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

Benoît Bonnet, Cristina Cipriani, Massimo Fornasier, Hui Huang

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]CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, 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, yielding a rigorous justification of a phenomenon that we call *coupled descent*, indicating the simultaneous decrease of generalization and training errors. We consider two approaches to the derivation of the mean-field maximum principle, including one that is based on a generalized Lagrange multiplier theorem on convex sets of spaces of measures, which is arguably much simpler than those currently available in the literature for mean-field optimal control problems. 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.bonnet@laas.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 computational approach that performs state-of-the-art on various relevant real-life applications such as speech [46] and image [47, 50] recognition, language translation [71], and which also serves as a basis for novel scientific computing methods [10, 34]. In unsupervised machine learning, deep neural networks have shown great success as well, for instance in image and speech generation [60, 61], and in reinforcement learning for solving control problems, such as mastering Atari games [59] or beating human champions at playing Go [68]. 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

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

where the matrices $W_\ell \in \mathbb{R}^{d_{\ell-1} \times d_\ell}$ represent collections of weights, the vectors $\tau_\ell \in \mathbb{R}^{d_\ell}$ are shifts/biases for each layer $\ell = 1, \dots, L$ and ρ is a scalar activation function acting componentwisely on vectors. 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 – is typically determined by means of heuristic considerations, whereas the weight matrices and the shifts are free parameters which are tuned in various possible ways by using a given training dataset.

Practical evidences towards certified benchmarks confirm that deep-learning algorithms are able to outperform many of the previously existing methods. Also, recent mathematical investigations [10, 27, 29, 31, 34, 43, 57, 58, 62, 66] have proven that deep artificial networks can approximate high dimensional functions without incurring in the curse of dimensionality, i.e. without needing 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 [44], the training phase itself, based on suitable optimization processes, remains a (black-)box with some levels of opacity. In fact, the latter procedure features a surprising and yet mostly unexplained phenomenon, which is in stark contrast with conventional statistics wisdom: in addition to providing a finer empirical data fitting, increasing the number of modelling parameters beyond that of training examples also tends to improve the *generalization error*, namely the prediction error on unseen data. We call the simultaneous decrease of both empirical and generalization errors the *coupled descent* phenomenon. Instead, from classical statistical learning theory [67], one would expect that overfitting should lead to a blow-up of the generalization error, owing to the wealth of complextiy of the underlying model [76]. Hence the prediction of the generalization error from data remains at large a fundamental open problem in deep learning. As one of the main results of this paper, we show that for certain classes of neural networks based on dynamical systems, whose training is reformulated as a convex optimal control problem, the newly defined coupled descent phenomenon can be rigorously explained.

1.2 Training of deep nets and residual blocks

In order to understand the context of our results, let us mention how the neural networks considered in this paper arise. We start by recalling how training of neural networks is performed and how it is facilitated by appropriate network architectures. The method that is most frequently used to train deep neural networks is the so-called *backpropagation of error* [53, 65, 73], which is justified by its tremendous empirical success. Inherently, all the practical advances recalled above are due to the efficacy of this method. The term backpropagation usually refers to the use of stochastic gradient descent[1] or some of its variants [69] to minimize a given loss function (e.g. mean-squared distances, Kullback-Leibler divergences, or Wasserstein distances) over the parameters of the network (the weights and biases), 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

[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 procedure resorting to such gradients. In many cases, these latter are computed using symbolic calculus.

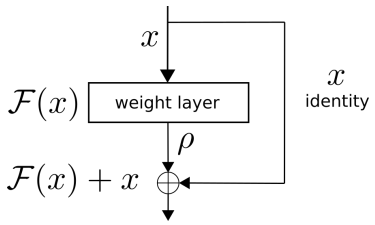

Figure 1: The layer update reads: $X^{n+1} = X^n + \mathcal{F}(X^n)$, see [47].

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 in turn very high-dimensional.

In their groundbreaking work [47], He et al. showed that the training error of the 56-layer CNN network remains worse than that 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 called the "Residual Block", see Figure 1. According to the analysis conveyed in [48], the use of identity mappings as skip connections and after-addition activations of the form

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

turns out to be beneficial to promote the smoothness of the information propagation. Therein, the authors present several 1000-layer deep networks that can be easily trained and achieve improved accuracy. Note that 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 originally the arguments in [48] that support the use of residual blocks are based on empirical considerations, a recent line of research has been devoted to a more mathematical and 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 E Weinan [32] and Haber-Ruthotto [45]. Later on, this dynamical approach has been greatly popularized in the machine learning community under the name of *NeurODE* by Chen et al. [26], see also [56]. The formulation starts by reinterpreting the iteration (1.2) as a step of the discrete-time Euler approximation [7] 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 $\rho := \tanh$ acting componentwisely on its entries. In [32, 33], the authors proposed a *stochastic control formulation* of the training of this nonlinear process, with a detailed analysis of the related optimality conditions. Therein, both the the Hamilton-Jacobi-Bellman equations [23] – based on the well-known dynamic programming principle – and the Pontryagin Maximum Principle [64] were studied in great generality. From another perspective, several recent works [1, 2, 70] in geometric control theory have aimed at explaining the efficiency of NeurODEs 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 measure theoretical reformulation of the general approach developed by E Weinan et al. [33], which allows us to derive more specific properties of the control problem, such as the existence, uniqueness, and smoothness of solutions to the Pontryagin Maximum Principle, and a strong form of 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 $\mu_0 \in \mathcal{P}(\mathbb{R}^{2d})$, 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 $\theta \in L^2([0,T];\mathbb{R}^m)$ 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

$$\inf_{\theta \in L^2([0,T];\mathbb{R}^m)} J(\theta) = \begin{cases} \inf\limits_{\theta \in L^2([0,T];\mathbb{R}^m)} \mathbb{E}_{\mu_0}\Big[\ell(X_T, Y_0)\Big] + \lambda \displaystyle\int_0^T |\theta_t|^2 \,\mathrm{d}t, \\ \text{s.t.} \begin{cases} \dot{X}_t = \mathcal{F}(t, X_t, \theta_t), \\ (X_t, Y_0)_{|t=0} \sim \mu_0. \end{cases} \end{cases} \tag{1.5}$$

The use of a regularization term of the type $\lambda \int_0^T |\theta_t|^2 \,\mathrm{d}t$ is very standard in machine learning, see e.g. [41, Chapter 7] or [51, Section 6]. In the absence of regularization, the resulting trained networks may have huge Lipschitz constants, rendering them extremely unstable and susceptible to adversarial attacks [42]. Additionally, the regularization may significantly help the usual training processes, by making the loss J increasingly more convex. As we shall see more in details below, such a standard regularization will allow us to establish the existence and uniqueness of solutions for (1.5), as well as their continuity with respect to the data, which provides a rigorous explaination to the stability of trained networks and what we name as coupled descent phenomenon. Conversely, we shall also demonstrate numerically in Section 5.2 that the lack of a sufficient regularization causes significant instabilities in the numerical solution of the optimal control problem (1.5), see Figure 7, rendering the latter absolutely essential from a practical standpoint. Other and more general regularizations are of course possible [51], but for the sake of simplicity and clarity in the exposition, we shall 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 [33], but which is not based on stochastic control considerations. We start by providing a measure-theoretic

reformulation of (1.5), which can be interpreted as a generalized optimal transport problem or mean-field optimal control problem. To the best of our knowledge, the present paper is the first in the literature to make such a connection. To this end, let us define a new stochastic process $Z_t := (X_t, Y_t)$ satisfying

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

with initial data (X_0, Y_0) distributed according to μ_0, and 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

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

understood in the sense of distributions as in Definition 2.2 below. With this transport equation at hand, we can recast the stochastic optimal control problem (1.5) as

$$\inf_{\theta \in L^2([0,T];\mathbb{R}^m)} J(\theta) = \begin{cases} \inf\limits_{\theta \in L^2([0,T];\mathbb{R}^m)} \displaystyle\int_{\mathbb{R}^{2d}} \ell(x,y) \mathrm{d}\mu_T(x,y) + \lambda \int_0^T |\theta_t|^2 \,\mathrm{d}t\,, \\ \text{s.t.} \begin{cases} \partial_t \mu_t + \nabla_x \cdot (\mathcal{F}(t,x,\theta_t)\mu_t) = 0\,, \\ \mu_t|_{t=0} = \mu_0\,. \end{cases} \end{cases} \tag{1.8}$$

Therein, the goal is again is to find the control signal θ for which $J(\theta)$ is minimal when μ satisfies the PDE constraint (1.7). Observe that when the initial 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 (1.8) reduces to a classical finite particle optimal control problem with ODE constraints.

Optimal control problems over spaces of probability measures of the form (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 [12, 25, 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 [11] based on the leader-follower model studied in [36]. In this work, the mean-field Pontryagin Maximum Principle is derived as limit of its classical finite-particle counterpart. The first general derivation of the PMP for mean-field optimal control problems was obtained in [18], and is based on a careful adaptation of the strategy of needle-variations to the abstract geometric structure of Wasserstein spaces. These results were further extended in [13] 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 [15], by adapting to the notion of multivalued dynamics in Wasserstein space introduced in [14] a methodology originally developed in [39], which relies on suitable linearisations of set-valued maps that produce admissible inner-perturbed trajectories. From a different standpoint, we also mention [21] in which 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 [24] inspired by the theory of mean-field games [52] (see also [3,9]). Similar methods, based on needle-variations in the space of measures are also leveraged in [33] and [49] for the derivation of the PMP for stochastic control problems of the form (1.5).

1.5 Contributions and organization of the paper

The contributions of this paper can be summarized as follows. From a global standpoint, we start by establishing existence and stability results for (1.8), based on compactness and Γ-convergence arguments. We then proceed by deriving general first-order optimality conditions for the measure-theoretic formulation of the optimal control of NeurODEs. Our modeling assumptions include the typical forward mappings (1.4) that appear throughout the literature related to neural networks, with for instance $\rho := \tanh$. As a matter of fact, most of the results available in the literature do not fully encompass this simple model, as they often require global Lipschitz bounds on the transport velocity field.

Let us now describe with more details the fundamental results of the paper. In Section 3, we start by showing that the mean-field optimal control problem (1.8) has solution when the regularization parameter $\lambda > 0$ is sufficiently large, and that the latter is in fact unique. By leveraging compactness arguments akin to that classically appearing in the theory of Γ-convergence, we also establish non-quantitative stability results for the training problem with respect to finite-samples, both at the level of the cost and of the controls. We then proceed by investigating first-order optimality conditions in Section 4. We initiate the discussion by providing in Section 4.1 a heuristic derivation of the following *mean-field Pontryagin Maximum Principle* ("PMP" in the sequel)

$$\begin{cases} \partial_t \mu_t + \nabla_x \cdot (\mathcal{F}(t,x,\theta_t)\mu_t) = 0, \quad \mu_t|_{t=0} = \mu_0\,, \\ \partial_t \psi + \nabla_x \psi \cdot \mathcal{F}(t,x,\theta_t) = 0, \quad \psi_t|_{t=T} = \ell\,, \\ \theta_t^\top = -\dfrac{1}{2\lambda} \displaystyle\int_{\mathbb{R}^{2d}} \nabla_x \psi \cdot \nabla_\theta \mathcal{F}(t,x,\theta_t) \mathrm{d}\mu_t(x,y)\,, \end{cases} \tag{1.9}$$

which characterizes optimal trajectory-control pairs (μ,θ) for (1.8). In Section 4.2, we show that the above optimality system is well-posed, and prove in Theorem 4.1 that it admits a unique control solution $\theta^* \in \mathrm{Lip}([0,T];\mathbb{R}^m)$. Consequently, we are able to show that the function $\mu_0 \to \theta^*$ which maps initial data distributions to the optimal parameters is single-valued, and to prove that it is also Lipschitz continuous with respect to the Wasserstein distance. Such a precise description of how data are encoded in the parameters of the network is a quite remarkable feature of our results. In particular, it allows us to establish a quantitative *generalization error* for finite samples in Corollary 4.4, which writes

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

In particular, (1.10) provides a rate of convergence that depends exclusively on the approximability of μ_0 by empirical measures μ_0^N. We should stress at this point the relevance of (1.10) as it is one of the few results in the literature that rigorously explains the *coupled descent* of both empirical and generalization error in the training of deep neural networks. In Section 5.2 we present numerical experiments fully confirming this phenomenon which is theoretically expected from (1.10), see Figure 5.

Remark 1.1 (Comparison with the existing literature on generalization errors)**.** *We point out that while the generalization errors established in* [49] *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 – i.e. probability measures over $\mathbb{R}^m$ –, that are forced to be non-deterministic by means of entropic regularization terms (see also* [25]*). On the contrary, the generalization errors that we obtain here relate to deterministic optimal controls with values in $\mathbb{R}^m$. A similar bound, yielding* (1.10)*, also appears in a completely different context in* [21, 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 4.3. At this stage, let it be noted that while part of our results may be derived by due adaptations from other approaches developed, e.g., in [21, 33] or [13, 15, 18], we are able to obtain a few stronger properties on the solutions of the optimal control problem than those generally presented in the literature. Whereas in [13, 15, 18] the first order optimality conditions are established in greater generality – but also with significant technical effort –, we propose in this paper a new and 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 4.1 –, we prove in Theorem 4.6 that the mean-field PMP (1.9) can be obtained by means of a generalized Lagrange Multiplier Theorem on the convex subset of Radon measures with unit mass. To this end, we use a new form of calculus recently introduced in [4], which is simpler than the calculus in Wasserstein spaces used in [21]. 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. This novel interpretation of the mean-field PMP as result of a Lagrange Multiplier Theorem in spaces of measures is in our view quite powerful, because it can be applied in other mean-field optimal control problems and be more easily understood by a broader community in optimization.

The main theoretical results of the paper can then be summarized as follows.

Theorem 1.1 (Main contributions of the article)**.** *Let $T > 0$ be given, consider a map $\mathcal{F}$ satisfying Assumptions 1 and 2 of Section 3, fix an initial data distribution $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$, and suppose that the regularization parameter $\lambda > 0$ is sufficiently large.*

Then, the mean-field optimal control problem (1.8) *admits solutions, and 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,*

the optimal control θ^* is uniquely determined, Lipschitz continuous in time, and depends continuously on the initial data distribution μ_0.

We then close the article by presenting numerical experiments to test the novel mean-field Pontryagin maximum principle that we propose, in which 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 derive a general semiconvexity estimate for the reduced cost functional, and provide sufficient conditions ensuring the existence and stability of its minimizers. In Section 4 we investigate the mean-field maximum principle by first studying its well-posedness and deriving the generalization error estimate (1.10), and then showing rigorously how it can be derived either by using a Lagrange multiplier theorem, or via a reduction of the Hamiltonian form. We finally present instructive numerical experiments in Section 5, where solutions of the mean-field maximum principle are computed by means of a shooting method. The Appendix contains proofs of auxiliary results, including the proof of a generalized Lagrange multiplier theorem, Theorem 4.5, 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 $\mathcal{M}(\mathbb{R}^d) \simeq (\mathcal{C}_0(\mathbb{R}^d))'$ 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 $(Z,\widehat{Z})$ whose laws are given by μ and ν respectively.

It is a well-known result in optimal transport theory that when $p=1$ and $\mu,\nu \in \mathcal{P}_c(\mathbb{R}^d)$, the following alternative representation holds for the Wasserstein distance

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

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 φ. In the sequel, we shall also use the signed generalized Wasserstein distance $\mathbb{W}_1^{1,1}$ introduced in [63], which coincides with the bounded Lipschitz distance. Given $\mu,\nu \in \mathcal{M}(\mathbb{R}^d)$, we set

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

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 $\psi \in \mathcal{C}^1_b(\mathbb{R}^{2d})$ *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* $\eta \in \mathcal{C}^1_c((0,T))$ *and* $\xi \in \mathcal{C}^1_c(\mathbb{R}^{2d})$ *is dense in* $\mathcal{C}^1_c((0,T)\times\mathbb{R}^{2d})$ *(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* $\mathcal{C}^1_b(\mathbb{R}^{2d})$ *instead of* $\mathcal{C}^1_c(\mathbb{R}^{2d})$ *or* $\mathcal{C}^1_0(\mathbb{R}^{2d})$ *in* (2.11).

Classical well-posedness results for (1.7) for arbitrary initial measures are usually established under the following type of standard Cauchy-Lipschitz assumptions (or minimal variations thereof).

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

(*i*) *For any fixed* $\theta \in \mathbb{R}^m$, *the map* $(t,x) \mapsto \mathcal{F}(t,x,\theta) \in \mathbb{R}^d$ *is continuous.*

(*ii*) *There exists a constant* $C_{\mathcal{F}} > 0$ *that may depend on* d, m *such that for every* $\theta \in \mathbb{R}^m$, *it holds*

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

(*iii*) *There exists a constant* $L_{\mathcal{F}} > 0$ *independent of* d, m *such that for every* $\theta \in \mathbb{R}^m$, *it holds*

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

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

(*iv*) *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* $C(d,m,R) > 0$ *such that*

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

Under the set of assumptions listed above, we can prove the well-posedness of (1.7) as stated in the following theorem. The proof of the latter is standard and deferred to Appendix A.

Theorem 2.3 (Classical well-posedness for continuity equation)**.** *Consider a measure* $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$ *with* $\mathrm{supp}(\mu_0) \subset B(R)$ *for some* $R > 0$, *and suppose that* $\mathcal{F}$ *satisfies 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* $C_{\mathcal{F}}$ *such that*

$$\mathrm{supp}(\mu_t) \subset B(R_T), \tag{2.12}$$

for all times $t \in [0,T]$, *and additionally it holds for any* $s,t \in [0,T]$ *that*

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

Denoting by μ^i *for* $i = 1,2$ *two solutions of* (1.7) *with initial data* μ^i_0 *satisfying the above assumptions, the following stability estimate*

$$W_1(\mu^1_t,\mu^2_t) \le e^{L_{\mathcal{F},T,\|\theta\|_1}} W_1(\mu^1_0,\mu^2_0), \tag{2.14}$$

holds for all times $t \in [0,T]$, *where* $C_{\mathcal{F}}$ *and* $L_{\mathcal{F},T,\|\theta\|_1}$ *are defined as in Assumption 1.*

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 F-differentiable 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

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

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-differentiable at some $e \in E$, then it is G-differentiable 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 Existence of minimizers and stability of solutions

In this section, we investigate sufficient conditions ensuring the existence of optimal solutions to the mean-field optimal control problem (1.8), as well as stability properties for the minimizers and costs stemming from large finite-sample training. Throughout the remainder of this article, we will use Assumption 1 and the following additional hypotheses to establish most of our results.

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

(i) 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

$$|\nabla_x \cdot \nabla_\theta \mathcal{F}(t, x, \theta)| + |\nabla_x \mathcal{F}(t, x, \theta)| + |\nabla_x^2 \mathcal{F}(t, x, \theta)| \leq C(d, m, R, |\theta|)\,. \tag{3.1}$$

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

$$|\mathcal{F}(t, x, \theta^1) - \mathcal{F}(s, x, \theta^2)| \leq C(d, m, R)\big(|t - s| + |\theta^1 - \theta^2|\big)\,. \tag{3.2}$$

(iii) For all fixed θ and $t \in [0, T]$, it holds

$$|\nabla_\theta \mathcal{F}(t, x, \theta) - \nabla_\theta \mathcal{F}(t, y, \theta)| \leq C(d, m, R, |\theta|)|x - y|, \tag{3.3}$$

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

Before moving on to the discussion pertaining to the existence and stability properties for solutions of (1.8), we highlight the adequacy of our working hypotheses in connection with classical machine learning models.

