x_0}, Y_t^{y_0}) + \nabla_x \psi(t, X_t^{x_0}, Y_t^{y_0}) \cdot \mathcal{F}(t, X_t^{x_0}, \theta_t) \Big) \mathrm{d}\mu_0(x_0,y_0) \\
&= \int_{\mathbb{R}^{2d}} \left(\partial_t \psi(t,x,y) + \nabla_x \psi(t,x,y) \cdot \mathcal{F}(t,x,\theta_t) \right) \mathrm{d}\mu_t(x,y) \,.
\end{aligned} \tag{4.11}$$

Integrating in time leads to formula (2.10). Furthermore, it follows from [23, Lemma 3.11] that for any $s,t \in [0,T]$ it holds

$$W_1(\mu_t, \mu_s) = W_1\big(\mathbf{\Phi}^{\mathcal{F}}_{(0,t)} \sharp \mu_o, \mathbf{\Phi}^{\mathcal{F}}_{(0,s)} \sharp \mu_0\big) \leq \Big\| \mathbf{\Phi}^{\mathcal{F}}_{(0,t)} - \mathbf{\Phi}^{\mathcal{F}}_{(0,s)} \Big\|_{L^\infty(\mathrm{supp}(\mu_0))} \leq C|t-s| \,, \tag{4.12}$$

due to the fact that

$$|\mathbf{\Phi}^{\mathcal{F}}_{(0,t)}(x_0,y_0) - \mathbf{\Phi}^{\mathcal{F}}_{(0,s)}(x_0,y_0)| = |(X_t^{x_0} - X_s^{x_0}, 0)| \leq C|t-s|, \tag{4.13}$$

for all $(x_0,y_0) \in \mathrm{supp}(\mu_0)$, where C depends only on R and $C_{\mathcal{F},T}$. Thus μ_t is continuous in time with respect to W_1 metric, and it has a compact support since (4.7). Moreover, supp $\mu_t \in B(R_T)$ for all $t \in [0,T]$, where R_T depends only on R and $C_{\mathcal{F},T}$.

Next we prove the stability estimate. Let μ_t^i, $i = 1,2$ be measure solutions to the equation (1.7) with initial data μ_0^i obtained above. Write $(X_t^i, Y_t^i) := \mathbf{\Phi}^{\mathcal{F}}_{(0,t)}(x_0^i, y_0^i)$ for $i = 1,2$ and $t \in [0,T]$, and observe that

$$\begin{aligned}
\big|(X_t^1, Y_t^1) - (X_t^2, Y_t^2)\big| &= \left| \left((x_0^1 - x_0^2) + \int_0^t \mathcal{F}(s, X_s^1, \theta_s) - \mathcal{F}(s, X_s^2, \theta_s) \,\mathrm{d}s \,,\, y_0^1 - y_0^2 \right) \right| \\
&\leq |(x_0^1 - x_0^2, y_0^1 - y_0^2)| + \int_0^t \big| \mathcal{F}(s, X_s^1, \theta_s) - \mathcal{F}(s, X_s^2, \theta_s) \big| \,\mathrm{d}s \\
&\leq |(x_0^1, y_0^1) - (x_0^2, y_0^2)| + \int_0^t g_R(1 + |\theta_s|) |X_s^1 - X_s^2| \,\mathrm{d}s \,.
\end{aligned}$$

Applying Gronwall's inequality leads to

$$\big|(X_t^1, Y_t^1) - (X_t^2, Y_t^2)\big| \leq \big|(x_0^1, y_0^1) - (x_0^2, y_0^2)\big| e^{\int_0^t g_R(1+|\theta_s|) \,\mathrm{d}s} = |(x_0^1, y_0^1) - (x_0^2, y_0^2)| e^{L_{\mathcal{F},T,R,\|\theta\|_1}} \,, \tag{4.14}$$

for all times $t \in [0,T]$. Therefore, $\mathbf{\Phi}^{\mathcal{F}}_{(0,t)}$ is Lipschitz on $B(0,R) \subset \mathbb{R}^{2d}$, and

$$|\mathbf{\Phi}^{\mathcal{F}}_{(0,t)}(x_0^1, y_0^1) - \mathbf{\Phi}^{\mathcal{F}}_{(0,t)}(x_0^2, y_0^2)| \leq L_{\mathcal{T}} |(x_0^1, y_0^1) - (x_0^2, y_0^2)| \quad \text{for all } t \in [0,T] \,, \tag{4.15}$$