Remark 3.1 (Adequacy for smooth sigmoidal activations)**.** *Assumptions 1 and 2 require smooth activation functions that exhibit also some boundedness properties with respect to the parameter θ, e.g. as in Assumption 1-(ii). These latter are needed both to express the PMP and to establish its well-posedness, as will become apparent in Section 4. Hence, some popular network models which use for instance ReLu activations are not covered by our results. However, we check here that the sets of hypotheses listed in Assumptions 1 and 2 include the popular subclass of feed-forwarding dynamics* (1.4) *involving sigmoidal-type activation functions, such as*

$$\mathcal{F}(t, x, \theta) = \mathcal{F}(x, \theta) := \tanh(\theta x) \in \mathbb{R}^d\,,$$

where $\theta \in \mathbb{R}^m = \mathbb{R}^{d \times d}$ and $x \in \mathbb{R}^d$. In that case, Assumption 1-(i) obviously holds, and since

$$\mathcal{F}_k(x, \theta) = \tanh\Big(\sum_{l=1}^{d} \theta_{k,l} x_l\Big)$$

for each $k \in \{1, \dots, d\}$ and $|\tanh(r)| \leq 1$ for all $r \in \mathbb{R}$, we have that $|\mathcal{F}(x, \theta)| \leq \sqrt{d}$ for all $(x, \theta) \in \mathbb{R}^d \times \mathbb{R}^{d \times d}$, and Assumption 1-(ii) also holds. This uniform boundedness property of the driving field implies in particular that the radius $R_T > 0$ given by Theorem 2.3 controlling the

support sizes of the solutions of (1.7) *will scale polynomially and not exponentially on* $d \geq 1$*, along with all the relevant constants depending polynomially thereon. Moreover, observe that*

$$\partial_{x_i} \mathcal{F}_k(x,\theta) = \tanh'\Big(\sum_{l=1}^{d} \theta_{k,l}\, x_l\Big)\theta_{k,i}$$

for each $i,k \in \{1,\dots,d\}$*, which implies in particular that* $|\nabla_x \mathcal{F}(x,\theta)| \leq |\theta|$ *for all* $(x,\theta) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}$ *by using the fact that* $|\tanh'(r)| = |1-\tanh(r)^2| \leq 1$ *for each* $r \in \mathbb{R}$*. By the mean-value theorem, this latter fact directly implies that*

$$|\mathcal{F}(t,x_1,\theta) - \mathcal{F}(t,x_2,\theta)| \leq |\theta||x_1 - x_2|$$

for all $\theta \in \mathbb{R}^{d\times d}$ *and* $x_1, x_2 \in \mathbb{R}^d$*, which verifies Assumption 1-*(iii)*. Concerning Assumption 1-*(iv)*, one has that*

$$\partial_{\theta_{ij}} \mathcal{F}_k(x,\theta) = \delta_{k,i} \tanh'\Big(\sum_{l=1}^{d} \theta_{k,l} x_l\Big) x_j$$

for each $i,j,k \in \{,1\dots,d\}$*, where* $\delta_{k,i}$ *refers here to the Kronecker symbol, which implies that* $|\nabla_\theta \mathcal{F}(x,\theta)| \leq \sqrt{d}|x|$ *for all* $\theta \in \mathbb{R}^m$ *and* $x \in \mathbb{R}^d$*. Furthermore, one can easily see that*

$$\partial^2_{\theta_{i,j},\theta_{m,n}} \mathcal{F}_k(x,\theta) = \delta_{k,m}\delta_{k,i} \tanh''\Big(\sum_{l=1}^{d} \theta_{k,l} x_l\Big) x_j x_n$$

for each $i,j,k,m,n \in \{1,\dots,d\}$*, which then yields* $|\nabla^2_\theta \mathcal{F}(x,\theta)| \leq 4\sqrt{d}\,|x|^2$ *for all* $(x,\theta) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}$ *since* $|\tanh''(r)| = |2\tanh(r)\,(\tanh(r)-1)\,| \leq 4$ *for every* $r \in \mathbb{R}^d$*. Thence, it holds*

$$\max_{(x,\theta)\in B(R)\times\mathbb{R}^m} |\nabla_\theta \mathcal{F}(x,\theta)| \leq \sqrt{d}R \qquad \text{and} \qquad \max_{(x,\theta)\in B(R)\times\mathbb{R}^m} |\nabla^2_\theta \mathcal{F}(x,\theta)| \leq 4\sqrt{d}R^2, \tag{3.4}$$

which completes the verification of Assumption 1.

We now shift our attention to the verification of Assumption 2. First of all, one has that

$$\partial^2_{x_i,x_j} \mathcal{F}_k(x,\theta) = \tanh''\Big(\textstyle\sum_{l=1}^{d} \theta_{k,l} x_l\Big)\theta_{k,i}\theta_{k,j}$$

for each $i,j,k \in \{1,\dots,d\}$*, which yields the estimate* $|\nabla^2_x \mathcal{F}(x,\theta)| \leq 4|\theta|^2$ *for all* $(x,\theta) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}$*. Moreover, one can check that*

$$\partial_{x_n}\partial_{\theta_{i,j}} \mathcal{F}_k(x,\theta) = \delta_{k,i} \tanh''\Big(\sum_{l=1}^{d} \theta_{k,l} x_l\Big) x_j \theta_{k,n} + \delta_{k,i}\delta_{j,n} \tanh'\Big(\sum_{l=1}^{d} \theta_{k,l} x_l\Big)$$

for each $i,j,k,n \in \{1,\dots,d\}$*. Thus, we obtain the estimates*

$$|\nabla_x \cdot \nabla_\theta \mathcal{F}(x,\theta)| \leq \sqrt{d}|\nabla_x \nabla_\theta \mathcal{F}(x,\theta)| \leq \sqrt{2}\sqrt{d}\big(4|x||\theta| + d\big)$$

for all $(x,\theta) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}$*, which leads to Assumption 2-*(i) *being fulfilled. Moreover, we can also deduce from the previous estimate that Assumption 2-*(iii) *holds, since*

$$|\nabla_\theta \mathcal{F}(t,x,\theta) - \nabla_\theta \mathcal{F}(t,y,\theta)| \leq \sqrt{2}(4R|\theta| + d)|x - y|$$

for all $\theta \in \mathbb{R}^{d\times d}$ *and* $x, y \in B(R)$*. Lastly, it follows from* (3.4) *that*

$$|\mathcal{F}(t,x,\theta^1) - \mathcal{F}(s,x,\theta^2)| \leq \sqrt{d}R|\theta^1 - \theta^2|$$

for all $\theta^1, \theta^2 \in \mathbb{R}^{d\times d}$ *and* $x \in \mathbb{R}^d$*, which equivalently means that Assumption 2-*(ii) *is satisfied and completes the verification of Assumption 2.*

3.1 Convexity of the reduced cost functional and existence of minimizers

As already recalled in the introduction, L^2-regularization of network parameters is a standard practice in machine learning which helps stabilizing the training procedure, while promoting the generalization capacities of networks [41, 51]. In this section, we show that for regularization parameters $\lambda > 0$ that are sufficiently large, the reduced cost of the problem is actually strictly convex, which in particular implies the existence and uniqueness of an optimal control $\theta^* \in L^2([0,T];\mathbb{R}^m)$ for the mean-field optimal control problem (1.8). Given the smoothness of the forward map $\mathcal{F}$, the convexity of J is perhaps not surprising, but it has never been noticed before in the literature in connection to mean-field optimal control problems, and appears to have far-reaching practical implications that we shall explore in the remainder of the paper.

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

$$\begin{cases} \partial_t \Phi^\theta_{(\tau,t)}(x) = \mathcal{F}\big(t, \Phi^\theta_{(\tau,t)}(x), \theta_t\big), \\ \Phi^\theta_{(\tau,\tau)}(x) = x, \end{cases} \tag{3.5}$$

for every $x \in \mathbb{R}^d$. It is a well-known result in the theory of non-linear dynamical systems (see e.g. [19, Theorem 2.3.2]) that under Assumption 1, 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 linearized Cauchy problem

$$\begin{cases} \partial_t w(t,x) = \nabla_x \mathcal{F}\big(t, \Phi^\theta_{(\tau,t)}(x), \theta_t\big) w(t,x) \\ w(\tau,x) = \mathrm{Id}. \end{cases} \tag{3.6}$$

This allows us to establish the following semiconvexity result for the reduced cost of (1.8).

Proposition 3.1 (Semiconvexity of the reduced cost functional)**.** *Let $T, R > 0$ and $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$ be such that $\mathrm{supp}(\mu_0) \subset B(R)$, and suppose that Assumptions 1 and 2 hold. Then, for every ball $\Gamma \subset L^2([0,T];\mathbb{R}^m)$, there exists a constant $\mathcal{L}(T,R,\Gamma) > 0$ such that the reduced cost functional*

$$J : \theta \in L^2([0,T];\mathbb{R}^m) \mapsto \begin{cases} \displaystyle\int_{\mathbb{R}^{2d}} \ell(x,y)\mathrm{d}\mu^\theta_T(x,y) + \lambda \int_0^T |\theta_t|^2 \mathrm{d}t, \\ \text{s.t.} \begin{cases} \partial_t \mu^\theta_t + \nabla_x\big(\mathcal{F}(t,x,\theta_t)\mu^\theta_t\big) = 0, \\ \mu^\theta_0 = \mu_0, \end{cases} \end{cases} \tag{3.7}$$

satisfies the semiconvexity estimate

$$J\big((1-\zeta)\theta^1 + \zeta\theta^2\big) \le (1-\zeta)J(\theta^1) + \zeta J(\theta^2) - (2\lambda - \mathcal{L}(T,R,\Gamma))\tfrac{\zeta(1-\zeta)}{2}\|\theta^1-\theta^2\|_2^2 \tag{3.8}$$

for any $\theta^1, \theta^2 \in \Gamma$ and all $\zeta \in [0,1]$. In particular if $\lambda > \frac{1}{2}\mathcal{L}(T,R,\Gamma)$, the reduced cost functional is then strictly convex over Γ.

The proof of this convexity estimate is almost entirely contained in the following regularity result, which itself relies on a series of technical properties for characteristic flows which are exposed in Appendix B.

Lemma 3.1 (Regularity of the reduced final cost)**.** *Let $T, R > 0$ and $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$ be such that* $\mathrm{supp}(\mu_0) \subset B(R)$*, and suppose that Assumptions 1 and 2 hold. Then, the reduced final cost*

$$J_\ell : \theta \in L^2([0,T];\mathbb{R}^m) \mapsto \begin{cases} \displaystyle\int_{\mathbb{R}^{2d}} \ell(x,y)\mathrm{d}\mu_T^\theta(x,y), \\ \text{s.t.} \begin{cases} \partial_t \mu_t^\theta + \nabla_x\big(\mathcal{F}(t,x,\theta_t)\mu_t^\theta\big) = 0, \\ \mu_0^\theta = \mu_0, \end{cases} \end{cases} \tag{3.9}$$

is Fréchet-differentiable. Moreover, denoting its gradient by $\nabla_\theta J_\ell(\theta) \in L^2([0,T];\mathbb{R}^m)$ and choosing $\theta^1, \theta^2 \in L^2([0,T];\mathbb{R}^m)$, there exists a constant $\mathcal{L}(T,R,\|\theta^1\|_1,\|\theta^2\|_1) > 0$ such that

$$\big\|\nabla_\theta J_\ell(\theta^1) - \nabla_\theta J_\ell(\theta^2)\big\|_2 \leq \mathcal{L}(T,R,\|\theta^1\|_1,\|\theta^2\|_1)\,\big\|\theta^1 - \theta^2\big\|_2.$$

Proof. We start by fixing a control signal $\theta \in L^2([0,T];\mathbb{R}^m)$. Following the discussion in Appendix A below, the unique solution $\mu^\theta \in \mathcal{C}([0,T];\mathcal{P}_c(\mathbb{R}^{2d}))$ of the controlled continuity equation can be expressed as $\mu_t^\theta = \mathbf{\Phi}_{(0,t)}^\theta \sharp \mu_0$, where

$$\mathbf{\Phi}_{(0,t)}^\theta(x,y) = \big(\Phi_{(0,t)}^\theta(x), y\big)$$

for all $(x,y) \in \mathbb{R}^{2d}$, with $(\Phi_{(0,t)}^\theta(\cdot))_{t\in[0,T]}$ being the characteristic flow defined in (3.5). In particular, this allows us to rewrite the reduced final cost as

$$J_\ell(\theta) = \int_{\mathbb{R}^{2d}} \ell\Big(\Phi_{(0,T)}^\theta(x), y\Big)\mathrm{d}\mu_0(x,y).$$

Given another control signal $\vartheta \in L^2([0,T];\mathbb{R}^m)$ and some $\varepsilon > 0$, we know by Proposition B.2 that the following Taylor expansion

$$\Phi_{(0,T)}^{\theta+\varepsilon\vartheta}(x) = \Phi_{(0,T)}^\theta(x) + \varepsilon \int_0^T \mathcal{R}_{(t,T)}^\theta(x)\nabla_\theta \mathcal{F}\big(t, \Phi_{(0,t)}^\theta(x), \theta_t\big)\vartheta_t \mathrm{d}t + o_\theta(\varepsilon) \tag{3.10}$$

holds for all $(t,x) \in [0,T]\times B(R)$, where $(\mathcal{R}_{(\tau,t)}^\theta(\cdot))_{t\in[0,T]} \subset \mathcal{C}^1(\mathbb{R}^d;\mathbb{R}^{d\times d})$ are the resolvent maps of the linearized Cauchy problem defined as in (B.2). Since the small-o in (3.10) is uniform in $x \in B(R)$, it holds by Lebesgue's dominated convergence and Fubini's theorems that

$$\begin{aligned} &\int_{\mathbb{R}^{2d}} \ell\Big(\Phi_{(0,T)}^{\theta+\varepsilon\vartheta}(x), y\Big)\mathrm{d}\mu_0(x,y) \\ &= \int_{\mathbb{R}^{2d}} \ell\Big(\Phi_{(0,T)}^{\theta}(x), y\Big)\mathrm{d}\mu_0(x,y) \\ &\quad + \varepsilon \int_0^T \Big\langle \int_{\mathbb{R}^{2d}} \Big(\mathcal{R}_{(t,T)}^\theta(x)\nabla_\theta \mathcal{F}\big(t, \Phi_{(0,t)}^\theta(x), \theta_t\big)\Big)^\top \nabla_x \ell\big(\Phi_{(0,T)}^\theta(x), y\big)\mathrm{d}\mu_0(x,y), \vartheta_t \Big\rangle \mathrm{d}t + o_\theta(\varepsilon), \end{aligned} \tag{3.11}$$

for every $\varepsilon > 0$ small enough. From the regularity estimates of Assumption 1, Proposition B.1 and Proposition B.2, we may infer that the Gateaux derivative expressed in (3.11) is continuous with respect to $\theta \in L^2([0,T];\mathbb{R}^m)$, so that the reduced final cost is Fréchet-differentiable, with

$$\nabla_\theta J_\ell(\theta) : t \in [0,T] \mapsto \int_{\mathbb{R}^{2d}} \Big(\mathcal{R}_{(t,T)}^\theta(x)\nabla_\theta \mathcal{F}\big(t, \Phi_{(0,t)}^\theta(x), \theta_t\big)\Big)^\top \nabla_x \ell\big(\Phi_{(0,T)}^\theta(x), y\big)\mathrm{d}\mu_0(x,y). \tag{3.12}$$

At this stage, by resorting again to Assumptions 1 and 2, Proposition B.1 and Proposition B.2, one can check that the previous expression is a (formal) product of quantities which are

bounded and Lipschitz with respect to θ on bounded subsets of $L^1([0,T];\mathbb{R}^m)$. Whence, for every pair $\theta^1, \theta^2 \in L^2([0,T];\mathbb{R}^m)$, there exists a constant $\mathcal{L}(T, R, \|\theta^1\|_1, \|\theta^2\|_1)$ such that

$$\left\|\nabla_\theta J_\ell(\theta^1) - \nabla_\theta J_\ell(\theta^2)\right\|_2 \leq \mathcal{L}(T, R, \|\theta^1\|_1, \|\theta^2\|_1) \left\|\theta^1 - \theta^2\right\|_2,$$

which ends the proof of our claim. □

We are now ready to move on to the proof of Proposition 3.1.

Proof of Proposition 3.1. First, observe that the reduced cost of the problem can be written as

$$J(\theta) = J_\ell(\theta) + \lambda\|\theta\|_2^2$$

for all $\theta \in L^2([0,T];\mathbb{R}^m)$, where $J_\ell(\theta)$ stands for the reduced final cost defined in (3.9). Whence, it can be easily checked as a consequence of Lemma 3.1 that the reduced cost is Fréchet-differentiable, with

$$\nabla_\theta J(\theta) = \nabla_\theta J_\ell(\theta) + 2\lambda\theta. \tag{3.13}$$

Let $\Gamma \subset L^2([0,T];\mathbb{R}^m)$ be a closed ball and $\theta^1, \theta^2 \in \Gamma$. By performing routine computations based on the integral version of Taylor's theorem (see e.g. [12, Lemma 6] for a detailed proof in the finite-dimensional case), one can show that

$$\begin{aligned} J_\ell\big((1-\zeta)\theta^1 + \zeta\theta^2\big) &\leq (1-\zeta)J_\ell(\theta^1) + \zeta J_\ell(\theta^2) + \mathrm{Lip}(\nabla_\theta J_\ell\,;\Gamma)\tfrac{\zeta(1-\zeta)}{2}\left\|\theta^1 - \theta^2\right\|_2^2 \\ &\leq (1-\zeta)J_\ell(\theta^1) + \zeta J_\ell(\theta^2) + \mathcal{L}(T, R, \Gamma)\tfrac{\zeta(1-\zeta)}{2}\left\|\theta^1 - \theta^2\right\|_2^2, \end{aligned}$$

for all $\zeta \in [0,1]$, where the constant $\mathcal{L}(T, R, \Gamma) := \mathcal{L}(T, R, \|\theta^1\|_1, \|\theta^2\|_1)$ is given as in Lemma 3.1. This, together with the standard fact of convex analysis in Hilbert spaces stating that

$$\left\|(1-\zeta)\theta^1 + \zeta\theta^2\right\|_2^2 \leq (1-\zeta)\|\theta^1\|_2^2 + \zeta\|\theta^2\|_2^2 - \tfrac{\zeta(1-\zeta)}{2}\left\|\theta^1 - \theta^2\right\|_2^2$$

allows us to conclude that the reduced cost functional satisfies the semiconvexity estimate (3.8) over Γ. □

By leveraging the semiconvexity result of Proposition 3.1, we are able to derive sufficient conditions for the existence of mean-field optimal controls.

Theorem 3.2 (Existence of minimizers)**.** *Let $T, R > 0$, $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$ be such that $\mathrm{supp}(\mu_0) \subset B(R)$, and $\Gamma \subset L^2([0,T];\mathbb{R}^m)$ be the closed ball of radius $C_\Gamma^{1/2} := \|\ell\|_{\mathcal{C}(B(R))} + 1$. If the regularization parameter is such that $\lambda > \frac{1}{2}\mathcal{L}(T, R, \Gamma)$ where the latter constant is given as in Lemma 3.1, then there exists a unique optimal control $\theta^* \in \Gamma$ for (1.8).*

Proof. The result follows from a standard application of the direct method of the calculus of variations. Given a minimizing sequence $(\theta^n) \subset L^2([0,T];\mathbb{R}^m)$ for which

$$J(\theta^n) \underset{n\to+\infty}{\longrightarrow} \inf_{\theta \in L^2([0,T];\mathbb{R}^m)} J(\theta), \tag{3.14}$$

it necessarily holds for $n \geq 1$ sufficiently large that

$$J(\theta^n) \leq J(0) + 1 \leq \|\ell\|_{\mathcal{C}(B(R))} + 1.$$

Recalling the expression (3.7) of the reduced cost, this implies in particular that $\|\theta^n\|_2 \leq C_\Gamma^{1/2}$ for each $n \geq 1$, or equivalently $(\theta^n) \subset \Gamma$. Remark now that $\Gamma \subset L^2([0,T];\mathbb{R}^m)$ is weakly compact since it is a closed ball in a Hilbert space (see e.g. [20, Theorem 3.17]), so that there exists an element $\theta^* \in \Gamma$ for which

$$\theta^{n_k} \underset{k\to+\infty}{\rightharpoonup} \theta^* \qquad \text{in } L^2([0,T];\mathbb{R}^m),$$

along an adequate subsequence. Moreover, it easily follows from Lemma 3.1 that $\theta \mapsto J(\theta) \in \mathbb{R}$ is continuous in the strong L^2-topology, as well as convex since we assumed that $\lambda > \frac{1}{2}\mathcal{L}(T,R,\Gamma)$. As such, it is weakly lower-semicontinuous (see e.g. [20, Corollary 3.9]), which together with (3.14) implies that

$$J(\theta^*) \leq \liminf_{n\to+\infty} J(\theta^n) = \inf_{\theta\in L^2([0,T];\mathbb{R}^m)} J(\theta).$$

Hence, we have shown that $\theta^* \in \Gamma$ is a solution of the mean-field optimal control problem (1.8), and its uniqueness follows straightforwardly from the strict convexity of the reduced cost. □

3.2 Stability of finitely-sampled costs and controls

In this section, we establish a general stability property for solutions of the mean-field optimal control problem (1.8) with respect to finite-samples. More precisely, assume that we are given a sample $\{(X_0^i, Y_0^i)\}_{i=1}^N$ of size $N \geq 1$ independently and identically distributed according to $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$, and let us consider the empirical loss minimization problem

$$\inf_{\theta\in L^2([0,T];\mathbb{R}^m)} J^N(\theta) := \begin{cases} \inf\limits_{\theta\in L^2([0,T];\mathbb{R}^m)} \dfrac{1}{N}\displaystyle\sum_{i=1}^N \ell(X_T^i, Y_T^i) + \lambda \int_0^T |\theta_t|^2 \mathrm{d}t \\ \text{s.t.} \begin{cases} \dot{X}_t^i = \mathcal{F}(t, X_t^i, \theta_t), & \dot{Y}_t^i = 0, \\ (X_t^i, Y_t^i)_{|t=0} = (X_0^i, Y_0^i), & i \in \{1,\dots,N\}. \end{cases} \end{cases} \tag{3.15}$$

By introducing the empirical measure $\mu_0^N \in \mathcal{P}_c^N(\mathbb{R}^{2d})$, defined by

$$\mu_0^N := \frac{1}{N}\sum_{i=1}^N \delta_{(X_0^i, Y_0^i)}, \tag{3.16}$$

the latter can be rewritten as the mean-field optimal control problem (1.8) with initial datum μ_0^N. In the following theorem, we show that when the regularization parameter $\lambda > 0$ is sufficiently large and the empirical samples satisfy

$$W_1(\mu_0^N, \mu_0) \underset{N\to+\infty}{\longrightarrow} 0, \tag{3.17}$$

then the minimizers and optimal values of the problems (3.15) converge in a suitable sense towards those of (1.8). Even though we do not resort explicitly to this terminology in the sequel, this stability result amounts to showing that the sequence (J^N) is Γ-converging towards J for the weak topology of $L^2([0,T];\mathbb{R}^m)$ in the sense e.g. of [28]. Although it bears some interest and provides insights on the finite data consistency of the problem, the result that follows is non-quantitative and purely based on compactness arguments. In order to obtain a quantitative version of this stability property, it is necessary to establish a smooth relation between optimal controls the θ^* and the data distributions μ_0. Such a connection will be realized through the fundamental formula (4.8) below, by leveraging the mean-field PMP studied in Section 4.

Theorem 3.3 (Stability of finitely sampled costs and controls)**.** *Let $T, R > 0$ be given, $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$ be such that $\mathrm{supp}(\mu_0) \subset B(R)$, and assume that Assumptions 1 and 2 hold. Moreover, suppose that $\lambda > 0$ is sufficiently large in the sense of Theorem 3.2.*

Then for every empirical approximating sequence (μ_0^N) satisfying (3.16)-(3.17)*, the corresponding sequence of optimal controls $(\theta^N) \subset L^2([0,T];\mathbb{R}^m)$ is such that*

$$\theta^N \underset{N\to+\infty}{\rightharpoonup} \theta^* \qquad \text{in } L^2([0,T];\mathbb{R}^m), \tag{3.18}$$

where $\theta^* \in L^2([0,T];\mathbb{R}^m)$ is the unique solution of (1.8)*. Moreover, the optimal values converge as well, in the sense that*

$$J^N(\theta^N) \underset{N\to+\infty}{\longrightarrow} J(\theta^*) = \min_{\theta\in L^2([0,T];\mathbb{R}^m)} J(\theta). \tag{3.19}$$

Before proving Theorem 3.3, we state a useful auxiliary lemma exhibiting the dependence of the reduced empirical cost with respect to the sample size $N \geq 1$.

Lemma 3.2 (Dependence of the reduced cost with respect to N)**.** *For every $\theta \in L^2([0,T];\mathbb{R}^m)$, there exists a constant $C(T,R,\|\theta\|_1) > 0$ such that*

$$\big|J(\theta) - J^N(\theta)\big| \leq C(T,R,\|\theta\|_1) W_1(\mu_0^N,\mu_0) \tag{3.20}$$

and

$$\big\|\nabla J(\theta) - \nabla J^N(\theta)\big\|_2 \leq C(T,R,\|\theta\|_1) W_1(\mu_0^N,\mu_0) \tag{3.21}$$

for each $N \geq 1$.

Proof. Let us denote by $\mu, \mu^N \in \mathcal{C}([0,T];\mathcal{P}_c(\mathbb{R}^{2d}))$ the solutions of (1.7) with control θ and initial data $\mu_0^N, \mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$ respectively. Under Assumptions 1, it follows from Theorem 2.3 that

$$\sup_{t\in[0,T]} W_1(\mu_t^N,\mu_t) \leq e^{L_{\mathcal{F},T,\|\theta\|_1}} W_1(\mu_0^N,\mu_0)$$

for some $L_{\mathcal{F},T,\|\theta\|_1} > 0$. This combined with Kantorovich's duality formula (2.6) implies that

$$\big|J(\theta) - J^N(\theta)\big| = \left|\int_{\mathbb{R}^{2d}} \ell(x,y)\mathrm{d}\big(\mu_T - \mu_T^N\big)(x,y)\right| \leq \mathrm{Lip}(\ell\,;B(R))\, e^{L_{\mathcal{F},T,\|\theta\|_1}} W_1(\mu_0^N,\mu_0),$$

for each $N \geq 1$. Analogously by leveraging the analytical expression (3.12) of the gradient of the reduced final cost, one also has that

$$\begin{aligned}&\big\|\nabla J(\theta) - \nabla J^N(\theta)\big\|_2^2 \\ &\qquad \leq \int_0^T \left|\int_{\mathbb{R}^{2d}} \Big(\mathcal{R}^\theta_{(t,T)}(x)\nabla_\theta \mathcal{F}\big(t,\Phi^\theta_{(0,t)}(x),\theta_t\big)\Big)^\top \nabla_x \ell\big(\Phi^\theta_{(0,T)}(x),y\big)\mathrm{d}\big(\mu_0 - \mu_0^N\big)(x,y)\right|^2 \mathrm{d}t.\end{aligned} \tag{3.22}$$

At this stage, one can check that as a consequence of Assumptions 1 and 2 along with the definition (B.2) of the resolvent maps that there exists a constant $C'(T,R,\|\theta\|_1) > 0$ such that

$$\int_0^T \Big\|\Big(\mathcal{R}^\theta_{(t,T)}(\cdot)\nabla_\theta \mathcal{F}\big(t,\Phi^\theta_{(0,t)}(\cdot),\theta_t\big)\Big)^\top \nabla_x \ell\big(\Phi^\theta_{(0,T)}(\cdot),\cdot\big)\Big\|^2_{\mathcal{C}^1(B(R))} \mathrm{d}t \leq C'(T,R,\|\theta\|_1)^2. \tag{3.23}$$

By combining (3.22) and (3.23) with an application of Kantorovich's duality formula (2.6), we finally obtain that

$$\big\|\nabla J(\theta) - \nabla J^N(\theta)\big\|_2 \leq C'(T, R, \|\theta\|_1) W_1(\mu_0^N, \mu_0)$$

for each $N \geq 1$, which concludes the proof of Lemma 3.2 by simply setting $C(T, R, \|\theta\|_1) := \max\big\{\mathrm{Lip}(\ell\,; B(R))\, e^{L_{\mathcal{F},T,\|\theta\|_1}}\,,\, C'(T, R, \|\theta\|_1)\big\}$. □

Building on these a priori estimates, we can move on to the proof of Theorem 3.3.

Proof of Theorem 3.3. Observe first that and because $\mathrm{supp}(\mu_0^N) \subset B(R)$ for each $N \geq 1$ and we assumed $\lambda > 0$ to be sufficiently large, there exists a unique optimal control $\theta^N \in L^2([0,T]; \mathbb{R}^m)$ solution of (3.15) as a consequence of Theorem 3.2. Noticing again that

$$J^N(\theta^N) \leq J^N(0) \leq \|\ell\|_{\mathcal{C}^1(B(R))} + 1$$

for each $N \geq 1$, the sequence (θ^N) is uniformly contained in the closed ball $\Gamma \subset L^2([0,T]; \mathbb{R}^m)$ whose radius is defined in Theorem 3.2, and as such it admits a subsequence (that we do not relabel) which converges weakly to some $\theta^* \in L^2([0,T]; \mathbb{R}^m)$.

Our goal is to show that θ^* is the unique minimizer of J and that the optimal values $(J^N(\theta^N))$ converge towards $J(\theta^*)$. To this end observe first that by Mazur's lemma (see e.g. [20, Corollary 3.8]), there exists a sequence $(\widetilde{\theta}^N)$ made of convex combinations of the elements of (θ^N) such that

$$\widetilde{\theta}^N \underset{N\to+\infty}{\longrightarrow} \theta^* \qquad \text{in } L^2([0,T]; \mathbb{R}^m).$$

Recalling that θ^N are minimizers of J^N and that these latter are uniformly equi-Lipschitz over Γ as a consequence of Lemma 3.1, it further holds that

$$\begin{aligned} J^N(\theta^N) &\leq J^N(\widetilde{\theta}^N) \\ &\leq J^N(\theta^*) + \big(\mathcal{L}(T, R, \Gamma) + 2\lambda\big)\big\|\theta^* - \widetilde{\theta}^N\big\|_2, \end{aligned}$$

for each $N \geq 1$. Using the stability estimate (3.20) of Lemma 3.2, we can pass to the limit in the previous expression and obtain that

$$\limsup_{N\to+\infty} J^N(\theta^N) \leq J(\theta^*). \tag{3.24}$$

In order to recover a similar inequality for the liminf, notice that the reduced costs J^N are convex by Proposition 3.1, which implies that

$$\begin{aligned} J^N(\theta^N) &\geq J^N(\theta^*) + \big\langle \nabla J^N(\theta^*), \theta^N - \theta^* \big\rangle_{L^2([0,T];\mathbb{R}^m)} \\ &\geq J^N(\theta^*) + \big\langle \nabla J(\theta^*), \theta^N - \theta^* \big\rangle_{L^2([0,T];\mathbb{R}^m)} + \big\langle \nabla J(\theta^*) - \nabla J^N(\theta^*), \theta^N - \theta^* \big\rangle_{L^2([0,T];\mathbb{R}^m)} \end{aligned} \tag{3.25}$$

for each $N \geq 1$. Observe now that by (3.21) in Lemma 3.2, one has that

$$\big\|\nabla J(\theta^*) - \nabla J^N(\theta^*)\big\|_2 \underset{N\to+\infty}{\longrightarrow} 0,$$

which together with the fact that $(\theta^N) \subset \Gamma$ is converging weakly towards θ^* then yields

$$\big\langle \nabla J(\theta^*) - \nabla J^N(\theta^*), \theta^* - \theta^N \big\rangle_{L^2([0,T];\mathbb{R}^m)} \underset{N\to+\infty}{\longrightarrow} 0,$$

by standard results on weak-strong convergence (see e.g. [20, Proposition 3.5]). Thus, by passing to the limit as $N \to +\infty$ in (3.25) while using (3.20) of Lemma 3.2, we recover

$$J(\theta^*) \leq \liminf_{N\to+\infty} J^N(\theta^N), \tag{3.26}$$

which together with (3.24) finally implies that

$$J^N(\theta^N) \underset{N\to+\infty}{\longrightarrow} J(\theta^*). \tag{3.27}$$

In order to conclude that θ^* is a minimizer of J, it is sufficient to consider a minimizing sequence $(\theta^n) \subset \Gamma$ for (1.8) and to observe that by Lemma 3.2 and (3.27), it holds that

$$J(\theta^n) = \lim_{N\to+\infty} J^N(\theta^n) \geq \lim_{N\to+\infty} J^N(\theta^N) = J(\theta^*)$$

and to let $n \to +\infty$. The strict convexity of J in turn provides the uniqueness of θ^*, from whence we can deduce that it is the weak limit of the whole sequence (θ^N). □

4 Mean-Field Maximum Principle

In this section, we investigate first-order optimality conditions for the mean-field optimal control problem (1.8), which take the form of a mean-field Pontryagin Maximum Principle ("PMP" for short). Their derivation – which is based on a Lagrange multiplier rule for the convex calculus introduced in Section 2 – is heuristically presented in Section 4.1. After studying the well-posedness of the optimality system in Section 4.2, we proceed to rigorously establish the PMP throughout Section 4.3.

4.1 Formal derivation of the Lagrangian maximum principle

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 (1.8), defined by

$$\begin{aligned}\mathcal{L}(\mu,\theta,\psi) = &\int_{\mathbb{R}^{2d}} \ell(x,y)\mathrm{d}\mu_T(x,y) + \lambda\int_0^T |\theta_t|^2\,\mathrm{d}t \\ &+ \int_{\mathbb{R}^{2d}} \psi(0,x,y)\mathrm{d}\mu_0(x,y) - \int_{\mathbb{R}^{2d}} \psi(T,x,y)\mathrm{d}\mu_T(x,y) \\ &+ \int_0^T \int_{\mathbb{R}^{2d}} \Big(\partial_t\psi(t,x,y) + \nabla_x\psi(t,x,y)\cdot\mathcal{F}(t,x,\theta_t)\Big)\mathrm{d}\mu_t(x,y)\,\mathrm{d}t\,. \end{aligned} \tag{4.1}$$

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

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

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 (1.8), we will show that there exists a Lagrange multiplier ψ^* such that

$$\frac{\delta \mathcal{L}}{\delta \mu}(\mu^*, \theta^*, \psi^*) = 0 \qquad \text{and} \qquad \frac{\delta \mathcal{L}}{\delta \theta}(\mu^*, \theta^*, \psi^*) = 0\,. \tag{4.2}$$

These latter will in turn provide us with the following backward adjoint dynamics

$$\partial_t \psi^* + \nabla_x \psi^* \cdot \mathcal{F}(t, x, \theta_t^*) = 0, \tag{4.3}$$

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

$$2\lambda \theta_t^{*\top} + \int_{\mathbb{R}^{2d}} \nabla_x \psi^* \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^*) \mathrm{d}\mu_t^*(x, y) = 0, \tag{4.4}$$

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

$$\partial_t \mu_t^* + \nabla_x \cdot (\mathcal{F}(t, x, \theta_t^*)\mu_t^*) = 0, \qquad \mu_t^*|_{t=0} = \mu_0\,. \tag{4.5}$$

We will see below that (4.3) is understood in the sense of (4.71), and that (4.4) is understood in the sense of (4.72).

4.2 Well-posedness of the maximum principle

This section is devoted to discussing 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

$$\begin{cases} \partial_t \mu_t^* + \nabla_x \cdot (\mathcal{F}(t, x, \theta_t^*)\mu_t^*) = 0, \quad \mu_t^*|_{t=0} = \mu_0, & (4.6) \\ \partial_t \psi^* + \nabla_x \psi^* \cdot \mathcal{F}(t, x, \theta_t^*) = 0, \quad \psi_t^*|_{t=T} = \ell, & (4.7) \\ \theta_t^{*\top} = -\dfrac{1}{2\lambda} \displaystyle\int_{\mathbb{R}^{2d}} \nabla_x \psi^* \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^*) \mathrm{d}\mu_t^*(x, y). & (4.8) \end{cases}$$

To do so, 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

$$\Gamma_{M,C} := \left\{ \theta \in \mathcal{C}([0,T]; \mathbb{R}^m) \;\middle|\; |\theta_t - \theta_s| \le M|t-s|,\ \|\theta\|_\infty \le C_\Gamma \right\}. \tag{4.9}$$

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

$$\Gamma_C := \left\{ \theta \in L^2([0,T]; \mathbb{R}^m) \mid \|\theta\|_2 \le C_\Gamma T^{\frac{1}{2}} \right\}. \tag{4.10}$$

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

Theorem 4.1. *For any given $T > 0$, take an initial data $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$ and a terminal condition ψ_T satisfying (4.14), let $\mathcal{F}$ be a map satisfying Assumptions 1 and 2, and suppose that $\lambda > 0$ is large enough.*

Then, there exists a triple $(\mu^*, \theta^*, \psi^*) \in \mathcal{C}([0,T]; \mathcal{P}_c(\mathbb{R}^{2d})) \times \mathrm{Lip}([0,T]; \mathbb{R}^m) \times \mathcal{C}^1([0,T]; \mathcal{C}_c^2(\mathbb{R}^{2d}))$ solution of (4.6)-(4.8). Moreover, the control solution θ^* is unique in $\Gamma_C \subset L^2([0,T]; \mathbb{R}^m)$ defined as in (4.10), and $\psi^* \in \mathcal{C}^1([0,T]; \mathcal{C}_c^2(\mathbb{R}^{2d}))$ is in characteristic form.

Remark 4.1. *If there exists an optimal control $\theta^* \in L^2([0,T];\mathbb{R}^m)$ satisfying the maximum principle* (4.6)-(4.8)*, then the uniqueness result in Theorem 4.1 ensures that θ^* coincides with a Lipschitz continuous function almost everywhere. This means that in such a case there exists a smooth optimal control $\theta^* \in \mathrm{Lip}([0,T];\mathbb{R}^m)$.*

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

Proposition 4.2. *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 1. 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* (4.6) *in the sense of Definition 2.2. Moreover, there exists some $R_T > 0$ depending only on R and $C_{\mathcal{F}}$, such that*

$$\mathrm{supp}(\mu_t^\theta) \subset B(R_T) \qquad \text{for all } t \in [0,T]\,. \tag{4.11}$$

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

$$W_1(\mu_t^\theta, \mu_s^\theta) \leq C(R, C_{\mathcal{F}})|t-s|\,. \tag{4.12}$$

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

$$W_1(\mu_t^{\theta,1}, \mu_t^{\theta,2}) \leq e^{L_{\mathcal{F},T,C_\Gamma}} W_1(\mu_0^1, \mu_0^2) \quad \text{for all } t \in [0,T]\,. \tag{4.13}$$

Here $C_{\mathcal{F}}$ and $L_{\mathcal{F},T,C_\Gamma}$ are defined as in Assumption 1 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 (4.7) as $\psi_T \in \mathcal{C}_c^2(\mathbb{R}^{2d})$ with

$$\mathrm{supp}(\psi_T) = B(R_T) \quad \text{and} \quad \psi_T(x,y) = \ell(x,y) \quad \text{for all } x,y \in B(R_T)\,. \tag{4.14}$$

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

Proposition 4.3. *Suppose that $\mathcal{F}$ satisfies Assumption 1. 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* (4.7) *whose terminal condition satisfies* (4.14)*. Moreover, it holds that*

$$\left\|\psi_t^\theta\right\|_{\mathcal{C}_c^2(\mathbb{R}^{2d})} \leq C(R', T, C_\Gamma, C_{\mathcal{F}}, L_{\mathcal{F},T,C_\Gamma}) \left\|\psi_T\right\|_{\mathcal{C}^2(B(R_T))}\,, \tag{4.15}$$

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 $R' = R + (R + C_{\mathcal{F}}T)e^{C_{\mathcal{F}}T}$.

The results of Proposition 4.3 are classical, and we postpone their proof to Appendix A.