where $L_{\mathcal{T}} = e^{L_{\mathcal{F},T,R,\|\theta\|_1}}$. Given an optimal transport plan π between μ_0^1 and μ_0^2, one can check that the measure $\gamma = (\mathbf{\Phi}^{\mathcal{F}}_{(0,t)} \times \mathbf{\Phi}^{\mathcal{F}}_{(0,t)})\sharp\pi$ has marginals $\mathbf{\Phi}^{\mathcal{F}}_{(0,t)}\sharp\mu_0^1$ and $\mathbf{\Phi}^{\mathcal{F}}_{(0,t)}\sharp\mu_0^2$. Hence, it holds

$$\begin{aligned} W_1\Big(\mathbf{\Phi}^{\mathcal{F}}_{(0,t)}\sharp\mu_0^1, \mathbf{\Phi}^{\mathcal{F}}_{(0,t)}\sharp\mu_0^2\Big) &\le \int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}} |x-y| \mathrm{d}\gamma(x,y) \\ &= \int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}} \big|\mathbf{\Phi}^{\mathcal{F}}_{(0,t)}(x) - \mathbf{\Phi}^{\mathcal{F}}_{(0,t)}(y)\big| \mathrm{d}\pi(x,y) \\ &\le L_{\mathcal{T}} \int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}} |x-y| \mathrm{d}\pi(x,y) = L_{\mathcal{T}} W_1(\mu_0^1, \mu_0^2)\,, \end{aligned}$$

which leads to

$$W_1(\mu_t^1, \mu_t^2) = W_1(\mathbf{\Phi}^{\mathcal{F}}_{(0,t)}\sharp\mu_0^1, \mathbf{\Phi}^{\mathcal{F}}_{(0,t)}\sharp\mu_0^2) \le L_{\mathcal{T}} W_1(\mu_0^1, \mu_0^2), \tag{4.16}$$

for all times $t \in [0,T]$. Thus we have completed the proof. □

Proof of Proposition 3.3. We shall use the standard characteristic method with backward propagation. Thanks to Cauchy-Lipschitz theorem on ODE, we know that for any terminal condition $(\overline{X}_T, \overline{Y}_T) = (x,y) \in B(R_T)$

$$\frac{\mathrm{d}\overline{X}_t}{\mathrm{d}t} = \mathcal{F}(t, \overline{X}_t, \theta_t), \quad \frac{\mathrm{d}\overline{Y}_t}{\mathrm{d}t} = 0\,, \tag{4.17}$$

admits a unique solution $t \in [0,T] \mapsto (\overline{X}_t, \overline{Y}_t) := \mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(x,y) \in \mathbb{R}^{2d}$ which can be written explicitly as

$$\mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(x,y) = \Big(x - \int_t^T \mathcal{F}(s, \overline{X}_s, \theta_s)\,\mathrm{d}s\,,\, y\Big).$$

for all $(x,y) \in B(R_T)$. Moreover, one has $\big|\mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(x,y)\big| \le (R_T + C_{\mathcal{F},T}T)e^{C_{\mathcal{F},T}T} + R_T$ by Gronwall's inequality as in (4.7). Namely one has $\mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(x,y)(B(R_T)) \subset B(R_T)$ with the diameter $R_T' = R + (R + C_{\mathcal{F},T}T)e^{C_{\mathcal{F},T}T}$.

Furthermore, the functions $\mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ are $\mathcal{C}^2$ diffeomorphisms for any $t \in [0,T]$, and the application $(t,x,y) \mapsto \mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(x,y) \in \mathbb{R}^{2d}$ is globally Lipschitz. Now we construct solution by using standard characteristic method, and we define

$$\psi^\theta(t,x,y) := \psi_T\big(\mathbf{\Phi}^{\mathcal{F}_\theta}_{(t,T)}(x,y)\big), \quad (x,y) \in \mathbb{R}^{2d}\,, \tag{4.18}$$

where $\psi_T \in \mathcal{C}^2_c(\mathbb{R}^{2d})$ satisfies (3.18). This implies that $\psi^\theta\big(t, \mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(x,y)\big) = \psi_T(x,y)$. Then we deduce

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}t}\psi^\theta\big(t, \mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(x,y)\big) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\psi^\theta(t, \overline{X}_t, \overline{Y}_t) \\ &= \partial_t\psi^\theta(t, \overline{X}_t, \overline{Y}_t) + \nabla_x\psi^\theta(t, \overline{X}_t, \overline{Y}_t)\cdot\frac{\mathrm{d}\overline{X}_t}{\mathrm{d}t} = \Big(\partial_t\psi^\theta + \nabla_x\psi^\theta\cdot\mathcal{F}\Big)\big(t, \mathbf{\Phi}^{\mathcal{F}_\theta}_{(0,t)}(x,y)\big)\,. \end{aligned}$$

for any $t \in [0,T)$ and $(x,y) \in \mathbb{R}^{2d}$. Since $\mathrm{supp}(\psi_T) = B(R_T)$, one has $\mathrm{supp}(\psi^\theta(t)) = \mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(B(R_T)) \subset B(R_T')$ for all $t \in [0,T]$. Thus we have constructed a function $\psi^\theta(t,x,y) = \psi_T\big(\mathbf{\Phi}^{\mathcal{F}_\theta}_{(t,T)}(x,y)\big)$ of class $\mathcal{C}^1([0,T]; \mathcal{C}^2_c(\mathbb{R}^{2d}))$ satisfying (3.11).