Remark 4.2. *Here, the fact that ψ^θ is a characteristic solution means that it is obtained via the characteristic method, and is of the form $\psi^\theta(t,x,y) = \psi_T(\mathbf{\Phi}^\theta_{(T,t)}(x,y))$. Therein, we denoted by $(\mathbf{\Phi}^\theta_{(\tau,t)})_{\tau,t\in[0,T]}$ the flow maps defined as in* (A.3) *with $\mathcal{F}(t,x) := \mathcal{F}(t,x,\theta_t)$. Characteristic solutions to* (4.8) *are unique because of the way they depends on the terminal condition* (4.15)*. 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* (4.8)*, 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 4.1. The existence of optimal controls θ^* in $\Gamma_{M,C}$ is based on the Schauder fixed point theorem [40, Theorem 11.1]. Then, the uniqueness will be obtained by additionally showing that the underlying fixed-point map is 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 (4.6) and by $\psi^\theta \in \mathcal{C}^1([0,T];\mathcal{C}^2_c(\mathbb{R}^{2d}))$ the unique characteristic solution of (4.7). In this context, we introduce the continuous mapping $\Lambda:\ \Gamma_{M,C} \to \mathcal{C}([0,T];\mathbb{R}^m)$, defined by

$$\Lambda(\theta)(t)^\top = -\frac{1}{2\lambda}\int_{\mathbb{R}^{2d}} \nabla_x \psi_t^\theta \cdot \nabla_\theta \mathcal{F}(t,x,\theta_t)\mathrm{d}\mu_t^\theta(x,y), \tag{4.16}$$

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 Assumption 1-(iii) and (4.15) that

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

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

$$\left\|\Lambda(\theta)\right\|_{L^\infty([0,T];\mathbb{R}^m)} \le C_\Gamma. \tag{4.17}$$

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

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

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

$$I_1 \le \frac{1}{2\lambda} C(R_T,T)|t-s|, \tag{4.18}$$

for all $s,t \in [0,T]$. Furthermore, it follows from assumption (3.2) and the estimate (4.15) that

$$\begin{aligned}
I_2 &\le \frac{1}{2\lambda} C(R_T) \sup_{t\in[0,T]} \big\|\psi_t^\theta\big\|_{\mathcal{C}^1(B(R'_T))}\big(|t-s| + |\theta_t - \theta_s|\big) \\
&\le \frac{1}{2\lambda} C(R_T) C(R'_T,T,C_\Gamma,C_\mathcal{F},L_{\mathcal{F},T,C_\Gamma}) \left\|\psi_T\right\|_{\mathcal{C}^1(B(R_T))} M|t-s|,
\end{aligned} \tag{4.19}$$

with $R'_T := R + (R + C_\mathcal{F} T)e^{C_\mathcal{F} T}$. Lastly by Kantorovich's duality formula (2.6), one has

$$I_3 \le \frac{1}{2\lambda}\mathrm{Lip}\big(\nabla_x \psi_s^\theta \cdot \nabla_\theta \mathcal{F}(s,\cdot,\theta_s)\,;B(R_T)\big)W_1(\mu_t,\mu_s), \tag{4.20}$$

and can further notice that

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

where we have used (4.15) and Assumption 2-(*iii*). This combined with (4.12) thus yields

$$I_3\le\frac{1}{2\lambda}C(R_T',T,C_\Gamma,C_{\mathcal{F}},L_{\mathcal{F},T,C_\Gamma},R_T)\,\|\psi_T\|_{\mathcal{C}^2(B(R_T))}\,|t-s|. \tag{4.21}$$

Collecting estimates (4.18), (4.19) and (4.21), we deduce that for $\lambda>0$ large enough, it holds

$$|\Lambda(\theta)(t)-\Lambda(\theta)(s)|\le M|t-s|. \tag{4.22}$$

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

$$\theta^{*\top}=-\frac{1}{2\lambda}\int_{\mathbb{R}^{2d}}\nabla_x\psi_t^{\theta^*}\cdot\nabla_\theta\mathcal{F}(t,x,\theta_t^*)\mathrm{d}\mu_t^{\theta^*}(x,y). \tag{4.23}$$

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

$$\|\theta^1-\theta^2\|_2=\|\Lambda(\theta^1)-\Lambda(\theta^2)\|_2\le\kappa\|\theta^1-\theta^2\|_2,$$

which leads to a contradiction for contraction constants satisfying $0\le\kappa<1$. 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 (4.6) driven by θ^1,θ^2 respectively, with the same initial condition μ_0. Similarly, denote by $\psi^{\theta^1},\psi^{\theta^2}$ the solutions of (4.7) 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

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

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

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

Hence, the estimate in (4.2) is equivalent to

$$|\Lambda(\theta^1)(t) - \Lambda(\theta^2)(t)| \leq \frac{1}{2\lambda} (|I_1| + |I_3| + |I_4|) \,. \tag{4.24}$$

Let us focus on each term separately, starting with 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 (4.21), we have that

$$\begin{aligned}
|I_1| &= \bigg| \int_{B(R_T)} \nabla_x \psi_t^{\theta^1} \cdot \nabla_\theta \mathcal{F}(t,x,\theta_t^1) (\mathrm{d}\mu_t^{\theta^1} - \mathrm{d}\mu_t^{\theta^2})(x,y) \bigg| \\
&\leq \mathrm{Lip} \left(\nabla_x \psi_t^{\theta^1} \cdot \nabla_\theta \mathcal{F}(t,x,\theta_t^1) \,; B(R_T) \right) W_1(\mu_t^{\theta^1}, \mu_t^{\theta^2}) \\
&\leq C(R_T', T, C_\Gamma, C_\mathcal{F}, L_{\mathcal{F},T,C_\Gamma}, R_T) \, \| \psi_T \|_{\mathcal{C}^2(B(R_T'))} \, W_1(\mu_t^{\theta^1}, \mu_t^{\theta^2}).
\end{aligned} \tag{4.25}$$

Observe now that following Appendix A, the curves $\mu_t^{\theta^1}$ and $\mu_t^{\theta^2}$ are characteristic solutions of (4.6), in the sense that

$$\mu_t^{\theta^i} = \mathbf{\Phi}_{(0,t)}^{\theta^i} \sharp \mu_0 \tag{4.26}$$

for all times $t \in [0, T]$, where $\mathbf{\Phi}_{(0,t)}^{\theta^i}(\cdot)$ are the flow maps of the underlying ODEs

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

for $i = 1, 2$. Then, it follows from Assumption 1 that

$$\begin{aligned}
\left| (X_t^1, Y_t^1) - (X_t^2, Y_t^2) \right| &= \left| \left(x_0 - x_0 + \int_0^t (\mathcal{F}(s, X_s^1, \theta_s^1) - \mathcal{F}(s, X_s^2, \theta_s^2)) \mathrm{d}s, y_0 - y_0 \right) \right| \\
&\leq \int_0^t \left| (\mathcal{F}(s, X_s^1, \theta_s^1) - \mathcal{F}(s, X_s^2, \theta_s^2) \right| \mathrm{d}s \\
&\leq \int_0^t \left| \mathcal{F}(t, X_s^1, \theta_s^1) - \mathcal{F}(t, X_s^2, \theta_s^1) \right| \mathrm{d}s + \int_0^t \left| \mathcal{F}(t, X_s^2, \theta_s^1) - \mathcal{F}(t, X_s^2, \theta_s^2) \right| \mathrm{d}s \\
&\leq L_{\mathcal{F},T,C_\Gamma} \int_0^t |X_s^1 - X_s^2| \mathrm{d}s + C(R_T, T) \int_0^t |\theta_s^1 - \theta_s^2| \mathrm{d}s.
\end{aligned} \tag{4.27}$$

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

$$W_1(\mu_t^{\theta^1}, \mu_t^{\theta^2}) \leq W_1 \Big(\mathbf{\Phi}_{(0,t)}^{\theta^1} \sharp \mu_0, \mathbf{\Phi}_{(0,t)}^{\theta^2} \sharp \mu_0 \Big) \leq C(R_T, T) e^{L_{\mathcal{F},T,C_\Gamma} T} \|\theta^1 - \theta^2\|_2, \tag{4.28}$$

and by using (4.28) in (4.25), it further holds that

$$|I_1| \leq C(R_T', T, C_\Gamma, C_\mathcal{F}, L_{\mathcal{F},T,C_\Gamma}, R_T) \, \|\psi_T\|_{\mathcal{C}^2(B(R_T))} \, \|\theta^1 - \theta^2\|_2. \tag{4.29}$$

We now shift our focus to the integral I_3. By Assumption 2-(i), we have that

$$\begin{aligned} |I_3| &= \left| \int_{B(R_T)} (\psi_t^{\theta^1} - \psi_t^{\theta^2}) \nabla_x \cdot \nabla_\theta \mathcal{F}(t, x, \theta_t^1) \mathrm{d}\mu_t^{\theta^2}(x, y) \right| \\ &\leq C(R_T, T, C_\Gamma) \sup_{t \in [0,T]} \left\| \psi_t^{\theta^1} - \psi_t^{\theta^2} \right\|_{\mathcal{C}(B(R_T'))}. \end{aligned} \tag{4.30}$$

Recalling that $\psi^{\theta^1}, \psi^{\theta^2}$ are characteristic solutions of (4.7) while using (A.16), one further has

$$\begin{aligned} \left\| \psi_t^{\theta^1} - \psi_t^{\theta^2} \right\|_{\mathcal{C}(B(R_T'))} &= \left\| \psi_T\big(\mathbf{\Phi}_{(t,T)}^{\theta^1}\big) - \psi_T\big(\mathbf{\Phi}_{(t,T)}^{\theta^2}\big) \right\|_{\mathcal{C}(B(R_T'))} \\ &\leq \|\psi_T\|_{\mathcal{C}^1(B(R_T))} \left\| \mathbf{\Phi}_{(t,T)}^{\theta^1} - \mathbf{\Phi}_{(t,T)}^{\theta^2} \right\|_{C(B(R_T))}. \end{aligned} \tag{4.31}$$

Besides, it simply follows from Proposition B.1 that

$$\sup_{t \in [0,T]} \left\| \mathbf{\Phi}_{(t,T)}^{\theta^1} - \mathbf{\Phi}_{(t,T)}^{\theta^2} \right\|_{\mathcal{C}(B(R_T))} \leq C(R_T, T) e^{L_{\mathcal{F},T,C_\Gamma} T} \|\theta^1 - \theta^2\|_2, \tag{4.32}$$

for some given constant $C(R_T, T)e^{L_{\mathcal{F},T,C_\Gamma}T} > 0$. Therefore, the term I_3 can be estimated as

$$|I_3| \leq C(T, R_T, C_\Gamma, C_\mathcal{F}) \, \|\psi_T\|_{\mathcal{C}^1((B(R_T))} \, \|\theta^1 - \theta^2\|_2 \, . \tag{4.33}$$

Lastly, we focus on the integral quantity I_4. Using Assumption (1)-(iv), we can write

$$\begin{aligned} |I_4| &\leq \int_{\mathbb{R}^{2d}} |\nabla_x \psi_T^{\theta^2}| \left| \nabla_\theta (\mathcal{F}(t, x, \theta_t^1) - \mathcal{F}(t, x, \theta_t^2)) \right| \mathrm{d}\mu_t^{\theta^2}(x, y) \\ &\leq \int_{\mathbb{R}^{2d}} |\nabla_x \psi_T^{\theta^2}| \left| \nabla_\theta^2 \mathcal{F}(t, x, \theta) \right| \left| \theta_t^1 - \theta_t^2 \right| \mathrm{d}\mu_t^{\theta^2}(x, y) \\ &\leq C(R_T, T) |\theta_t^1 - \theta_t^2| \sup_{t \in [0,T]} \left\| \psi_t^{\theta^2} \right\|_{\mathcal{C}^1(B(R_T'))} \\ &\leq C(R_T, T, R_T', C_\Gamma, C_\mathcal{F}, L_{\mathcal{F},T,C_\Gamma}) \|\psi_T\|_{\mathcal{C}^1(B(R_T))} |\theta_t^1 - \theta_t^2| \, . \end{aligned} \tag{4.34}$$

Collecting the estimates from (4.29), (4.33) and (4.34), we can conclude

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

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 (4.6)-(4.8), and it is unique in Γ_C. □

Remark 4.3. *As it was shown in the proof above, the size condition imposed on λ depends on some constant $C(|R_T'|, R_T, T, C_\mathcal{F}, C_\Gamma, L_{\mathcal{F},T,C_\Gamma})$ and $\|\psi_T\|_{\mathcal{C}^2(R_T)}$. Especially for the case $\mathcal{F}(t, x, \theta) := \tanh(\theta x)$, 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 4.4. *For any $T, > 0$, let $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$ be such that $\mathrm{supp}(\mu_0) \subset B(R)$ and and ψ_T be a terminal condition satisfying (4.14), and suppose Assumptions 1 and 2 hold. Moreover, assume that for each $N \geq 1$ we are given an approximating empirical measure of the form*

$$\mu_0^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{(X_0^i, Y_0^i)} \in \mathcal{P}_c^N(\mathbb{R}^{2d}),$$

such that

$$\lim_{N\to\infty} W_1(\mu_0^N, \mu_0) = 0.$$

Let $\lambda > 0$ be sufficiently large 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 (4.6)-(4.8) with initial conditions μ_0 and μ_0^N respectively. Then

$$\max\left\{ \|\theta^N - \theta^*\|_p \,,\, \sup_{t\in[0,T]} W_1(\mu_t^N, \mu_t^*) \,,\, \|\psi^N - \psi^*\|_{\mathcal{C}([0,T]\times B(R_T))} \right\} \leq C W_1(\mu_0^N, \mu_0), \tag{4.35}$$

for a constant $C > 0$ which only depends on the parameters of the model and $p \in [1, +\infty]$, and where $R_T > 0$ is defined as in Proposition 4.2 above. In particular, we obtain the following quantitative generalization error estimate

$$\left| \int_{\mathbb{R}^{2d}} \ell(x,y)\, \mathrm{d}\mu_T^*(x,y) - \frac{1}{N}\sum_{i=1}^{N} \ell\big(X_T^i, Y_T^i\big) \right| \leq C W_1(\mu_0^N, \mu_0). \tag{4.36}$$

Proof. By using similar arguments as in the proof of Theorem 4.1, see in particular (4.13) and (4.27)-(4.28), we can prove the stability estimate

$$\begin{aligned}
\sup_{t\in[0,T]} W_1(\mu_t^N, \mu_t^*) &\leq \sup_{t\in[0,T]} W_1(\mu_t^N, \mu_t^{\theta^N}) + \sup_{t\in[0,T]} W_1(\mu_t^{\theta^N}, \mu_t^*) \\
&\leq C\left(W_1(\mu_0^N, \mu_0) + \int_0^T |\theta_t^N - \theta_t^*|\, \mathrm{d}t \right) \qquad (4.37) \\
&\leq C\Big(W_1(\mu_0^N, \mu_0) + \|\theta^N - \theta^*\|_p \Big), \qquad (4.38)
\end{aligned}$$

where $\mu_t^{\theta^N}$ is the unique solution of (4.5) driven by θ^N with initial datum μ_0, and $C > 0$ is an overloaded constant depending on the data of the problem. Similarly, from (4.27), (4.31) and (4.32), we have that

$$\|\psi^N - \psi^*\|_{\mathcal{C}([0,T]\times B(R_T))} \leq C \int_0^T |\theta_t^N - \theta_t^*|\, \mathrm{d}t \leq C\|\theta^N - \theta^*\|_p, \tag{4.39}$$

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 4.1, see in particular (4.24), (4.25), (4.30) and (4.34), we obtain

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

where we applied (4.37) and (4.39) in the last inequality. Hence for $\lambda > 0$ large enough, it holds

$$\|\theta^N - \theta^*\|_p \leq C W_1(\mu_0^N, \mu_0). \tag{4.40}$$

Combining now (4.37), (4.39) and (4.40) finally yields (4.35). The generalization error displayed in (4.36) follows from (4.37) and (4.40), since

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

This completes the proof of Corollary 4.4. □

Remark 4.4 (Data bounds, regularization parameters and error estimates)**.** *The estimate* (4.36) *is in the worst case affected by the curse of dimension, although it will not be the case in practice e.g. for networks driven by sigmoid activation functions. The constant C in* (4.36) *is encoding the complexity of the NeurODE and is derived as a consequence of* (4.40) *as*

$$C = C_1(1 - C_0/\lambda) > 0.$$

Therein, the constant $C_0 > 0$ may depend exponentially on the constants $C_\mathcal{F}$ and $L_\mathcal{F}$ appearing in Assumptions 1 – and in particular on the dimension $d \geq 1$ of the state space –, and polynomially on those of Assumptions 2, owing to the pessimistic nature of deterministic Grönwall estimates. Thus, as long as the worst-case Grönwall estimates do indeed reflect the actual stability of the PMP, the constant $C_0 > 0$ may be extremely large. Nevertheless, in the case of sigmoidal-type activation functions such as $\rho := \tanh$, we detailed in Remark 3.1 how the uniform boundedness of the velocity field $\mathcal{F}$ implied a polynomial dependence of all the relevant constants of the problem with respect to the state space dimension. Therefore, in that particular yet relevant case, the quantity C_0 will in fact scale polynomially and not exponentially with d.

For arbitrary initial measures μ_0, it is known that empirical measures μ_N supported on finite samples satisfy the estimate

$$\mathbb{E}\big[W_1(\mu_0^N, \mu_0)\big] \leq C N^{-1/d},$$

see for instance [30, 38]*, which scales quite badly with the dimension $d \geq 1$ of the state space. However, if μ_0 is concentrated around manifolds of lower dimension, then the factor $C > 0$ depends favorably on that intrinsic lower dimension* [72]*. In practice, it is expected that data distributions do concentrate around such lower-dimensional structures.*

4.3 Rigorous derivation of the mean-field maximum principle

The previous section, we proved the well-posedness of the mean-field PMP (4.6)-(4.8) in the class of control that are Lipschitz continuous with respect to time. Under this assumption, we rigorously derive in what follows the optimality conditions by using a generalized Lagrange multiplier theorem over convex sets. The method we present is to a certain extent a standard calculus of variations approach, and allows to bypass the more technical ones based either on the abstract differential calculus of Wasserstein as in [13, 15, 18], or on the fine structural results for continuity equations leveraged in [21].

Let it be stressed that the requirement of continuity of the control is purely technical, and stems from our use of [63, Theorem 1] concerning the well-posedness of transport equations with sources. Were such results available in the case where the source terms are merely measurable in time – which seems true but is not written anywhere yet –, we could then remove the continuity assumption and prove the mean-field PMP in its full generality using the Lagrangian approach.

4.3.1 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 $x^* \in E$, we introduce the notation

$$DG(x^*) := \Big\{ L \in \mathcal{L}(X_E, Y) \;\Big|\; L \text{ satisfies (2.15)} \Big\}. \tag{4.41}$$

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 [75, 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 Appendix C.

Theorem 4.5. *Let $x^* \in E$ be a solution of the constrained optimization problem*

$$\begin{cases} \inf\limits_{x \in E} J(x), \\ \text{s.t. } G(x) = 0. \end{cases} \tag{4.42}$$

Suppose moreover 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

$$\langle J'(x^*), z \rangle + \langle G'(x^*)z, p^* \rangle = 0 \tag{4.43}$$

for all $z \in \overline{X}_E$.

4.3.2 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 1.

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

$$\begin{aligned}
\|\partial_t \mu_t\|_{(\mathcal{C}_b^1(\mathbb{R}^{2d}))'} &= \sup_{\|\varphi\|_{\mathcal{C}_b^1} \leq 1} |\langle \partial_t \mu_t, \varphi \rangle| && (4.44)\\
&= \sup_{\|\varphi\|_{\mathcal{C}_b^1} \leq 1} |\langle \mathcal{F}(t,\cdot,\theta_t)\mu_t, \nabla_x \varphi \rangle| \\
&\leq \|\mathcal{F}\|_{L^\infty(\mathrm{supp}(\mu_t))} \leq C_{\mathcal{F}}(1+|R_T|). && (4.45)
\end{aligned}$$

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

$$\begin{aligned}
\|\partial_t \mu_t - \partial_s \mu_s\|_{(\mathcal{C}_b^1(\mathbb{R}^{2d}))'} &= \sup_{\|\varphi\|_{\mathcal{C}_b^1} \leq 1} |\langle \partial_t \mu_t - \partial_s \mu_s, \varphi \rangle| && (4.46)\\
&= \sup_{\|\varphi\|_{\mathcal{C}_b^1} \leq 1} |\langle \mathcal{F}(t,\cdot,\theta_t)\mu_t - \mathcal{F}(s,\cdot,\theta_s)\mu_s, \nabla_x \varphi \rangle| \\
&\leq \sup_{\|\varphi\|_{\mathcal{C}_b^1} \leq 1} \Big| \big\langle (\mathcal{F}(t,\cdot,\theta_t) - \mathcal{F}(s,\cdot,\theta_s))\mu_t, \nabla_x \varphi \big\rangle \Big| && (4.47)\\
&\quad + \sup_{\|\varphi\|_{\mathcal{C}_b^1} \leq 1} |\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla_x \varphi \rangle| && (4.48)\\
&\leq C|t-s| + \sup_{\|\varphi\|_{\mathcal{C}_b^1} \leq 1} |\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla_x \varphi \rangle|, && (4.49)
\end{aligned}$$

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

$$\begin{aligned}
&\sup_{\|\varphi\|_{\mathcal{C}_b^1} \leq 1} |\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla_x \varphi \rangle| \\
&\leq \sup_{\|\varphi\|_{\mathcal{C}_b^1} \leq 1} \big| \big\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), (\nabla_x \varphi - \nabla_x \varphi^n) \big\rangle \big| + \sup_{\|\varphi\|_{\mathcal{C}_b^1} \leq 1} |\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla_x \varphi^n \rangle| \\
&\leq C\|\varphi^n - \varphi\|_{\mathcal{C}_b^1(\mathbb{R}^{2d})} + \mathrm{Lip}\big(\mathcal{F}(t,\cdot,\theta_t)\cdot \nabla_x \varphi^n\big) W_1(\mu_t,\mu_s) && (4.50)\\
&\leq C\|\varphi^n - \varphi\|_{\mathcal{C}_b^1(\mathbb{R}^{2d})} + C_n|t-s|, && (4.51)
\end{aligned}$$

where we have used the Kantorovitch duality (2.6) and (A.7), which further yields that

$$\lim_{s \to t} \sup_{\|\varphi\|_{\mathcal{C}_b^1} \leq 1} |\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla_x \varphi \rangle| \leq \|\varphi^n - \varphi\|_{\mathcal{C}_b^1(\mathbb{R}^{2d})}, \tag{4.52}$$

for every $n \in \mathbb{N}$. Therefore letting $n \to +\infty$ in (4.52), we can conclude

$$\sup_{\|\varphi\|_{\mathcal{C}_b^1} \leq 1} |\langle \mathcal{F}(s,\cdot,\theta_s)(\mu_t - \mu_s), \nabla_x \varphi \rangle| \underset{s \to t}{\longrightarrow} 0. \tag{4.53}$$

This combined with (4.46) 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 shall 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 (1.8), 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 $\partial_t \mu \in \mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')$.

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

$$V := \widetilde{\mathcal{C}}([0,T];\mathcal{M}_{1,c}(\mathbb{R}^{2d})) \cap \mathcal{C}^1([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))') \qquad \text{and} \qquad Q := \mathcal{C}([0,T];\mathbb{R}^m), \tag{4.54}$$

where

$$\widetilde{\mathcal{C}}([0,T];\mathcal{M}_{1,c}(\mathbb{R}^{2d})) := \Big\{ \mu \in \mathcal{C}([0,T];\mathcal{M}_{1,c}(\mathbb{R}^{2d})) \;\Big|\; \mathrm{supp}(\mu_t) \subset S_\mu \text{ for all } t \in [0,T] \\ \text{where } S_\mu \subset \mathbb{R}^d \text{ is a compact set} \Big\}, \tag{4.55}$$

and fix

$$E := V \times Q \;=\; \widetilde{\mathcal{C}}([0,T];\mathcal{M}_{1,c}(\mathbb{R}^{2d})) \cap \mathcal{C}^1([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))') \times \mathcal{C}([0,T];\mathbb{R}^m). \tag{4.56}$$

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

$$X := U \times Q = \mathcal{C}^1([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))') \times \mathcal{C}([0,T];\mathbb{R}^m). \tag{4.57}$$

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 use the notation $U_V := \mathbb{R}(V-V)$ as well as the identity

$$U_V := \widetilde{\mathcal{C}}([0,T];\mathcal{M}_{0,c}(\mathbb{R}^{2d})) \cap \mathcal{C}^1([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))'). \tag{4.58}$$

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

$$U_\nu := \mathbb{R}_+(V-\nu) \subset U_V\,, \tag{4.59}$$

in keeping with the concepts 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

$$X_E := U_V \times Q \;=\; \widetilde{\mathcal{C}}([0,T];\mathcal{M}_{0,c}(\mathbb{R}^{2d})) \cap \mathcal{C}^1([0,T];(\mathcal{C}^1_b(\mathbb{R}^{2d}))') \times \mathcal{C}([0,T];\mathbb{R}^m). \tag{4.60}$$

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 4.5.

∘ **The setup of maps.** For any $(\mu,\theta) \in E$, we denote the full cost functional of (1.8) by

$$J(\mu,\theta) := \int_{\mathbb{R}^{2d}} \ell(x,y)\mathrm{d}\mu_T(x,y) + \lambda \int_0^T |\theta_t|^2 \,\mathrm{d}t, \tag{4.61}$$

and observe that it is a map from E into $\mathbb{R}_+$. We also introduce the notation

$$G(\mu,\theta) := -\partial_t \mu - \nabla_x \cdot (\mathcal{F}(t,x,\theta)\mu)\,. \tag{4.62}$$

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

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

By performing density arguments similar to those of (4.50)-(4.53), one has that

$$\sup_{\|\varphi\|_{\mathcal{C}_b^1}\leq 1} |\langle \mathcal{F}(s,x,\theta_s)(\mu_t-\mu_s),\nabla\varphi\rangle| \leq C\|\mu_t-\mu_s\|_{(\mathcal{C}^1)'}. \tag{4.63}$$

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

$$\mathrm{supp}(\mu_t)\subset S_\mu \quad \text{ for all } t\in[0,T]\,. \tag{4.64}$$

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

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

This allows us to define the Banach space

$$Y:=\overline{Y}_0=\overline{\widetilde{\mathcal{C}}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))'_{0,c})}, \tag{4.65}$$

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

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

where we have again used density arguments similar to that of (4.50)-(4.53). Thus, we have proven that

$$\begin{aligned} \big\|G(\mu^1,\theta^1)-G(\mu^2,\theta^2)\big\|_{\mathcal{C}([0,T];(\mathcal{C}_b^1(\mathbb{R}^{2d}))')} &\leq C\big\|\mu^1-\mu^2\big\|_{\mathcal{C}^1([0,T];\mathcal{C}_b^1(\mathbb{R}^{2d}))}\\ &\quad+C(R_T,T)\|\theta_1-\theta_2\|_{\mathcal{C}([0,T])}\,, \end{aligned} \tag{4.66}$$

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

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

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

$$\langle \mathrm{d}_\mu G(\mu,\theta)(\nu), \varphi\rangle = \left\langle \lim_{\varepsilon\to 0^+} \frac{G(\mu+\varepsilon\nu,\theta)-G(\mu,\theta)}{\varepsilon}, \varphi \right\rangle \tag{4.67}$$

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

$$= \langle -\partial_t\nu - \nabla_x\cdot(\mathcal{F}(t,x,\theta)\nu), \varphi\rangle < +\infty. \tag{4.68}$$

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

$$\langle G'_\theta(\mu,\theta)(\alpha),\varphi\rangle = \lim_{\varepsilon\to 0^+} \frac{\langle G(\mu,\theta+\varepsilon\alpha),\varphi\rangle - \langle G(\mu,\theta),\varphi\rangle}{\varepsilon} = \langle -\nabla_x\cdot(\nabla_\theta\mathcal{F}(t,x,\theta)\alpha\mu),\varphi\rangle < +\infty\,. \tag{4.69}$$

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

$$J'_\mu(\mu,\theta)(\nu) = \int_{\mathbb{R}^{2d}} \ell(x,y)\mathrm{d}\nu_T \qquad \text{and} \qquad J'_\theta(\mu,\theta)(\alpha) = \int_0^T 2\lambda\theta_t\cdot\alpha_t\,\mathrm{d}t\,, \tag{4.70}$$

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

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

We are now ready to present the derivation of the first order optimality condition (4.6)-(4.8) in the class of continuous controls, by means of a Lagrange multiplier rule tailored to the calculus for convex functions introduced in Section 2.3.

Theorem 4.6 (Abstract Lagrange multiplier theorem)**.** *Let $(\mu^*,\theta^*) \in E \subset X = U\times Q$ be a solution to the optimal control problem* (1.8)*. Then there exists $p^* \in Y'$ such that*

$$\begin{cases} \langle G'_\mu(\mu^*,\theta^*)(\nu), p^*\rangle + J'_\mu(\mu^*,\theta^*)(\nu) = 0, & \text{for all } \nu \in \overline{U}_V, \qquad (4.71)\\ \langle G'_\theta(\mu^*,\theta^*)(\alpha), p^*\rangle + J'_\theta(\mu^*,\theta^*)(\alpha) = 0, & \text{for all } \alpha \in Q\,. \qquad (4.72)\end{cases}$$

Remark 4.5. *The solution $\psi^* = p^* \in \mathcal{C}^1([0,T];\mathcal{C}_c^2(\mathbb{R}^{2d}))$ constructed in Proposition 4.3 is in Y'. This comes from the fact that, 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 4.5 whose 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 (4.56) and (4.60) 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 element

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

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

$$G'_\mu(\mu^*,\theta^*)(\nu)=\eta\,, \tag{4.73}$$

which is understood in the sense of

$$\langle G'_\mu(\mu^*_t,\theta^*_t)(\nu_t),\varphi\rangle=\langle\eta_t,\varphi\rangle\qquad\text{for all }\varphi\in\mathcal{C}^1_b(\mathbb{R}^{2d})\,. \tag{4.74}$$

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

$$\partial_t\nu_t+\nabla_x\cdot(\mathcal{F}(t,x,\theta^*_t)\nu_t)=-\eta_t\,, \tag{4.75}$$

with source term $(-\eta)$ and initial condition $\nu_0\in U_{\mu_0}$. 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

$$\sup_{t\in[0,T]}|\langle\eta^n_t-\eta_t,\varphi\rangle|\underset{n\to+\infty}{\longrightarrow}0. \tag{4.76}$$

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 [63, Theorem 1] that there exists a unique measure solution $\mu^{1,n}\in V$ to the following transport equation

$$\partial_t\mu^{1,n}_t+\nabla_x\cdot(\mathcal{F}(t,x,\theta^*_t)\mu^{1,n}_t)=-\eta^n_t,\qquad\mu^{1,n}_t|_{t=0}=\mu^1_0\in\mathcal{P}_c(\mathbb{R}^{2d})\,, \tag{4.77}$$

understood analogously to (2.11) in the sense of distribution, 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 $i\in\{0,1,\cdots,2^k-1\}$, we denote by $\mathcal{F}_{i\Delta t}=\mathcal{F}(i\Delta t,x,\theta^*_{i\Delta t})$ and set

$$\mu^{1,n,(k)}_t=\Gamma^{\mathcal{F}_{i\Delta t}}_{t-i\Delta t}\sharp\mu^{1,n,(k)}_{i\Delta t}-(t-i\Delta t)\eta^n_{i\Delta t},\quad t\in[i\Delta t,(i+1)\Delta t]\,, \tag{4.78}$$

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

$$\begin{cases}\partial_t f_t+\nabla\cdot(\mathcal{F}_{i\Delta t}f_t)=0,\qquad t\in(i\Delta t,(i+1)\Delta t],\\ f_{i\Delta t}=\mu^{1,n,(k)}_{i\Delta t}\,,\end{cases} \tag{4.79}$$

which is is explicitly written as a pushforward through a characteristic flow. From (4.78), we know the sequence $(\mu^{1,n,(k)}_t)_{k\in\mathbb{N}}$ has uniformly bounded support, since

$$\mathrm{supp}(\mu^{1,n,(k)}_t)\subset B(R_T)\cup S_{\eta^n} \tag{4.80}$$

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 (4.79) and the support of the source term. Similarly, it holds for $t \in [i\Delta t, (i+1)\Delta t]$

$$\|\mu_t^{1,n,(k)}\|_{(\mathcal{C}_b^1)'} \leq \|\Gamma_{t-i\Delta t}^{\mathcal{F}_{i\Delta t}} \sharp \mu_{i\Delta t}^{1,n,(k)}\|_{(\mathcal{C}_b^1)'} + \Delta t \|\eta_{i\Delta t}^n\|_{(\mathcal{C}_b^1)'} \leq \|\mu_{i\Delta t}^{1,n,(k)}\|_{(\mathcal{C}_b^1)'} + \Delta t \|\eta_{i\Delta t}^n\|_{(\mathcal{C}_b^1)'} \,. \quad (4.81)$$

This provides us with the following upper-bound

$$\sup_{t\in[0,T]} \|\mu_t^{1,n,(k)}\|_{(\mathcal{C}_b^1)'} \leq \|\mu_0^1\|_{(\mathcal{C}_b^1)'} + T \sup_{t\in[0,T]} \|\eta_t^n\|_{(\mathcal{C}_b^1)'} < +\infty \,, \quad (4.82)$$

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 (4.77) such that

$$\sup_{t\in[0,T]} \mathbb{W}_1^{1,1}(\mu^{1,n}, \mu_t^{1,n,(k)}) \underset{k\to+\infty}{\longrightarrow} 0. \quad (4.83)$$

Recall that the generalized Wasserstein metric introduced in [63] is equivalent to the bounded-Lipschitz norm $\|\cdot\|_{BL}$, so that the limit curves $(\mu^{1,n})_{n\in\mathbb{N}}$ satisfy

$$\mathrm{supp}(\mu_t^{1,n}) \subset B(R_T) \cup S_{\eta^n} \quad \text{and} \quad \|\mu_t^{1,n}\|_{(\mathcal{C}_b^1)'} < +\infty \quad (4.84)$$

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 [63, 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

$$\int_{\mathbb{R}^{2d}} \varphi(x,y) \mathrm{d}\mu_{t_2}^1(x,y) - \int_{\mathbb{R}^{2d}} \varphi(x,y) \mathrm{d}\mu_{t_1}^1(x,y) \quad (4.85)$$

$$= \int_{t_1}^{t_2} \int_{\mathbb{R}^{2d}} \nabla_x \varphi(x,y) \cdot \mathcal{F}(s,x,\theta_s^*) \mathrm{d}\mu_s^1(x,y) \, \mathrm{d}s - \int_{t_1}^{t_2} \int_{\mathbb{R}^{2d}} \varphi(x,y) \mathrm{d}\eta_s(x,y) \, \mathrm{d}s \,. \quad (4.86)$$

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

$$\partial_t \mu_t^* + \nabla_x \cdot (\mathcal{F}(t,x,\theta_t^*)\mu_t^*) = 0, \qquad \mu_t^*|_{t=0} = \mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d}) \,, \quad (4.87)$$

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 (4.75), 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

$$G'(\mu^*, \theta^*)(\nu, 0) = G'_\mu(\mu^*, \theta^*)(\nu) + G'_\theta(\mu^*, \theta^*)(0) = \eta \,. \quad (4.88)$$

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

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

The goal of this subsection is to show that solutions $(\mu^*, \theta^*) \in \mathcal{C}([0,T]; \mathcal{P}_c(\mathbb{R}^d)) \times L^2([0,T]; \mathbb{R}^m)$ with a priori discontinuous controls satisfy the optimality condition (4.6)-(4.8) by using the Pontryagin Maximum Principle in Wasserstein spaces studied in [13, 15, 18].

In the sequel, we suppose that the optimal control problem (1.8) admits an optimal trajectory-control pair $(\mu^*, \theta^*) \in \mathrm{Lip}([0,T]; \mathcal{P}_c(\mathbb{R}^{2d}) \times L^2([0,T]; \mathbb{R}^m)$. 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 with the optimal control problem is defined by

$$\mathbb{H}(t, \nu, \theta) := \int_{\mathbb{R}^{4d}} \langle r, \mathcal{F}(t, x, \theta) \rangle \mathrm{d}\nu(x, y, r, s) - \lambda |\theta|^2, \tag{4.89}$$

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 [15] was adapted to unbounded control sets in [16], and can be written in context as follows.

Theorem 4.7 (Pontryagin Maximum Principle)**.** *There exists a radius $R_T' > 0$ and a uniquely determined state-costate curve $\nu^* \in \mathrm{Lip}([0,T], \mathcal{P}_c(\mathbb{R}^{4d}))$ with $\mathrm{supp}(\nu_t^*) \subset B(R_T') \times B(R_T')$ for all times $t \in [0,T]$, such that the following holds.*

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

$$\begin{cases} \partial_t \nu_t^* + \nabla_{(x,y,r,s)} \cdot \left(\mathbb{J}_{4d} \nabla_\nu \mathbb{H}(t, \nu_t^*, \theta_t^*) \nu_t^* \right) = 0, \\ \pi^1_\# \nu_t^* = \mu_t^* \qquad \text{for all times } t \in [0,T], \\ \nu_T^* = (\mathrm{Id}, -\nabla_x \ell) \sharp \mu_T^*, \end{cases} \tag{4.90}$$

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 $(x, y, r, s) \in B(R_T') \times B(R_T')$.

(ii) The maximization condition

$$\mathbb{H}(t, \nu_t^*, \theta_t^*) = \max_{\theta \in \mathbb{R}^m} \mathbb{H}(t, \nu_t^*, \theta), \tag{4.91}$$

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 [18] to build solutions to (4.90), while the necessary part has been

established more recently in [17]. Following the notations of Section 3 and Appendix A, 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 $(x,y) \in B(R'_T)$, where $(\Phi^*_{(\tau,t)})_{\tau,t\in[0,T]}$ is the characteristic flow defined via (3.5) with $\theta_t := \theta^*_t$ being the optimal control.

Proposition 4.8 (Representation formula for state-costate curves)**.** *A state-costate curve $\nu^* \in \mathrm{Lip}([0,T], \mathcal{P}_c(\mathbb{R}^{4d}))$ solves the forward-backward system* (4.90) *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 $t \in [0,T] \mapsto \nu^T_t \in \mathcal{P}_c(\mathbb{R}^{4d})$ is built via the disintegration formula*

$$\nu^T_t := \int_{\mathbb{R}^{2d}} \sigma^*_{t,x,y}(t)\mathrm{d}\mu^*_T(x,y),$$

for all times $t \in [0,T]$. Therein for μ^*_T-almost every $(x,y) \in \mathbb{R}^{2d}$, the curve $t \in [0,T] \mapsto \sigma^*_{t,x,y} \in \mathcal{P}_c(\mathbb{R}^{2d})$ is chosen as the unique solution of the backward adjoint dynamics

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

where

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

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

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 (4.91).

Lemma 4.1 (Fixed-point expression for the optimal control)**.** *Let (μ^*, θ^*) be an optimal pair for the problem* (1.8)*, and ν^* be the corresponding state-costate curve given by Theorem 4.7. Then for $\lambda > 0$ large enough, it holds that*

$$\theta^*_t = \frac{1}{2\lambda} \int_{\mathbb{R}^{4d}} \nabla_\theta \mathcal{F}(t,x,\theta^*_t)^\top r \,\mathrm{d}\nu^*_t(x,y,r,s), \tag{4.92}$$

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

Proof. As a consequence Assumptions 1-(*iv*), 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 $\mathrm{supp}(\nu^*_t) \subset B(R'_T) \times B(R'_T)$, there exists a constant $C(R'_T) > 0$ such that

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

Hence for $\lambda > C(R'_T)$, the Hamiltonian is a concave function of θ, and the optimal control θ^* satisfies the pointwise maximization condition (4.91) if and only if

$$\nabla_\theta \mathbb{H}(t, \nu^*_t, \theta^*_t) = 0 \qquad \text{for a.e. } t \in [0,T], \tag{4.93}$$

which is equivalent to the fixed-point equation (4.92). □

For all times $t \in [0,T]$, we shall denote by $(x,y) \in B(R'_T) \mapsto \overline{\sigma}^*(t,x,y) \in \mathbb{R}^d$ 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 (4.92) can be rewritten as

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

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 4 satisfies

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

which is exactly (4.4). This is the object of the following proposition, whose proof relies on the explicit characterization of the adjoint of the differential of a flow that we recall in the following lemma. 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 Appendix A for the sake of completeness.

Lemma 4.2. *For every $x \in \mathbb{R}^d$ and $\theta \in L^2([0,T];\mathbb{R}^m)$, the map $t \in [0,T] \mapsto \nabla_x \Phi^\theta_{(t,T)}(\Phi^\theta_{(T,t)}(x))^\top$ is the unique solution of the backward adjoint Cauchy problem*

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

Proposition 4.9 (Rigorous link between the Hamiltonian and Lagrangian adjoint states)**.** *Let $\psi^* \in \mathcal{C}^1([0,T];\mathcal{C}^2_c(\mathbb{R}^{2d}))$ be the unique characteristic solution of the formal adjoint equation* (4.7) *associated with 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*

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

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 (4.6)-(4.8).

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 linearized adjoint dynamics associated with the controlled velocity field $\mathcal{F} : [0,T] \times \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$.