Now for all $(x,y) \in \mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(x,y)(B(R_T)) \subset B(R'_T)$, let us consider the following ODEs

$$\frac{\mathrm{d}X_s}{\mathrm{d}s} = \mathcal{F}(s, X_s, \theta_s), \quad \frac{\mathrm{d}Y_s}{\mathrm{d}s} = 0\,, \tag{4.19}$$

with the initial data $(X_s, Y_s)\big|_{s=t} = (x,y)$. Then

$$\mathbf{\Phi}^{\mathcal{F}_\theta}_{(t,T)}(x,y) = (X_T, Y_T) = \left(x + \int_t^T \mathcal{F}(s, X_s, \theta_s)\,\mathrm{d}s\,,\, y\right)$$

It follows from similar arguments as in (4.14) that

$$\left|\mathbf{\Phi}^{\mathcal{F}_\theta}_{(t,T)}(x_1,y_1) - \mathbf{\Phi}^{\mathcal{F}_\theta}_{(t,T)}(x_2,y_2)\right| \leq (|x_1-x_2| + |y_1-y_2|)e^{L_{F,T,C_\Gamma}T},$$

which combined with *(1)* in Assumption 3, according to [69, Lemma 2.3], implies that

$$\left\|\mathbf{\Phi}^{\mathcal{F}_\theta}_{(t,T)}(\cdot,\cdot)\right\|_{\mathcal{C}^2(\mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(B(R_T)))} \leq C(R'_T, T, C_\Gamma, C_{\mathcal{F},T}, L_{F,T,C_\Gamma})\,. \tag{4.20}$$

Thus we have for all $t \in [0,T]$

$$\begin{aligned}\left\|\psi^\theta_t(\cdot,\cdot)\right\|_{\mathcal{C}^2_c(\mathbb{R}^{2d})} = \left\|\psi^\theta_t\right\|_{\mathcal{C}^2(\mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(B(R_T)))} &= \left\|\psi_T(\mathbf{\Phi}^{\mathcal{F}_\theta}_{(t,T)}(\cdot,\cdot))\right\|_{\mathcal{C}^2(\mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(B(R_T)))} \\ &\leq C\left(\left\|\mathbf{\Phi}^{\mathcal{F}_\theta}_{(t,T)}(\cdot,\cdot)\right\|_{\mathcal{C}^2(\mathbf{\Phi}^{\mathcal{F}_\theta}_{(T,t)}(B(R_T)))}\right)\|\psi_T\|_{\mathcal{C}^2(B(R_T))}\,.\end{aligned} \tag{4.21}$$

This concludes the proof of (3.19). □

Proof of Theorem 3.5.

• **Step 1.** We first want to show that

$$G'(x^*)h = 0 \qquad \text{implies} \qquad DJ(x^*)h = 0\,, \tag{4.22}$$

for all $h \in \overline{X}_E$. To this end, let $h \in \overline{X}_E$ be given such that $G'(x^*)h = 0$. Here $DJ(x^*)$ is the multivalued F-differential of J at x^* as in Definition 2.4. Consider the operator

$$\Psi(\varepsilon, u) := \overline{G}(x^* + \varepsilon h + u)\,, \tag{4.23}$$

where (ε, u) is in some neighborhood of $(0,0)$ in $\mathbb{R}\times\overline{X}_E$, and $\overline{G}$ is the unique extension of G to $\overline{E}$. Indeed, for any $h, u \in \overline{X}_E$, there exists sequences $(h^n)_{n\in\mathbb{N}}, (u^n)_{n\in\mathbb{N}} \subset X_E$ such that $h^n \to h$ and $u^n \to u$. According to the assumption it necessarily holds that $(x^* + \varepsilon h^n + u^n) \in x^* + X_E \subset E$, so one can uniquely define

$$\Psi(\varepsilon, u) = \overline{G}(x^* + \varepsilon h + u) := \lim_{n\to\infty} G(x^* + \varepsilon h^n + u^n)\,. \tag{4.24}$$

In the sequel we will not differentiate G from $\overline{G}$.