Lemma 4.3. *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*

$$\begin{cases} \partial_t \overline{\sigma}^*(t,x,y) = -\nabla_x \mathcal{F}\big(t, \Phi^*_{(T,t)}(x), \theta_t^*\big)^\top \, \overline{\sigma}^*(t,x,y) \\ \overline{\sigma}^*(T,x,y) = -\nabla_x \ell(x,y). \end{cases} \tag{4.94}$$

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 4.8, it holds for any $\xi \in \mathcal{C}^1_c(\mathbb{R}^{2d})$ that

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^{2d}} \xi(r,s) \mathrm{d}\sigma^*_{t,x,y}(r,s) = \int_{\mathbb{R}^{2d}} \Big\langle \nabla_r \xi(r,s), -\nabla_x \mathcal{F}\big(t, \theta^*_t, \Phi^*_{(T,t)}(x)\big)^\top r \Big\rangle \mathrm{d}\sigma^*_{t,x,y}(r,s) \tag{4.95}$$

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 $\zeta, \phi \in \mathcal{C}^1_c(\mathbb{R}^{2d})$. Then given an arbitrary $h \in \mathbb{R}^d$, consider ζ, ϕ to be smooth functions such that

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

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 $(r,s) \in B(R'_T)$, which upon recalling that $\mathrm{supp}(\sigma^*_{t,x,y}) \subset B(R'_T)$ for all times $t \in [0,T]$ yields together with (4.95) 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 (4.94). The uniqueness follows from Assumption 1 together with classical Grönwall estimates. □

Proof of Proposition 4.9. Following Proposition 4.3, 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, we further obtain that

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

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 4.2, 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

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

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 4.10. *For any given $T > 0$, let $\mathcal{F}$ satisfy the Assumption 1 and 2, the initial data $\mu_0 \in \mathcal{P}_c(\mathbb{R}^{2d})$, and the terminal condition ψ_T satisfy* (4.14)*. 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* (4.6)-(4.8) *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 4.1, Theorems 4.6-4.7 and Proposition 4.9. □

5 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 [8, 45, 54, 55]. 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.

5.1 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 (4.6),(4.7) and (4.8) 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 the convergence of the method is achieved. The operator Λ has been introduced in the proof of Theorem 4.1, where we showed that the optimal control is its unique fixed point. In particular, we proved therein that such iterations are contractive as soon as they remain bounded, and provided that the regularization parameter $\lambda > 0$ is large enough. Therefore, by construction, the convergence of the shooting scheme is automatically guaranteed in our setting for bounded iterations. Moreover, Corollary 4.4 also ensures the convergence of the empirical solutions obtained for finite samples as $N \to \infty$. Hence, the combination of the results of Theorem 4.1 and Corollary 4.4 provides a theoretically guaranteed convergence for the shooting method, which is summarized in Algorithm 1.

Algorithm 1 Shooting Technique

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

2: **for** $k = 0 \ldots$ number of iterations **do**

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

$$\partial_t \mu_t^k + \nabla_x \cdot \left(\mathcal{F}(t,x,\theta_t^k)\mu_t^k\right) = 0, \qquad \mu_t^k|_{t=0} = \mu_0\,. \tag{5.1}$$

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

$$\partial_t \psi_t^k + \nabla_x \psi_t^k \cdot \mathcal{F}(t,x,\theta_t^k) = 0, \qquad \psi_t^k(x,y)|_{t=T} = |x-y|^2\,. \tag{5.2}$$

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

$$\theta_t^{k+1} + \frac{1}{2\lambda}\int_{\mathbb{R}^{2d}} \nabla_\theta \mathcal{F}(t,x,\theta_t^{k+1})^\top \nabla_x \psi_t^k(x,y)\mathrm{d}\mu_t^k(x,y) = 0\,. \tag{5.3}$$

6: **end for**

Forward Equation. As already mentioned in the introduction, the dynamics (5.1) is a linear transport equation that describes the forward pass of the initial data through the network. We investigate various ways to solve such a forward equation: our first approach, very much inspired by [55] and the deep learning task that we aim to solve, is a particle method. Given an initial distribution μ_0, we sample N particles and their corresponding labels and evolve them in time for $t \in (0,T]$ according to their governing ODEs

$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = \mathcal{F}(t,X_t^i,\theta_t), \qquad \frac{\mathrm{d}Y_t^i}{\mathrm{d}t} = 0\,, \tag{5.4}$$

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

In order to demonstrate that the convergence and contractivity of the method is independent of the number of particles/samples N and to highlight the power of our mean-field result, we also employ a Monte Carlo method. The idea in this case is to "break up" the particles trajectories by performing density estimations and resamplings at every time step. Namely, we start by sampling N particles from the initial distribution μ_0, let them evolve according to the governing ODE (5.4) during a small time, and then perform a kernel density estimation in order to compute the distribution of the evolved particles, i.e. μ_1^m. The apex m indicates that this process sampling-moving-estimating is repeated for a certain number of repetitions M in order to obtain a result that is independent of the initial sample of particles. Then, the distribution μ_1 is computed as the mean over all the repetitions μ_1^m with $m = 0, ..., M$. Clearly, this process needs to be repeated for every layer $t \in [0,T]$. More rigorously, the method is summarized in Algorithm 2, for a generic iteration k of the shooting method.

By using the Monte Carlo method, we are not only highlighting the mean-field nature of our algorithm (since we can sample as many particles as we want), but also distinguish our method from the ODE-based algorithm in [55]. Indeed, the main difference with their approach

Algorithm 2 Monte Carlo Method

1: **for** $t \in [0, T]$ **do**
2: **for** $m = 0 \dots M$ **do**
3: Sample N particles from μ_t
4: Evolve the N particles according to the ODE (5.4)
5: Use kernel density estimation to compute μ_{t+1}^m
6: **end for**
7: Define $\mu_{t+1} = \frac{1}{M} \sum_{m=1}^{M} \mu_{t+1}^m$
8: **end for**

is that we are considering a mean-field version of the maximum principle, wherein the dynamics is written in terms of PDEs rather than ODEs, and for which the Monte Carlo method is a suitable solver.

In the spirit of the latter issue, we also solve the forward equation with a classical finite volume method, which is a well-known numerical scheme to efficiently tackle generic conservation laws in any dimension. This approach is based on a mesh partition of the domain, and on the integration of the PDE over each control volume, i.e. each element of the mesh, in order to obtain a balance equation that is then discretized. One of the fundamental issues of this context lies in the discretization of the fluxes, which have to be conservative and consistent in order to produce an efficient method. In our case, since the flux depends on the function $\mathcal{F}$, we discretize it by means of an upwind spacial scheme. The drawback of this method is that it is highly dependent on the space and time discretization steps, which are very important parameters whose role will be discussed at the end of this section.

Backward Equation. The backward equation (5.2) is independent of the forward evolution (5.1) and, as such, it can be solved simultaneously. Observe that (5.2) 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, we decide to use finite differences in space and an explicit time-scheme to solve this latter. As it happened for the resolution of (5.1) with finite volumes, the upwind method has been used to perform the space discretization of the velocity of the backward equation. Not only is this method suitable for transport equations, but it is also ideal in the case where the velocity $\mathcal{F}$ depends on both space and time, i.e. when it can change at every point of the domain. Notice that we could solve (5.2) using a finite volume method akin to that described for the forward equation, but this may prove to be inefficient because of the oscillations of ψ_t for some choices of the algorithm parameters. Hence, we chose to focus our attention on the finite difference method, which produces very good results in the low dimensional test cases considered here.

Parameter Update. Finally, we solve (5.3) which allows to update the set of layers. Given the primal-dual solutions (μ_t, ψ_t) of equations (5.1)-(5.2), we can solve (5.3) by computing the root of the following non-linear function

$$f(\theta_t) = \theta_t + \frac{1}{2\lambda} \int_{\mathbb{R}^{2d}} \nabla_\theta \mathcal{F}(t, x, \theta_t)^\top \nabla_x \psi_t(x, y) \mathrm{d}\mu_t(x, y). \tag{5.5}$$

for each $t \in [0, T]$. Inspired by the particle method employed to solve (5.1), the integral with respect to μ_t can be simply computed by means of a particle approximation as μ_t is an empirical distribution in our context. Moreover, given the discrete values of $\psi_t(x, y)$ that have been computed as a by-product of the finite difference scheme used to solve the backward equation (5.2), the function ψ_t and its gradient $\nabla_x \psi_t$ can be interpolated, e.g. using splines, in order to be able to evaluate these latter in whatever position X_t^i the particles may be located at in the domain. Ultimately, the fixed point equation (5.3) can be therefore be approximated by

$$f(\theta_t) \approx \theta_t + \frac{1}{2\lambda N} \sum_{i=0}^{N} \nabla_\theta \mathcal{F}(t, X_i(t), \theta_t)^\top \nabla_x \psi_t(X_i(t), Y_i(t)), \tag{5.6}$$

and its roots can be computed using any classical non-linear equation solver such as Newton-Raphson, Bisection, or Brent's method, depending on the particular test case at and. Notice that here, the only source of approximation is the interpolation error of ψ_t.

In the case where the forward equation has been solved with a Monte Carlo method, the approximation of the integral needs to be performed many times (as for the forward equation) in order to obtain a result which is independent of the initial particle sample, with same number of repetitions $M \geq 1$. Finally, if one chooses to solve the forward equation with a finite volume method, the result μ_t for each $t \in [0, T]$ is not obtained through particle approximations, but as a function on a spatially discretized domain and, as such, it is reasonable to approximate the integral using classical numerical quadrature methods. Unfortunately, those high-accuracy methods require a fine space discretization, which involves the introduction of a spline interpolation also for the forward function $\mu_t(x)$, as it was previously done for ψ_t and its gradient, which adds a new source of error on top of that arising from the interpolations of ψ_t and μ_t. For this reason, we also opted for particle and Monte Carlo methods to approximate the integrals, rather than using its spatial values.

5.2 Results

In this section, we will show how the three optimality conditions, namely forward, backward, and parameter update ((5.1), (5.2), (5.3) 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 vector in the corresponding orthant, while a negative orthant label vector 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 (5.1) and the backward (5.2) equations. We will consider the layer forward map $\mathcal{F}(t, x, \theta_t) = \tanh(W_t\, x + \tau_t)$ and $\theta_t = (W_t, \tau_t)$, where $W_t \in \mathbb{R}^{d \times d}$ and $\tau_t \in \mathbb{R}^d$. However, in some of the experiments reported below, we used also a forward map without shifts $\mathcal{F}(t, x, \theta_t) = \tanh(W_t x)$, so that simply $\theta_t = W_t$, where $W_t \in \mathbb{R}^{d \times d}$. The test cases for the initial distribution are the following:

- *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 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.

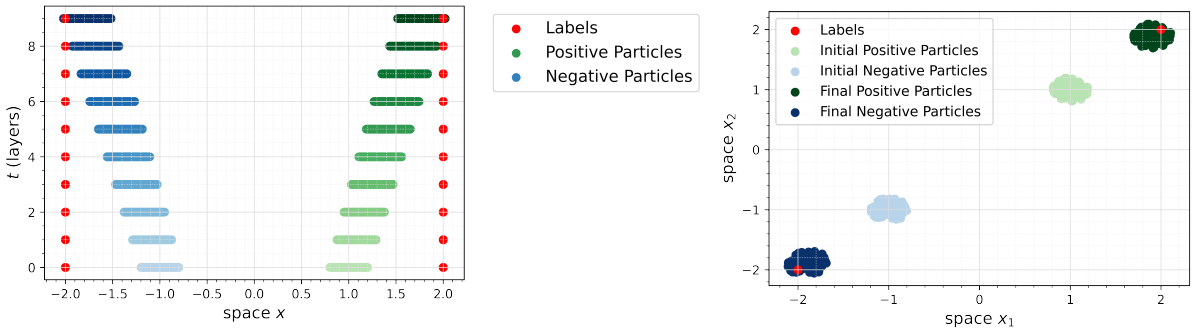

Figure 2: Left: Evolution in time of the particles from the monodimensional initial bimodal distribution μ_0 to μ_T; Right: Plot of the initial bidimensional bimodal distribution μ_0 and the final distribution μ_T.

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 (5.1) is solved using the particle method with $N = 200$ points, and the backward equation (5.2) 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 (5.6) 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 pre-

sented 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 10^{-3} in the monodimensional case, and to 10^{-4} in the bidimensional one. The reason for this will be explained below when the role of λ will be discussed. 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 nicely 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. This is more likely to happen when the "wrong value" for λ is chosen and, since it is more difficult to tune it in the bidimensional case, it is possible to see those incorrectly classified particles on the right of Figure 3.

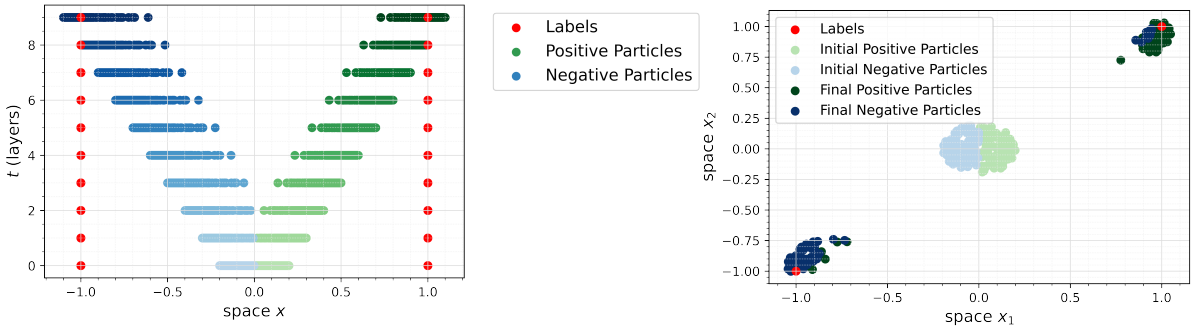

Figure 3: Left: Evolution in time of the particles from the monodimensional initial unimodal distribution μ_0 to μ_T; Right: Plot of the initial bidimensional unimodal distribution μ_0 and the final distribution μ_T.

Comparing the resolution methods for the forward equation. For the monodimensional example, it is easy to check how the various resolution methods described above perform relative to each other in solving the forward equation. As already explained, the particle and Monte Carlo methods are more similar and based on a discrete-sampling description of the dynamics. The Monte Carlo method is more sensitive than its particle counterpart, and needs many repetitions to produce a result θ_t that is stable over shooting iterations. In the first row of Figure 4, the evolution of the estimated distributions is plotted in the case of particle method, on the left, and Monte Carlo method, on the right. It is natural to expect the Monte Carlo scheme to be more diffusive, which stems from the high stochasticity of the algorithm. However, in both cases the final distribution is the one that we expect, i.e. both distributions are concentrated around the labels. The same happens in case of resolution with finite volume method, presented on the bottom left of Figure 4 . In this case the solution is not subject to the high stochasticity of the Monte Carlo method and hence it does not show as much diffusion,

but it is not as smooth as the solution obtained with particle method. This is due to the fact that the time and space discretizations are correlated and can't be chosen freely, so a relatively big space discretization needs to be chosen to compare experiments with the same number of layers (i.e. time discretization). Moreover, as illustrated by the plot on the bottom right of Figure 4, the optimal control solution θ_t does not vary significantly from an algorithm to the other. The solutions indicated in this graphics are the empirical expected values over multiple shooting iterations for every algorithm, and their standard deviation is also depicted around the lines representing the means. Clearly, the algorithm that has more variations in terms of shooting iterations is the Monte Carlo one, due to its stochasticity, while the particle method and the finite volume method are inherently sharper.

Hence, in terms of computational speed and stability over iterations (especially in the more difficult case of unimodal initial distribution μ_0), the particle method is the one that performs best, while being also the most suited one for a deep learning task, which in general implies very high-dimensional data. That being said, the experiments conducted using the Monte Carlo method and the finite volume method do allow us to confirm numerically that the shooting method based on our mean-field optimality conditions converge on the space of probability measures as expected by the theory, independently of the number of particles. Indeed, all the modelling parameters, and in particular the regularization constant $\lambda > 0$, can be chosen independently of N. Moreover, the iteration does not need any batching of the data, as it is usually done in deep learning, that is, we can take a very large number of particles (as in the Monte Carlo scheme) or a small one (as in particle method), and in both cases our algorithm will return the optimal solution θ_t for every layer $t \in [0, T]$.

Let us now focus on the particle method and for that, analyze the statistical behavior of our algorithm.

Statistical behavior. The power of our mean-field maximum principle relies on the results presented in the previous paragraph regarding the independence of all parameters from the number of particles, but also on its ability to provide a strong quantitative generalization error (4.36). This means that, if we trained our network and obtained an optimal solution θ_t^*, we have the extra advantage of knowing through (4.36) how well the latter will be able to perform on test data, i.e. when sampling new, unseen particles from μ_0. Denoting by $J^N(\theta^*)$ the empirical error as in (3.15), the generalization error consists in computing the same quantity, but with an empirical measure made by sampling new particles from μ_0 that were not used for the training phase and, possibly, a significantly larger number of them. Similarly, we can define the accuracy as the empirical probability that the output of the network will be in a small ball around the corresponding label vectors, again for all the new particles that can be resampled from μ_0. Figure 5 presents the expected *coupled descent curve* of the empirical and generalization error, and the corresponding increase of the accuracy. Since both generalization error and accuracy are measured on newly sampled test data, we perform various samplings and calculations of these quantites and report in Figure 5 their mean values and their standard deviation in form of a "cloud" of the same color. The numerical results nicely confirm the theoretically predicted phenomenon which we call *coupled descent*.

Now that the resolution methods are clarified and the resulting algorithm is understood from

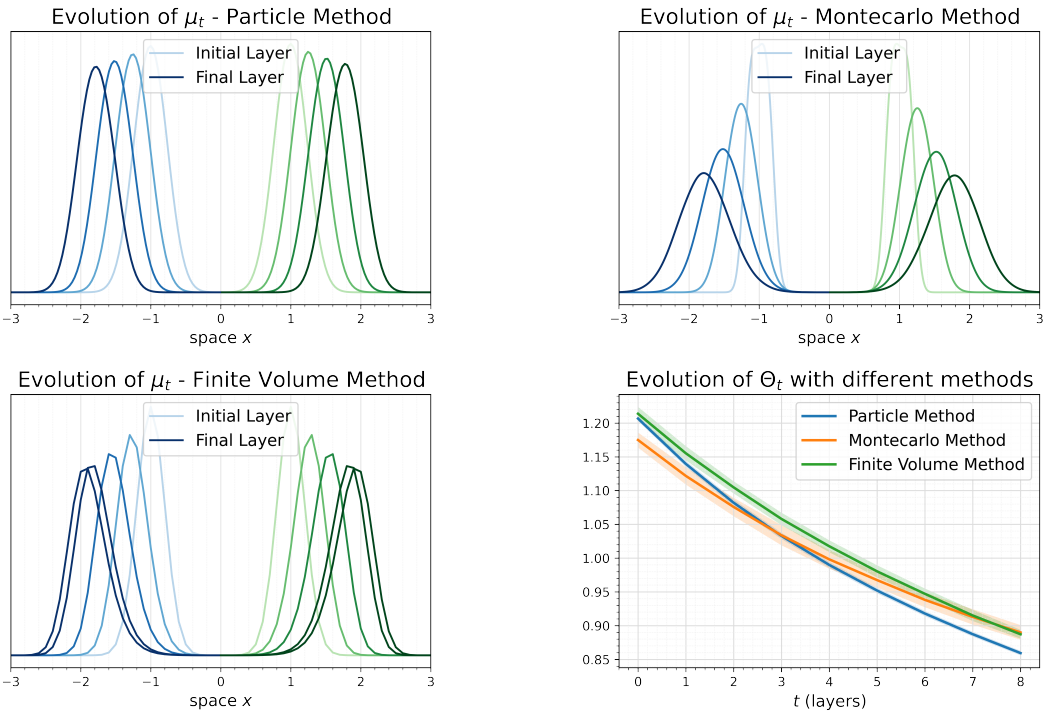

Figure 4: Top Left: evolution of the estimated μ_t obtained with particle method; Top Right: evolution of the estimated μ_t obtained with Monte Carlo method; Bottom Left: evolution of the estimated μ_t obtained with finite volume method; Bottom Right: comparison of the optimal control θ_t obtained with the three different resolution methods of the forward.

a numerical and statistical perspective, we shift our focus towards expounding the influence of the parameters that are playing a role in our method, by first considering the number of outer iterations and then the interesting role of the regularization parameter λ which acts as a learning rate. Finally, the necessary number of layers (i.e. the time discretization) may be examined in relation to the space discretization.