Note that if x^* solves (3.47), one has

$$\Psi(0,0) = G(x^*) = 0\,. \tag{4.25}$$

By the definition of F-derivatives, we note that

$$\lim_{y\to 0} \frac{\|\Psi(0,y) - \Psi(0,0) - G'(x^*)y\,\|_Y}{\|y\,\|_X} = \lim_{y\to 0} \frac{\|G(x^*+y) - G(x^*) - G'(x^*)y\,\|_Y}{\|y\,\|_X} = 0. \tag{4.26}$$

This means that $G'(x^*) \in D\Psi_u(0,0)$. Thus there exists some $\Psi'_u(0,0) \in D\Psi_u(0,0)$ such that

$$\Psi'_u(0,0) = G'(x^*), \tag{4.27}$$

Moreover $\Psi'_u(0,0)$ is surjective on $\overline{X}_E \to Y$, since $G'(x^*)$ is surjective on $\overline{X}_E \to Y$.

∘ *Step 1.1.* From above, we know that $\Psi'_u(0,0)$ is surjective on $\overline{X}_E \to Y$. Thus, there exists a number $\kappa > 0$ such that, for each $y \in Y$, there is a point $\omega(y) \in \overline{X}_E \subset X$ satisfying

$$\Psi'_u(0,0)\omega(y) = y \qquad \text{and} \qquad \|\omega(y)\|_X \le \kappa \|y\|_Y \,, \tag{4.28}$$

where the second inequality follows from Banach's continuous inverse theorem. We define

$$f(\varepsilon, u) := \Psi'_u(0,0)u - \Psi(\varepsilon, u)\,. \tag{4.29}$$

Let $\varepsilon \le \rho$ and $\|u\|_X, \|v\|_X \le r$, and observe that for some $f'_u(\varepsilon, u) \in Df_u(\varepsilon, u)$, it holds

$$f'_u(\varepsilon, u) = \Psi'_u(0,0) - \Psi'_u(\varepsilon, u)\,. \tag{4.30}$$

Since $f'_u(\varepsilon, u)$ is continuous at $(0,0)$ and $f'_u(0,0) = 0$, Taylor's theorem implies that

$$\|f(\varepsilon,u) - f(\varepsilon,v)\| \le \sup_{0\le\tau\le 1} \|f'_u(\varepsilon, u+\tau(v-u))\|\|u-v\|_X = o(1)\|\|u-v\|_X \,, \tag{4.31}$$

as $\rho, r \to 0$. In addition since $f(0,0) = 0$ and f is continuous at $(0,0)$, we also get

$$\|f(\varepsilon,u)\|_Y \le \|f(\varepsilon,u) - f(\varepsilon,0)\|_Y + \|f(\varepsilon,0)\|_Y \le o(1)\|u\|_X + \|f(\varepsilon,0)\|_Y \,, \tag{4.32}$$

as $\rho, r \to 0$. For a given $\varepsilon \in \mathbb{R}^+$ with $\varepsilon < \rho$, we consider following iterative method

$$\Psi'_u(0,0)u_{m+1} = f(\varepsilon, u_m), \quad m = 0, 1, 2, \cdots \,, \tag{4.33}$$

where $u_0 = 0$ and $u_{m+1} = \omega(f(\varepsilon, u_m))$. Since $\|u_{m+1}\|_X \le \kappa\|f(\varepsilon, u_m)\|_Y$, it follows from (4.31) and (4.32) that for sufficiently small ρ and r, one has

$$\|u_m\|_X \le o(1)r + o(1),\ \rho \to 0 \quad \text{and} \quad \|u_{m+2} - u_{m+1}\|_X \le \frac{1}{2}\|u_{m+1} - u_m\|_X \text{ for all } m = 0, 1, \cdots \,, \tag{4.34}$$

which means that $\{u_m\}_{m\ge 0}$ is a Cauchy sequence in the Banach space $\overline{X}_E$, and hence there exists some $u \in \overline{X}_E$ such that

$$u_m \to u \text{ as } m \to \infty\,. \tag{4.35}$$

Moreover we have that $\|u\|_X \le r$ and $\Psi'_u(0,0)u = f(\varepsilon, u)$ because of (4.33), and thus $\Psi(\varepsilon, u) = 0$. Lastly, we let $m \to \infty$ in

$$\|u_{m+2}\|_X \le \kappa\|f(\varepsilon, u_{m+1})\|_Y = \kappa \left\|\Psi'_u(0,0)u_{m+1} - \Psi(\varepsilon, u_{m+1})\right\|_Y \,, \tag{4.36}$$

then it follows that $\|u\|_X \le \kappa\|\Psi'_u(0,0)u\|_Y$.