Contribution of the number of iterations of the shooting method. In what follows, we test how many iterations of the shooting method are necessary to obtain a good result, starting first from the initial guess $\theta_t^0 \equiv 0$, and then from the 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 towards 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 appropriate values. Hence, for both 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 6, the L^2 distance between shooting method

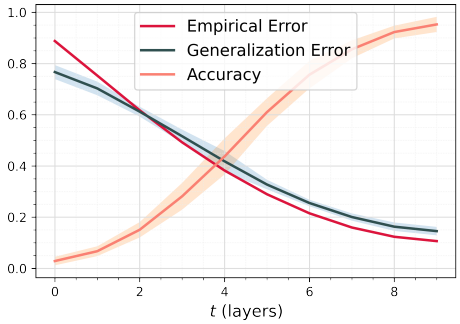

Figure 5: Statistical behaviour of the algorithm resulting from the mean-field optimality conditions.

solutions, denoted by $\varepsilon(k) = \left\|\theta_t^{k+1} - \theta_t^k\right\|_2$, is plotted for each $k = 0, ...,$number of iterations, starting from different initial guesses θ_t^0. It appears that independently of the initial guess, the distance between consecutive solutions goes to zero in a few iterations (which is also shown on the right of Figure 6 where, after the second iteration, it becomes impossible to distinguish between consecutive solutions), with different velocities depending on the initial guess.

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 the position of the corresponding label. The dynamics of the iterations is depicted in the plot on the right of Figure 6, in the one dimensional case where an initial bimodal Gaussian is fed to a network with layer forward map $\mathcal{F}(t, x, \theta_t) = \tanh(\theta_t x)$, and where the initial guess is $\theta^0 \equiv 1$.

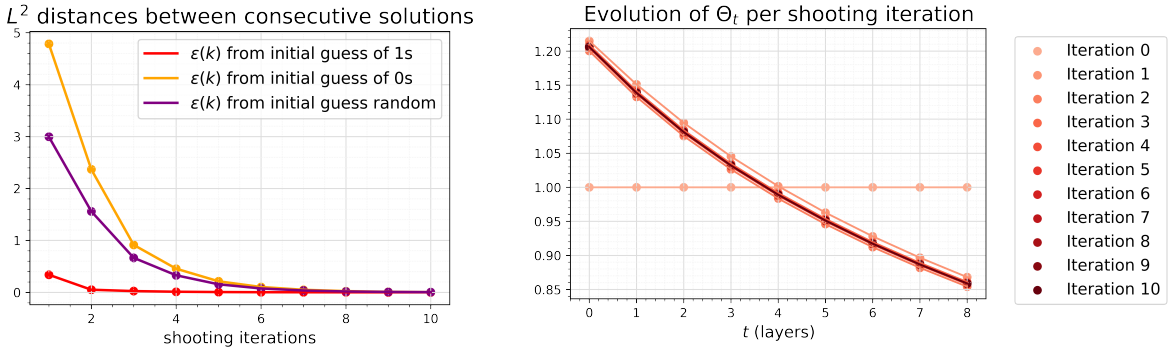

Figure 6: Left: L^2 distance between successive solutions of the shooting method over the number of iterations, and starting from different initial guess, namely $\theta_t^0 = 1$, $\theta_t^0 = 0$, and $\theta_t^0 = r_t$, $r_t \sim \mathcal{U}(0, 1)$ for all t; Right: values of θ_t over time, starting from initial guess $\theta_t^0 = 1$ for all t.

On the effect of the regularization parameter λ. A fundamental factor that has to be taken into consideration is that of the impact of the regularization parameter λ, appearing in the fixed-point equation (5.3) of the optimality conditions. The latter is a real positive number decided a priori, which determines the competing influence of the regularization term in the loss function (1.5), and hence controls 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. smaller than 0.1 in our examples –, to allow $\|\theta\|_2$ to be large enough to reach the labels, otherwise the particles will not have enough speed to get to the correct location at time $T > 0$. However, always choosing a small λ is not a good choice either, because that would destroy the convexity of the problem and lead, as we discuss below, to numerical instabilities in the learning process. Indeed, our experiments show that small values of λ may cause the mapping $f(\theta_t)$ defined in (5.6) 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 functions $f(\theta_t)$ with multiple roots, which may cause the algorithm to lose stability and to oscillate between solutions, also reflecting the loss of convexity. That being said, this issue can be overcome at the price of increasing the total number of layers of the network, as evidenced by the discussion on the role of discretization parameters detailed hereinbelow.

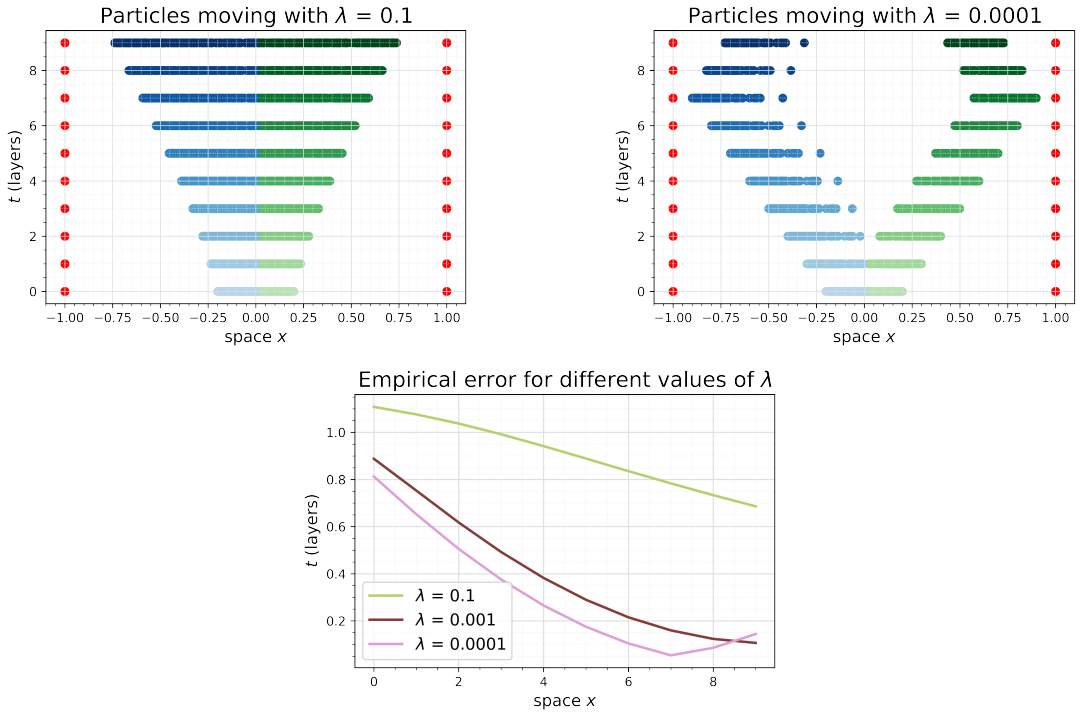

Figure 7: Top Left: unimodal initial distribution, case with $\lambda = 0.1$; Top Right: unimodal initial distribution, case with $\lambda = 0.0001$; Bottom: Resulting empirical error for different values of the learning rate λ.

Let us now look at an instructive example, in which a unimodal monodimensional Gaussian centered at the origin is fed to a neural network that has layer forward map without biases. In the left plot of Figure 3, λ is set to 0.01, leading to a correct solution. But in Figure 7, 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 7 where the particles are moving in the correct direction but not fast enough to reach the label. On the contrary, if λ is too small, the control θ_t obtained at every iteration of the shooting method

leads to an unstable and oscillating behavior between the correct result and another solution, which is shown on the right of Figure 7. 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 allow for 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 (4.3), 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, e.g. $\lambda = 10^{-3}$ in the case on the left of Figure 3). Besides, these considerations still hold in case of activation function with biases. Indeed in this case, the parameter can be split into 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 8 shows an experiments in dimension 2: 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 plots in Figure 8 is that different numbers of layers are employed, i.e., $\mathrm{d}t = 0.2$ and $\mathrm{d}t = 0.05$, 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 of 10 layers, not depicted here). These issues can clearly be overcomed 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.

Moreover, we need to keep in mind that the time step $\mathrm{d}t$ has to be chosen in accordance with the space step $\mathrm{d}x$ appearing in the backward equation as well, as the Courant number needs to be kept below 1 in order for the CFL condition to be satisfied and to guarantee the convergence and stability of the numerical scheme. It is interesting to notice that in the case of unimodal distribution, increasing the space discretization to $\mathrm{d}x = \mathrm{d}y = 0.2$ 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 either of them, depends on the function $\mathcal{F}(t, x, \theta_t)$, and since all the particles X_0^i are initially close to zero, this number tends to be too small. Hence, a better convergence rate is obtained when the space discretization is increased.

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

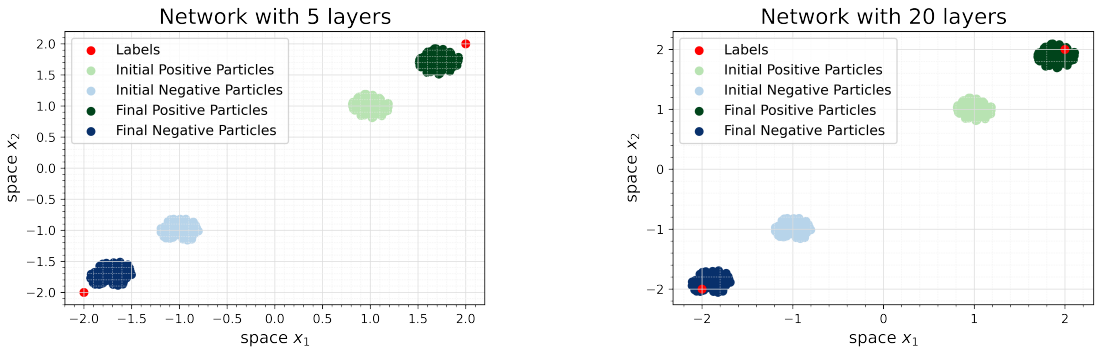

Figure 8: Left: bimodal initial distribution in 2D with dt=0.2; Right: bimodal initial distribution in 2D with dt=0.05.

fieldPMP-NeurODE-training.

Appendices

In the following series of appendices, we recollect some auxiliary results appearing earlier in the paper, and detail the proofs of some intermediate steps in our previous arguments.

A Well-posedness continuity equations and properties of characteristic flows

Proof of Theorem 2.3. In what follows, we shall study qualitative properties of the ODEs

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = \mathcal{F}(t, X_t, \theta_t) \qquad \text{and} \qquad \frac{\mathrm{d}Y_t}{\mathrm{d}t} = 0. \tag{A.1}$$

Since for any given $\theta \in L^2([0,T];\mathbb{R}^m)$ the velocity field $(t,x) \mapsto \mathcal{F}(t,x,\theta_t)$ satisfies the regularity and growth conditions of Assumption 1, it follows from standard results 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 that

$$|X_t| \leq (R + C_{\mathcal{F}} T) e^{C_{\mathcal{F}} T} \quad \text{and} \qquad Y_t = y_0\,, \tag{A.2}$$

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

$$\mathbf{\Phi}^{\theta}_{(\tau,t)} : (x_\tau, y_\tau) \in \mathbb{R}^{2d} \mapsto (X_t^{x_\tau}, Y_t^{y_\tau}) \in \mathbb{R}^{2d}, \tag{A.3}$$

where $t \in [0,T] \mapsto (X_t^{x_0}, Y_t^{y_0})$ is the unique solution of (A.1) starting from $(x_\tau, y_\tau) \in \mathbb{R}^{2d}$ at time $\tau \in [0,T]$. Given an initial datum $\mu_0 \in \mathcal{P}_c^a(\mathbb{R}^{2d})$, we can use the characteristic flow to define the following curve of measures

$$\mu_t := \mathbf{\Phi}^{\theta}_{(0,t)} \sharp \mu_0\,, \tag{A.4}$$

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

$$\int_{\mathbb{R}^{2d}} \psi(t,x,y) \mathrm{d}\mu_t(x,y) = \int_{\mathbb{R}^{2d}} \psi(t, X_t^{x_0}, Y_t^{y_0}) \mathrm{d}\mu_0(x_0,y_0)\,. \tag{A.5}$$

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

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^{2d}}\psi(t,x,y)\mathrm{d}\mu_t(x,y) &= \int_{\mathbb{R}^{2d}}\frac{\mathrm{d}}{\mathrm{d}t}\psi(t,X_t^{x_0},Y_t^{y_0})\mathrm{d}\mu_0(x_0,y_0)\\ &= \int_{\mathbb{R}^{2d}}\Big(\partial_t\psi(t,X_t^{x_0},Y_t^{y_0})+\nabla_x\psi(t,X_t^{x_0},Y_t^{y_0})\cdot\mathcal{F}(t,X_t^{x_0},\theta_t)\Big)\mathrm{d}\mu_0(x_0,y_0)\\ &= \int_{\mathbb{R}^{2d}}\Big(\partial_t\psi(t,x,y)+\nabla_x\psi(t,x,y)\cdot\mathcal{F}(t,x,\theta_t)\Big)\mathrm{d}\mu_t(x,y), \qquad (A.6)\end{aligned}$$

and an integration with respect to the time variable leads to (2.10). Furthermore, it follows e.g. from [22, Lemma 3.11] that for any $s,t\in[0,T]$, it holds

$$W_1(\mu_t,\mu_s) = W_1\big(\mathbf{\Phi}^\theta_{(0,t)}\sharp\mu_0,\mathbf{\Phi}^\theta_{(0,s)}\sharp\mu_0\big) \le \big\|\mathbf{\Phi}^\theta_{(0,t)}-\mathbf{\Phi}^\theta_{(0,s)}\big\|_{L^\infty(\mathrm{supp}(\mu_0))} \le C|t-s|\,, \qquad (A.7)$$

due to the fact that

$$|\mathbf{\Phi}^\theta_{(0,t)}(x_0,y_0)-\mathbf{\Phi}^\theta_{(0,s)}(x_0,y_0)| = |(X_t^{x_0}-X_s^{x_0},0)| \le C|t-s|, \qquad (A.8)$$

for all $(x_0,y_0)\in\mathrm{supp}(\mu_0)$, where C depends only on R, T and $C_\mathcal{F}$. Thus, the curve μ is Lipschitz continuous with respect to W_1-metric, and it is such that $\mathrm{supp}(\mu_t)\in B(R_T)$ for all $t\in[0,T]$ as a consequence of (A.2), where $R_T>0$ depends only on R, T and $C_\mathcal{F}$.

Next we prove the stability estimate. For $i=1,2$, denote by μ^i be two measure solutions of (1.7) with initial data μ_0^i. Introducing the notation $(X_t^i,Y_t^i) := \mathbf{\Phi}^\theta_{(0,t)}(x_0^i,y_0^i)$ for $t\in[0,T]$ and $(x_0^i,y_0^i)\in\mathrm{supp}(\mu_0^i)$, it holds that

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

which by applying Gronwall's Lemma then leads to

$$\big|(X_t^1,Y_t^1)-(X_t^2,Y_t^2)\big| \le \big|(x_0^1,y_0^1)-(x_0^2,y_0^2)\big|e^{\int_0^t L_\mathcal{F}(1+|\theta_s|)\,\mathrm{d}s} = |(x_0^1,y_0^1)-(x_0^2,y_0^2)|e^{L_{\mathcal{F},T,\|\theta\|_1}}, \qquad (A.9)$$

for all times $t\in[0,T]$. This provides us with the following Lipschitz estimate

$$|\mathbf{\Phi}^\theta_{(0,t)}(x_0^1,y_0^1)-\mathbf{\Phi}^\theta_{(0,t)}(x_0^2,y_0^2)| \le L_\mathcal{T}|(x_0^1,y_0^1)-(x_0^2,y_0^2)|, \qquad (A.10)$$

for all times $t\in[0,T]$, where $L_\mathcal{T} := e^{L_{\mathcal{F},T,\|\theta\|_1}}$. Given an optimal transport plan π_0 between μ_0^1 and μ_0^2, one can check that the measure $\pi := (\mathbf{\Phi}^\theta_{(0,t)}\times\mathbf{\Phi}^\theta_{(0,t)})\sharp\pi$ has marginals $\mathbf{\Phi}^\theta_{(0,t)}\sharp\mu_0^1$ and $\mathbf{\Phi}^\theta_{(0,t)}\sharp\mu_0^2$. Whence, it holds

$$\begin{aligned}W_1\Big(\mathbf{\Phi}^\theta_{(0,t)}\sharp\mu_0^1,\mathbf{\Phi}^\theta_{(0,t)}\sharp\mu_0^2\Big) &\le \int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}}|x-y|\mathrm{d}\gamma(x,y)\\ &= \int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}}\big|\mathbf{\Phi}^\theta_{(0,t)}(x_0^1,y_0^1)-\mathbf{\Phi}^\theta_{(0,t)}(x_0^2,y_0^2)\big|\mathrm{d}\pi(x_0^1,y_0^1,x_0^2,y_0^2)\\ &\le L_\mathcal{T}\int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}}\big|(x_0^1,y_0^1)-(x_0^2,y_0^2)\big|\mathrm{d}\pi(x_0^1,y_0^1,x_0^2,y_0^2)\\ &= L_\mathcal{T}\,W_1(\mu_0^1,\mu_0^2)\,,\end{aligned}$$

which leads to

$$W_1(\mu_t^1, \mu_t^2) = W_1(\mathbf{\Phi}^\theta_{(0,t)}\sharp\mu_0^1, \mathbf{\Phi}^\theta_{(0,t)}\sharp\mu_0^2) \leq L_{\mathcal{T}} W_1(\mu_0^1, \mu_0^2), \tag{A.11}$$

for all times $t \in [0,T]$, and completes the proof of Theorem 3.2. □

Proof of Proposition 4.3. We shall use the standard characteristic method with backward propagation. For any terminal condition $(X_T, Y_T) = (x,y) \in B(R_T)$, we know thanks to the classical Cauchy-Lipschitz theory that the ODEs

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = \mathcal{F}(t, X_t, \theta_t) \qquad \text{and} \qquad \frac{\mathrm{d}Y_t}{\mathrm{d}t} = 0, \tag{A.12}$$

admit a unique solution $t \in [0,T] \mapsto (X_t, Y_t) := \mathbf{\Phi}^\theta_{(T,t)}(x,y) \in \mathbb{R}^{2d}$ which can be written explicitly as

$$\mathbf{\Phi}^\theta_{(T,t)}(x,y) = \left(x - \int_t^T \mathcal{F}(s, X_s, \theta_s)\,\mathrm{d}s\,,\, y \right). \tag{A.13}$$

for all $(x,y) \in B(R_T)$. Moreover, one has that

$$\left|\mathbf{\Phi}^\theta_{(T,t)}(x,y)\right| \leq (R_T + C_{\mathcal{F}} T) e^{C_{\mathcal{F}} T} + R_T$$

by Gronwall's inequality as in (A.2), which equivalently means that $\mathbf{\Phi}^\theta_{(T,t)}(B(R_T)) \subset B(R_T)$ with $R'_T := R + (R + C_{\mathcal{F}} T) e^{C_{\mathcal{F}} T}$. Furthermore under Assumptions 1 and 2, the functions $\mathbf{\Phi}^\theta_{(T,t)} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ are $\mathcal{C}^2$ diffeomorphisms for any $t \in [0,T]$, and the application $(t,x,y) \mapsto \mathbf{\Phi}^\theta_{(T,t)}(x,y) \in \mathbb{R}^{2d}$ is locally Lipschitz.