∘ *Step 1.2.* It follows from Step 1.1 above that there exists numbers $\rho > 0$ and $r > 0$ such that for any $\varepsilon \in \mathbb{R}^+$ and $\varepsilon \le \rho$, there exists $u(\varepsilon) \in \overline{X}_E$ with $\|u(\varepsilon)\|_X \le r$ such that

$$\Psi(\varepsilon, u(\varepsilon)) = G(x^* + \varepsilon h + u(\varepsilon)) = 0 \tag{4.37}$$

and

$$\|u(\varepsilon)\|_X \le \kappa \left\|\Psi'_u(0,0)u(\varepsilon)\right\|_Y = \kappa \left\|G'(x^*)u(\varepsilon)\right\|_Y \tag{4.38}$$

along with $\|u(\varepsilon)\|_X \to 0$ as $\varepsilon \to 0$.

By the definition of F- derivative, one has

$$G(x^*+k) - G(x^*) - G'(x^*)k = o(\|k\|_X), \quad k \to 0\,. \tag{4.39}$$

Let $k = \varepsilon h + u(\varepsilon)$, we have

$$G(x^* + \varepsilon h + u(\varepsilon)) - G(x^*) - \varepsilon G'(x^*)h - G'(x^*)u(\varepsilon) = o(\|\varepsilon h + u(\varepsilon)\|_X), \quad \varepsilon \to 0\,. \tag{4.40}$$

Therefore

$$G'(x^*)u(\varepsilon) = o(1)\|\varepsilon h + u(\varepsilon)\|_X\,, \quad \varepsilon \to 0\,. \tag{4.41}$$

By (4.38), we obtain $\|u(\varepsilon)\|_X \le o(1)\|\varepsilon h + u(\varepsilon)\|_X$, which is

$$\|u(\varepsilon)\| = o(\varepsilon), \quad \varepsilon \to 0\,. \tag{4.42}$$

Since x^* is the minimizer of J, one has

$$J(x^* + \varepsilon h + u(\varepsilon)) \ge J(x^*)\,, \tag{4.43}$$

which yields

$$DJ(x^*)(\varepsilon h + u(\varepsilon)) + o(\|\varepsilon h + u(\varepsilon)\|_X) \ge 0, \quad \varepsilon \to 0\,. \tag{4.44}$$

Dividing by ε and letting $\varepsilon \to \pm 0$, one has $DJ(x^*)h \ge 0$ and $DJ(x^*)h \le 0$. In other words

$$DJ(x^*)h = 0\,. \tag{4.45}$$

• *Step 2.* In Step 1 we have proven that if $G'(x^*)h = 0$ for some $h \in \overline{X}_E$, then $DJ(x^*)h = 0$. This can be written in the more compact operator form

$$DJ(x^*) \subset [\mathcal{N}(G'(x^*))]^\perp = \left\{x' \in \overline{X}'_E \mid \langle x', h\rangle = 0 \text{ for all } h \in \mathcal{N}(G'(x^*)) \subset \overline{X}_E\right\}. \tag{4.46}$$

Then, it follows from the closed range theorem in Banach spaces that

$$[\mathcal{N}(G'(x^*))]^\perp = \mathcal{R}(G'(x^*)^\top)\,. \tag{4.47}$$

which implies that

$$DJ(x^*) \subset \mathcal{N}(G'(x^*))^\perp = \mathcal{R}(G'(x^*)^\top)\,.$$

Therefore, there exists a covector $p^* \in Y'$ such that $J'(x^*) = G'(x^*)^\top p^*$ for any $J'(x^*) \in DJ(x^*)$. In other words

$$\langle J'(x^*), z\rangle = \langle G'(x^*)^\top p^*, z\rangle = \langle p^*, G'(x^*)z\rangle \quad \text{ for all } z \in \overline{X}_E\,, \tag{4.48}$$

which completes the proof of Theorem 3.5. □

Proof of Lemma 3.1. By construction of the semigroups $(\Phi_{(\tau,t)})_{\tau,t\in[0,T]}$, it holds for all $(t,x) \in [0,T]\times\mathbb{R}^d$ that

$$\Phi_{(t,T)} \circ \Phi_{(T,t)}(x) = x, \tag{4.49}$$

where "$\circ$" stands for the standard composition operation between functions. Thus by differentiating with respect to $x \in \mathbb{R}^d$ in (4.49), we obtain

$$\nabla_x \Phi_{(t,T)}(\Phi_{(T,t)}(x))\nabla_x \Phi_{(T,t)}(x) = \mathrm{Id},$$

for every $y \in \mathbb{R}^d$. Thus, recalling that $\nabla_x \Phi_{(T,t)}(x)$ is invertible by construction, one further has

$$\nabla_x \Phi_{(t,T)}(\Phi_{(T,t)}(x)) = \nabla_x \Phi_{(T,t)}(x)^{-1}, \tag{4.50}$$