Building on these insights, we can construct solutions of (4.7) via the standard characteristic method, by setting

$$\psi^\theta(t,x,y) := \psi_T\big(\mathbf{\Phi}^\theta_{(T,t)}(x,y)\big), \tag{A.14}$$

for all $(t,x,y) \in [0,T]\mathbb{R}^{2d}$, where $\psi_T \in \mathcal{C}^2_c(\mathbb{R}^{2d})$ satisfies (4.14). This implies that in particular that

$$\psi^\theta\big(t, \mathbf{\Phi}^\theta_{(T,t)}(x,y)\big) = \psi_T(x,y),$$

for all times $t \in [0,T]$, from whence we can deduce

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

for any $t \in [0,T)$ and $(x,y) \in \mathbb{R}^{2d}$. Since $\mathrm{supp}(\psi_T) = B(R_T)$, one has that

$$\mathrm{supp}(\psi^\theta(t)) = \mathbf{\Phi}^\theta_{(T,t)}(B(R_T)) \subset B(R'_T)$$

for all times $t \in [0,T]$. Thus, we have constructed a function $\psi^\theta(t,x,y) = \psi_T\big(\mathbf{\Phi}^\theta_{(T,t)}(x,y)\big)$ of class $\mathcal{C}^1([0,T]; \mathcal{C}^2_c(\mathbb{R}^{2d}))$ satisfying (4.7).

At this stage by considering the analytical expression (A.13), it follows from arguments similar to those leading to (A.9) that

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

which combined Assumption 2-(i), according to [74, Lemma 2.3] further implies that

$$\left\|\mathbf{\Phi}^{\theta}_{(T,t)}\right\|_{\mathcal{C}^2(\mathbf{\Phi}^{\theta}_{(T,t)}(B(R_T)))} \leq C(R'_T, T, C_\Gamma, C_{\mathcal{F}}, L_{\mathcal{F},T,C_\Gamma})\,. \tag{A.15}$$

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

$$\begin{aligned}\left\|\psi^{\theta}_t\right\|_{\mathcal{C}^2_c(\mathbb{R}^{2d})} = \left\|\psi^{\theta}_t\right\|_{\mathcal{C}^2(\mathbf{\Phi}^{\theta}_{(T,t)}(B(R_T)))} &= \left\|\psi_T(\mathbf{\Phi}^{\theta}_{(T,t)})\right\|_{\mathcal{C}^2(\mathbf{\Phi}^{\theta}_{(T,t)}(B(R_T)))} \\ &\leq C\left(\left\|\mathbf{\Phi}^{\theta}_{(T,t)}\right\|_{\mathcal{C}^2(\mathbf{\Phi}^{\theta}_{(T,t)}(B(R_T)))}\right)\|\psi_T\|_{\mathcal{C}^2(B(R_T))},\end{aligned} \tag{A.16}$$

which concludes the proof of (4.15). □

We now end this first appendix section by detailing the proof of Lemma 4.2.

Proof of Lemma 4.2. By construction of the semigroups $(\Phi^{\theta}_{(\tau,t)})_{\tau,t\in[0,T]}$, it holds for all $(t,x) \in [0,T]\times\mathbb{R}^d$ that

$$\Phi^{\theta}_{(t,T)} \circ \Phi^{\theta}_{(T,t)}(x) = x, \tag{A.17}$$

where "$\circ$" stands for the standard composition operation between functions. Thus by differentiating with respect to $x \in \mathbb{R}^d$ in (A.17), we obtain

$$\nabla_x \Phi^{\theta}_{(t,T)}(\Phi^{\theta}_{(T,t)}(x))\nabla_x \Phi^{\theta}_{(T,t)}(x) = \mathrm{Id},$$

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

$$\nabla_x \Phi^{\theta}_{(t,T)}(\Phi^{\theta}_{(T,t)}(x)) = \nabla_x \Phi^{\theta}_{(T,t)}(x)^{-1}, \tag{A.18}$$

for every $(t,x) \in [0,T]\times\mathbb{R}^d$. Differentiating with respect to $t \in [0,T]$ in (A.18) while recalling the ODE characterization derived in (3.6) for $t \in [0,T] \mapsto \nabla_x \Phi^{\theta}_{(T,t)}(x)$ then yields

$$\begin{aligned}\partial_t\Big(\nabla_x \Phi^{\theta}_{(t,T)}(\Phi^{\theta}_{(T,t)}(x))\Big) &= -\nabla_x \Phi^{\theta}_{(T,t)}(x)^{-1}\partial_t\Big(\nabla_x \Phi^{\theta}_{(T,t)}(x)\Big)\nabla_x \Phi^{\theta}_{(T,t)}(x)^{-1} \\ &= -\nabla_x \Phi^{\theta}_{(T,t)}(x)^{-1}\nabla_x \mathcal{F}\big(t, \Phi^{\theta}_{(T,t)}(x), \theta_t\big) \\ &= -\nabla_x \Phi^{\theta}_{(t,T)}(\Phi^{\theta}_{(T,t)}(x))\nabla_x \mathcal{F}\big(t, \Phi^{\theta}_{(T,t)}(x), \theta_t\big),\end{aligned}$$

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

$$\begin{cases}\partial_t\Big(\nabla_x \Phi^{\theta}_{(t,T)}(\Phi^{\theta}_{(T,t)}(x))^\top\Big) = -\nabla_x \mathcal{F}\big(t, \Phi^{\theta}_{(T,t)}(x), \theta_t\big)^\top \nabla_x \Phi^{\theta}_{(t,T)}(\Phi^{\theta}_{(T,t)}(x))^\top, \\ \quad \nabla_x \Phi^{\theta}_{(T,T)}(\Phi^{\theta}_{(T,T)}(x))^\top = \mathrm{Id},\end{cases}$$

which ends the proof of our claim. □

B Regularity of ODE flows with respect to the control variables

In this second Appendix section, we recollect somewhat elementary results concerning the regularity of the flows of diffeomorphisms $(\Phi^{\theta}_{(0,t)})_{t\in[0,T]} \subset \mathcal{C}(\mathbb{R}^d, \mathbb{R}^d)$ defined in (3.5) with respect to the control variable $\theta \in L^2([0,T], \mathbb{R}^m)$.

Proposition B.1 (Lipschitz and supremum bound for controlled flows)**.** *For any given $T > 0$, suppose that $\mathcal{F}$ satisfies Assumptions 1 and 2. Then for every $R > 0$ and any pair of control signals $\theta_1, \theta_2 \in L^2([0,T], \mathbb{R}^m)$, there exists a constant $C(T, R, \|\theta^1\|) > 0$ such that*

$$\sup_{t\in[0,T]} \big\|\Phi^{\theta^1}_{(0,t)}\big\|_{\mathcal{C}(B(R))} \leq C(T, R, \|\theta^1\|_1)$$

and

$$\sup_{t\in[0,T]} \big\|\Phi^{\theta^1}_{(0,t)} - \Phi^{\theta^2}_{(0,t)}\big\|_{\mathcal{C}(B(R))} \leq C(T, R, \|\theta^1\|_1)\|\theta^1 - \theta^2\|_2.$$

Proof. These estimates follows from our quantitative regularity assumptions together with a standard application of Grönwall's lemma. □

Proposition B.2 (Regularity of the flow with respect to the control variable)**.** *For any given $T > 0$, suppose that $\mathcal{F}$ satisfies Assumptions 1 and 2. Then for every $\theta, \vartheta \in L^2([0,T], \mathbb{R}^m)$, the following Taylor expansion*

$$\Phi^{\theta+\varepsilon\vartheta}_{(0,t)}(x) = \Phi^{\theta}_{(0,t)}(x) + \varepsilon \int_0^t \mathcal{R}^{\theta}_{(s,t)}(x)\nabla_\theta \mathcal{F}\big(t, \Phi^{\theta}_{(0,s)}(x), \theta_s\big)\vartheta_s \mathrm{d}s + o_\theta(\varepsilon), \tag{B.1}$$

holds in $\mathcal{C}([0,T] \times B(R), \mathbb{R}^{2d})$, where for each $(\tau, x) \in [0,T]\times \in \mathbb{R}^d$ the resolvent map $t \in [0,T] \mapsto \mathcal{R}^{\theta}_{(\tau,t)}(\cdot) \in \mathcal{C}^1(\mathbb{R}^d; \mathbb{R}^{d\times d})$ is the unique solution of the linearized Cauchy problem

$$\begin{cases} \partial_t \mathcal{R}^{\theta}_{(\tau,t)}(x) = \nabla_x \mathcal{F}\big(t, \Phi^{\theta}_{(0,t)}(x), \theta_t\big)\mathcal{R}^{\theta}_{(\tau,t)}(x), \\ \mathcal{R}^{\theta}_{(\tau,\tau)}(x) = \mathrm{Id}\,. \end{cases} \tag{B.2}$$

Moreover, for any $\theta^1, \theta^2 \in L^2([0,T], \mathbb{R}^m)$, there exists a constant $C'(T, R, \|\theta^1\|_1) > 0$ such that

$$\sup_{t\in[0,T]} \big\|\mathcal{R}^{\theta^1}_{(0,t)}\big\|_{\mathcal{C}(B(R),\mathbb{R}^{d\times d})} \leq C'(T, R, \|\theta^1\|_1) \tag{B.3}$$

and

$$\sup_{t\in[0,T]} \big\|\mathcal{R}^{\theta^1}_{(0,t)} - \mathcal{R}^{\theta^2}_{(0,t)}\big\|_{\mathcal{C}(B(R),\mathbb{R}^{d\times d})} \leq C'(T, R, \|\theta^1\|_1)\|\theta^1 - \theta^2\|_2. \tag{B.4}$$

In particular, the map $\theta \in L^2([0,T], \mathbb{R}^m) \mapsto \Phi^\theta \in C^0([0,T]\times B(R), \mathbb{R}^{2d})$ is Fréchet-differentiable.

Proof. By reproducing the parametrised fixed-point argument detailed in [19, Theorem 2.3.1], one can prove that the following Taylor expansion

$$\Phi^{\theta+\varepsilon\vartheta}_{(0,t)}(x) = \Phi^{\theta}_{(0,t)}(x) + \varepsilon\Psi^{\theta,\vartheta}_{(0,t)}(x) + o_\theta(\varepsilon) \tag{B.5}$$

holds for all $(t, x) \in [0,T] \times B(R)$ and each $\varepsilon > 0$, where the map $t \in [0,T] \mapsto \Psi^{\theta,\vartheta}_{(0,t)}(x) \in \mathbb{R}^d$ is the unique solution of the linearized Cauchy problem

$$\begin{cases} \partial_t \Psi^{\theta,\vartheta}_{(0,t)}(x) = \nabla_x \mathcal{F}\big(t, \Phi^{\theta}_{(0,t)}(x), \theta_t\big)\Psi^{\theta,\vartheta}_{(0,t)}(x) + \nabla_\theta \mathcal{F}\big(t, \Phi^{\theta}_{(0,t)}(x), \theta_t\big)\vartheta_t, \\ \Psi^{\theta,\vartheta}_{(0,0)}(x) = 0. \end{cases} \tag{B.6}$$

By a simple application of the constant variation formula (see e.g. [19, Theorem 2.2.3]), it can be shown that it can in fact be expressed as

$$\Psi^{\theta,\vartheta}_{(0,t)}(x) = \int_0^t \mathcal{R}^{\theta}_{(s,t)}(x)\nabla_\theta \mathcal{F}\big(s, \Phi^{\theta}_{(0,s)}(x), \theta_s\big)\vartheta_s \mathrm{d}s,$$

for all times $t \in [0,T]$, where the resolvent map $t \in [0,T] \mapsto \mathcal{R}^{\theta}_{(\tau,t)}(x) \in \mathbb{R}^{d\times d}$ is defined as in (B.2). The regularity bounds displayed in (B.3)-(B.4) easily follow by combining the regularity hypotheses of Assumption 1 and 2 with the arguments detailed in [19, Theorem 2.2.4]. □

C Proof of Theorem 4.5

In this third appendix section, we provide a proof of the abstract Lagrange multiplier rule stated in Theorem 4.5.

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

$$G'(x^*)h = 0 \qquad \text{implies} \qquad DJ(x^*)h = 0\,, \tag{C.1}$$

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

$$\Psi(\varepsilon, u) := \overline{G}(x^* + \varepsilon h + u)\,, \tag{C.2}$$

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

$$\Psi(\varepsilon, u) = \overline{G}(x^* + \varepsilon h + u) := \lim_{n\to\infty} G(x^* + \varepsilon h^n + u^n)\,. \tag{C.3}$$

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

Note that if x^* solves (4.42), one has

$$\Psi(0,0) = G(x^*) = 0\,. \tag{C.4}$$

By the definition of F-derivatives, we note that

$$\lim_{y\to 0} \frac{\|\Psi(0,y) - \Psi(0,0) - G'(x^*)y\,\|_Y}{\|y\,\|_X} = \lim_{y\to 0} \frac{\|G(x^*+y) - G(x^*) - G'(x^*)y\,\|_Y}{\|y\,\|_X} = 0. \tag{C.5}$$

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

$$\Psi'_u(0,0) = G'(x^*), \tag{C.6}$$

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

$$\Psi'_u(0,0)\omega(y) = y \qquad \text{and} \qquad \|\omega(y)\|_X \le \kappa\|y\|_Y\,, \tag{C.7}$$

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

$$f(\varepsilon, u) := \Psi'_u(0,0)u - \Psi(\varepsilon, u)\,. \tag{C.8}$$

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

$$f'_u(\varepsilon, u) = \Psi'_u(0,0) - \Psi'_u(\varepsilon, u)\,. \tag{C.9}$$

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

$$\|f(\varepsilon, u) - f(\varepsilon, v)\| \le \sup_{0\le\tau\le 1} \|f'_u(\varepsilon, u + \tau(v-u))\|\|u - v\|_X = o(1)\|\|u - v\|_X\,, \tag{C.10}$$

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

$$\|f(\varepsilon,u)\|_Y \le \|f(\varepsilon,u) - f(\varepsilon,0)\|_Y + \|f(\varepsilon,0)\|_Y \le o(1)\|u\|_X + \|f(\varepsilon,0)\|_Y\,, \tag{C.11}$$

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

$$\Psi'_u(0,0)u_{m+1} = f(\varepsilon,u_m), \quad m = 0,1,2,\cdots, \tag{C.12}$$

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 (C.10) and (C.11) that for sufficiently small ρ and r, one has

$$\|u_m\|_X \le o(1)r + o(1),\ \rho \to 0 \quad \text{and} \quad \|u_{m+2} - u_{m+1}\|_X \le \frac{1}{2}\|u_{m+1} - u_m\|_X \text{ for all } m = 0,1,\cdots, \tag{C.13}$$

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

$$u_m \to u \text{ as } m \to \infty\,. \tag{C.14}$$

Moreover we have that $\|u\|_X \le r$ and $\Psi'_u(0,0)u = f(\varepsilon,u)$ because of (C.12), and thus $\Psi(\varepsilon,u) = 0$. Lastly, we let $m \to \infty$ in

$$\|u_{m+2}\|_X \le \kappa\|f(\varepsilon,u_{m+1})\|_Y = \kappa\left\|\Psi'_u(0,0)u_{m+1} - \Psi(\varepsilon,u_{m+1})\right\|_Y\,, \tag{C.15}$$

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

$$\Psi(\varepsilon,u(\varepsilon)) = G(x^* + \varepsilon h + u(\varepsilon)) = 0 \tag{C.16}$$

and

$$\|u(\varepsilon)\|_X \le \kappa\left\|\Psi'_u(0,0)u(\varepsilon)\right\|_Y = \kappa\left\|G'(x^*)u(\varepsilon)\right\|_Y \tag{C.17}$$

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

By the definition of F- derivative, one has

$$G(x^* + k) - G(x^*) - G'(x^*)k = o(\|k\|_X), \quad k \to 0\,. \tag{C.18}$$

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

$$G(x^* + \varepsilon h + u(\varepsilon)) - G(x^*) - \varepsilon G'(x^*)h - G'(x^*)u(\varepsilon) = o(\|\varepsilon h + u(\varepsilon)\|_X), \quad \varepsilon \to 0\,. \tag{C.19}$$

Therefore

$$G'(x^*)u(\varepsilon) = o(1)\|\varepsilon h + u(\varepsilon)\|_X\,, \quad \varepsilon \to 0\,. \tag{C.20}$$

By (C.17), we obtain $\|u(\varepsilon)\|_X \le o(1)\|\varepsilon h + u(\varepsilon)\|_X$, which is

$$\|u(\varepsilon)\| = o(\varepsilon), \quad \varepsilon \to 0\,. \tag{C.21}$$

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

$$J(x^* + \varepsilon h + u(\varepsilon)) \ge J(x^*)\,, \tag{C.22}$$

which yields

$$DJ(x^*)(\varepsilon h + u(\varepsilon)) + o(\|\varepsilon h + u(\varepsilon)\|_X) \geq 0, \quad \varepsilon \to 0\,. \tag{C.23}$$

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

$$DJ(x^*)h = 0\,. \tag{C.24}$$

• *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

$$DJ(x^*) \subset [\mathcal{N}(G'(x^*))]^\perp = \left\{ x' \in \overline{X}'_E \mid \langle x', h\rangle = 0 \text{ for all } h \in \mathcal{N}(G'(x^*)) \subset \overline{X}_E \right\}. \tag{C.25}$$

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

$$[\mathcal{N}(G'(x^*))]^\perp = \mathcal{R}(G'(x^*)^\top)\,. \tag{C.26}$$

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

$$\langle J'(x^*), z\rangle = \langle G'(x^*)^\top p^*, z\rangle = \langle p^*, G'(x^*)z\rangle \quad \text{ for all } z \in \overline{X}_E\,, \tag{C.27}$$

which completes the proof of Theorem 4.5.

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

[8] 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.

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

[10] 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, 631–657.

[11] 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.

[12] B. Bonnet and F. Rossi, *Intrinsic Lipschitz Regularity of Mean-Field Optimal Controls*, SIAM Journal on Control and Optimization **59** (2021), no. 3, 2011–2046.

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

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

[15] Benoît Bonnet and Hélène Frankowska, *Necessary Optimality Conditions for Optimal Control Problems in Wasserstein Spaces*, To appear in Applied Mathematics and Optimization **84** (2021), 1281–1330.

[16] Benoît Bonnet and Hélène Frankowska, *On the Properties of the Value Function Associated to a Mean-Field Optimal Control Problem of Bolza Type*, Proceedings of the 2021 60th conference on decision and control (cdc), 2021, pp. 4558–4563.

[17] Benoît Bonnet and Hélène Frankowska, *Semiconcavity and Sensitivity Analysis in Mean-Field Optimal Control and Applications*, Journal de Mathématiques Pures et Appliquées **157** (2022), 282–345.

[18] 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.

[19] 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.

[20] H. Brézis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext, Springer, 2010.

[21] 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.

[22] 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.

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

[24] René Carmona and François Delarue, *Forward–backward stochastic differential equations and controlled McKean–Vlasov dynamics*, The Annals of Probability **43** (2015), no. 5, 2647 –2700.

[25] 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).

[26] 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. 6572–6583.

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

[28] Gianni Dal Maso, *An introduction to Γ-convergence*, Progress in Nonlinear Differential Equations and their Applications, 8, Birkhäuser Boston Inc., Boston, MA, 1993.

[29] Ingrid Daubechies, DeVore Ronal, Simon Foucart, Boris Hanin, and Petrova Guergana, *Nonlinear approximation and (deep) relu networks* (2019), available at `1905.02199`.

[30] Steffen Dereich, Michael Scheutzow, and Reik Schottstedt, *Constructive quantization: Approximation by empirical measures*, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques **49** (2013), no. 4, 1183 –1203.

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

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

[33] 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.

[34] 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).

[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] Nicolas Fournier and Arnaud Guillin, *On the rate of convergence in wasserstein distance of the empirical measure*, Probability Theory and Related Fields **162** (2015), no. 3, 707–738.

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

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

[41] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT Press, 2016. Book in preparation for MIT Press.

[42] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, *Explaining and harnessing adversarial examples*, International conference on learning representations, 2015.

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

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

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

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

[47] 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.

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

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

[50] 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.

[51] J. Kukačka, V. Golkov, and D. Cremers, *Regularization for deep learning: A taxonomy*, arXiv preprint arXiv:1710.10686 (2017), available at `1710.10686`.

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

[53] 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)).

[54] 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, 5998–6026.

[55] 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.

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

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

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

[59] 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.

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

[61] 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).

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

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

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

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

[66] 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.

[67] Shai Shalev-Shwartz and Shai Ben-David, *Understanding machine learning - from theory to algorithms.*, Cambridge University Press, 2014.

[68] 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–.

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

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

[71] 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.

[72] Jonathan Weed and Francis Bach, *Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance*, Bernoulli **25** (2019), no. 4A, 2620 –2648.

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

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

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

[76] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, *Understanding deep learning requires rethinking generalization*, International conference on learning representations, 2017.