for every $(t,x) \in [0,T]\times\mathbb{R}^d$. Differentiating with respect to $t \in [0,T]$ in (4.50) while recalling the ODE characterization derived in (3.94) for $t \in [0,T] \mapsto \nabla_x \Phi_{(T,t)}(x)$ then yields

$$\begin{aligned}
\partial_t\Big(\nabla_x \Phi_{(t,T)}(\Phi_{(T,t)}(x))\Big) &= -\nabla_x \Phi_{(T,t)}(x)^{-1}\partial_t\Big(\nabla_x \Phi_{(T,t)}(x)\Big)\nabla_x \Phi_{(T,t)}(x)^{-1} \\
&= -\nabla_x \Phi_{(T,t)}(x)^{-1}\nabla_x v(t,\Phi_{(T,t)}(x)) \\
&= -\nabla_x \Phi_{(t,T)}(\Phi_{(T,t)}(x))\nabla_x v(t,\Phi_{(T,t)}(x)),
\end{aligned}$$

where we used the classical characterization of the differential of the inverse mapping over matrices. Taking the transpose in the previous expression while using the fact that the process of adjoining a matrix is linear, we can conclude that

$$\begin{cases}
\partial_t\Big(\nabla_x \Phi_{(t,T)}(\Phi_{(T,t)}(x))^\top\Big) = -\nabla_x v(t,\Phi_{(T,t)}(x))^\top \nabla_x \Phi_{(t,T)}(\Phi_{(T,t)}(x))^\top, \\
\quad \nabla_x \Phi_{(T,T)}(\Phi_{(T,T)}(x))^\top = \mathrm{Id},
\end{cases}$$

which ends the proof of our claim. □

Acknowledgments

C.C., H.H., and M.F. acknowledge the support of the DFG Project "Identification of Energies from Observation of Evolutions" and the DFG SPP 1962 "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization". C.C. and M.F. acknowledge also the partial support of the project "Online Firestorms And Resentment Propagation On Social Media: Dynamics, Predictability and Mitigation" of the TUM Institute for Ethics in Artificial Intelligence.

References

[1] Andrei Agrachev and Andrey Sarychev, *Control in the spaces of ensembles of points*, SIAM Journal on Control and Optimization **58** (2020), no. 3, 1579–1596.

[2] Andrei Agrachev and Andrey Sarychev, *Control on the manifolds of mappings as a setting for deep learning*, arXiv preprint arXiv:2008.12702 (2020).

[3] Giacomo Albi, Young-Pil Choi, Massimo Fornasier, and Dante Kalise, *Mean field control hierarchy*, Applied Mathematics & Optimization **76** (2017), no. 1, 93–135.

[4] Luigi Ambrosio, Massimo Fornasier, Marco Morandotti, and Giuseppe Savaré, *Spatially inhomogeneous evolutionary games*, Communications on Pure and Applied Mathematics **74** (2021), no. 7, 1353–1402.

[5] Luigi Ambrosio, Nicola Fusco, and Diego Pallara, *Functions of bounded variation and free discontinuity problems*, Courier Corporation, 2000.

[6] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows in metric spaces and in the space of probability measures*, Second, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2008.

[7] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu, *A convergence analysis of gradient descent for deep linear neural networks*, arXiv preprint arXiv:1810.02281 (2018).

[8] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo, *Implicit regularization in deep matrix factorization*, Advances in neural information processing systems, 2019, pp. 7413–7424.

[9] Benny Avelin and Kaj Nyström, *Neural odes as the deep limit of resnets with constant weights*, Vol. 19, World Scientific, 2021.

[10] Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg, *Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers*, arXiv preprint arXiv:1910.05505 (2019).

[11] Martin Benning, Elena Celledoni, Matthias J Ehrhardt, Brynjulf Owren, and Carola-Bibiane Schönlieb, *Deep learning as optimal control problems: Models and numerical methods*, Journal of Computational Dynamics **6** (2019), 171.

[12] Alain Bensoussan, Jens Frehse, Phillip Yam, et al., *Mean field games and mean field type control theory*, Vol. 101, Springer, 2013.

[13] Julius Berner, Philipp Grohs, and Arnulf Jentzen, *Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of black–scholes partial differential equations*, SIAM Journal on Mathematics of Data Science **2** (2020Jan), no. 3, 631657.

[14] Mattia Bongini, Massimo Fornasier, Francesco Rossi, and Francesco Solombrino, *Mean-field pontryagin maximum principle*, Journal of Optimization Theory and Applications **175** (2017), no. 1, 1–38.

[15] Benoît Bonnet, *A pontryagin maximum principle in wasserstein spaces for constrained optimal control problems*, ESAIM: Control, Optimisation and Calculus of Variations **25** (2019), 52.

[16] Benoît Bonnet and Hélène Frankowska, *Differential inclusions in wasserstein spaces: The cauchy-lipschitz framework*, Journal of Differential Equations **271** (2021), 594–637.

[17] Benoît Bonnet and Hélène Frankowska, *Necessary Optimality Conditions for Optimal Control Problems in Wasserstein Spaces*, To appear in Applied Mathematics and Optimization (2021).

[18] Benoît Bonnet and Hélène Frankowska, *On the Properties of the Value Function Associated to a Mean-Field Optimal Control Problem of Bolza Type*, Submitted (2021).

[19] Benoît Bonnet and Hélène Frankowska, *Semiconcavity and Sensitivity Analysis in Mean-Field Optimal Control and Applications*, In revision (2021).

[20] Benoît Bonnet and Francesco Rossi, *The Pontryagin maximum principle in the Wasserstein space*, Calculus of Variations and Partial Differential Equations **58** (2019), no. 1, 1–36.

[21] Alberto Bressan and Benedetto Piccoli, *Introduction to the mathematical theory of control*, AIMS Series on Applied Mathematics, vol. 2, American Institute of Mathematical Sciences (AIMS), Springfield, MO, 2007.

[22] Martin Burger, René Pinnau, Claudia Totzeck, and Oliver Tse, *Mean-field optimal control and optimality conditions in the space of probability measures*, SIAM Journal on Control and Optimization **59** (2021), no. 2, 977–1006.

[23] José A Canizo, José A Carrillo, and Jesús Rosado, *A well-posedness theory in measures for some kinetic models of collective motion*, Mathematical Models and Methods in Applied Sciences **21** (2011), no. 03, 515–539.

[24] Piermarco Cannarsa and Carlo Sinestrari, *Semiconcave functions, Hamilton-Jacobi equations, and optimal control*, Vol. 58, Springer Science & Business Media, 2004.

[25] René Carmona and François Delarue, *Forwardbackward stochastic differential equations and controlled McKeanVlasov dynamics*, The Annals of Probability **43** (2015), no. 5, 2647 –2700.

[26] Giulia Cavagnari, Stefano Lisini, Carlo Orrieri, and Giuseppe Savaré, *Lagrangian, eulerian and kantorovich formulations of multi-agent optimal control problems: Equivalence and gamma-convergence*, arXiv preprint arXiv:2011.07117 (2020).

[27] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud, *Neural ordinary differential equations*, Proceedings of the 32nd international conference on neural information processing systems, 2018, pp. 65726583.

[28] Alexander Cloninger and Timo Klock, *Relu nets adapt to intrinsic dimensionality beyond the target domain*, arXiv preprint arXiv:2008.02545 (2020).

[29] Ingrid Daubechies, Ronald DeVore, Simon Foucart, Boris Hanin, and Guergana Petrova, *Nonlinear approximation and (deep) relu networks*, 2019.

[30] Ronald DeVore, Boris Hanin, and Guergana Petrova, *Neural network approximation*, arXiv preprint arXiv:2012.14501 (2020).

[31] Weinan E, *A proposal on machine learning via dynamical systems*, Communications in Mathematics and Statistics **5** (2017), no. 1, 1–11.

[32] Weinan E, Jiequn Han, and Qianxiao Li, *A mean-field optimal control formulation of deep learning*, Research in the Mathematical Sciences **6** (2019), no. 1, 10.

[33] Dennis Elbrächter, Philipp Grohs, Arnulf Jentzen, and Christoph Schwab, *Dnn expression rate analysis of high-dimensional pdes: Application to option pricing*, arXiv preprint arXiv:1809.07669 (2020).

[34] Aleksei Fedorovich Filippov, *Differential equations with discontinuous righthand sides: control systems*, Vol. 18, Springer Science and Business Media, 2013.

[35] Massimo Fornasier, Stefano Lisini, Carlo Orrieri, and Giuseppe Savaré, *Mean-field optimal control as Gamma-limit of finite agent controls*, European Journal of Applied Mathematics **30** (2019), no. 6, 1153–1186.

[36] Massimo Fornasier, Benedetto Piccoli, and Francesco Rossi, *Mean-field sparse optimal control*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **372** (2014), no. 2028, 20130400, 21.

[37] Massimo Fornasier and Francesco Solombrino, *Mean-field optimal control*, ESAIM: Control, Optimisation and Calculus of Variations **20** (2014), no. 4, 1123–1152.

[38] Halina Frankowska, *A priori estimates for operational differential inclusions*, Journal of differential equations **84** (1990), no. 1, 100–128.

[39] David Gilbarg and Neil S Trudinger, *Elliptic partial differential equations of second order*, springer, 2015.

[40] Philipp Grohs, Dmytro Perekrestenko, Dennis Elbrächter, and Helmut Bölcskei, *Deep neural network approximation theory*, arXiv preprint arXiv:1901.02220 **1** (2020).

[41] Ingo Gühring, Mones Raslan, and Gitta Kutyniok, *Expressivity of deep neural networks*, 2020.

[42] Eldad Haber and Lars Ruthotto, *Stable architectures for deep neural networks*, Inverse Problems **34** (2017dec), no. 1, 014004.

[43] Awni Hannun, Carl Case, Jared Casper, et al., *Deep speech: Scaling up end-to-end speech recognition*, arXiv preprint arXiv:1412.5567 (2014).

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Deep residual learning for image recognition*, Proceedings of the ieee conference on computer vision and pattern recognition, 2016, pp. 770–778.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Identity mappings in deep residual networks*, Springer, 2016.

[46] Jean-François Jabir, David Šiška, and Łukasz Szpruch, *Mean-field neural odes via relaxed optimal control*, arXiv preprint arXiv:1912.05475 (2019).

[47] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 2012, pp. 1097–1105.

[48] Jean-Michel Lasry and Pierre-Louis Lions, *Mean field games.*, Jpn. J. Math. (3) **2** (2007), no. 1, 229–260.

[49] Yann Lecun, *Une procedure d'apprentissage pour reseau a seuil asymmetrique (a learning scheme for asymmetric threshold networks)*, Proceedings of cognitiva 85, paris, france, 1985, pp. 599–604 (English (US)).

[50] Qianxiao Li, Long Chen, Cheng Tai, and E. Weinan, *Maximum principle based algorithms for deep learning*, J. Mach. Learn. Res. **18** (January 2017), no. 1, 59986026.

[51] Qianxiao Li and Shuji Hao, *An optimal control approach to deep learning and applications to discrete-weight neural networks*, Proceedings of the 35th international conference on machine learning, 201810, pp. 2985–2994.

[52] Guan-Horng Liu and Evangelos A Theodorou, *Deep learning theory review: An optimal control and dynamical systems perspective*, arXiv preprint arXiv:1908.10920 (2019).

[53] Song Mei, Andrea Montanari, and Phan-Minh Nguyen, *A mean field view of the landscape of two-layer neural networks*, Proceedings of the National Academy of Sciences **115** (2018), no. 33, E7665–E7671.

[54] Hrushikesh N Mhaskar and Tomaso Poggio, *Deep vs. shallow networks: An approximation theory perspective*, Analysis and Applications **14** (2016), no. 06, 829–848.

[55] Hrushikesh N Mhaskar and Tomaso Poggio, *Function approximation by deep networks.*, Communications on Pure & Applied Analysis **19** (2020), no. 8.

[56] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis, *Human-level control through deep reinforcement learning*, Nature **518** (2015), no. 7540, 529–533.

[57] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu, *Pixel recurrent neural networks*, 2016, pp. 1747–1756.

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

[59] Philipp Petersen and Felix Voigtlaender, *Optimal approximation of piecewise smooth functions using deep relu neural networks*, Neural Networks **108** (2018), 296 –330.

[60] Benedetto Piccoli, Francesco Rossi, and Magali Tournus, *A wasserstein norm for signed measures, with application to non local transport equation with source term* (2019).

[61] Lev Semenovich Pontryagin, *Mathematical theory of optimal processes*, CRC press, 1987.

[62] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, *Learning internal representations by error propagation*, MIT Press, Cambridge, MA, USA, 1986.

[63] Uri Shaham, Alexander Cloninger, and Ronald R Coifman, *Provable approximation properties for deep neural networks*, Applied and Computational Harmonic Analysis **44** (2018), no. 3, 537–557.

[64] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis, *Mastering the game of go without human knowledge*, Nature **550** (October 2017), 354–.

[65] Ruoyu Sun, *Optimization for deep learning: theory and algorithms*, arXiv preprint arXiv:1912.08957 (2019).

[66] Paulo Tabuada and Bahman Gharesifard, *Universal approximation power of deep residual neural networks via nonlinear control theory*, arXiv preprint arXiv:2007.06007 (2020).

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, Advances in neural information processing systems 30, 2017, pp. 5998–6008.

[68] Paul Werbos, *Beyond regression: New tools for prediction and analysis in the behavioral sciences*, Harvard University, 1975.

[69] Lexing Ying and Emmanuel J Candes, *The phase flow method*, Journal of Computational Physics **220** (2006), no. 1, 184–215.

[70] Eberhard Zeidler, *Applied functional analysis*, Springer Science and Business Media, 1995.

[71] Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, and Inderjit S. Dhillon, *Recovery guarantees for one-hidden-layer neural networks*, Proceedings of the 34th international conference on machine learning, 2017, pp. 4140–4149.