View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by AMS Tesi di Laurea

Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Matematica

Consensus-Based Optimization on Hypersurfaces

Relatore:
Chiar.ma Prof.ssa
Valeria Simoncini

Presentata da:
Giacomo Borghi

Correlatore:
Chiar.mo Prof.
Massimo Fornasier

Sessione IV
Anno Accademico 2019-2020

Abstract

In the present work we introduce a Consensus-Based algorithm for global optimization on hypersurfaces. The method constitutes a metaheuristic optimization technique where a set of interacting particles are driven by instantaneous stochastic and deterministic decisions in order to establish a consensus among particles on the location of a global minimizer within the domain. The dynamics is represented by a system of SDEs and it is studied under the formal framework of kinetic theory for individual-based models.

First, we demonstrate the well-posedness of the system and formally derive the mean-field limit. Next, we study analytically and computationally the consensus mechanism focusing on the difficulties the constrained optimization setting entails. We conclude with computational experiments on benchmark functions.

Sommario

In questo elaborato viene presentato un algoritmo Consensus-Based per l'ottimizzazione vincolata a ipersuperfici. Il metodo consiste in una tecnica di ottimizzazione di tipo metaeuristico dove un insieme di particelle interagenti si muove secondo un meccanismo che unisce movimenti deterministici e stocastici per creare un consenso attorno ad un luogo del dominio dove è presente un minimo della funzione. La dinamica è governata da un sistema di SDE ed è studiata attraverso il formalismo della teoria cinetica per modelli di particelle interagenti.

Innanzitutto, viene dimostrato che il sistema è ben posto e viene formalmente derivato il suo limite di campo medio. Il meccanismo di consenso viene poi studiato analiticamente e computazionalmente soffermandosi sulle difficoltà che il rispetto del vincolo comporta. Infine, vengono condotti esperimenti computazionali su classiche funzioni test.

Contents

Introduction

Optimization plays an important role in several fields of science, engineering, economics, and industry. Lately, with the progress made in data analytics, developing efficient algorithms for the optimization of high dimensional functions has become a crucial problem as optimization is a key component of most of Machine Learning techniques. Indeed, the learning process often consists of finding a minimum of the so-called cost function, a typically non-convex, non-differentiable function which makes the task extremely challenging and computationally expensive.

Even though, for such problems, gradient-based algorithms have been dominating the field thanks to their low computational complexity, they are not naturally defined when applied to mixed-integer problems or to the optimization of non-differentiable functions. Moreover, they are local search algorithms and, hence, they privilege the exploitation of the current solution above the exploration of new and unknown areas in the search space.

Exploration and *exploitation* are, indeed, two contradictory strategies and a good search algorithm must find a trade-off between these two [29]. Metaheuristic is a class of alternative algorithms that implement nature-inspired heuristic methods to combine these two strategies. These algorithms are designed to find global or near optimal solutions within acceptable search time, at reasonable computational cost [31]. Keeping in mind that all optimization techniques are often biased towards a specific class of problems ("no free lunch" Theorem [43]), it is important to explore different approaches in order to get more insight into the optimization problem, for instance the landscape

of the loss function in Machine Learning, and eventually complement the conventional gradient-based algorithms.

For this purpose, a new metaheuristic gradient-free Consensus-Based Optimization (CBO) method has been introduced and studied, presenting empirical success in the optimization of high dimensional functions [8,19,28,33]. The purpose of this Thesis is to further analyze this method in the context of constrained optimization on hypersurfaces.

Consensus-Based Optimization is an optimization technique in the area of Swarm Intelligence, a class of metaheuristic algorithms that is mostly inspired by biological systems [32, 35]. These algorithms are typically made of a system of particles that are placed in the search space of some problems or functions, and each particle evaluates the objective function at its current location. Each particle, then, determines its movement through the search space by interacting with the other particles and following a specific mechanism, different for every algorithm. Eventually, the set of particles, or swarm, is likely to move close to a minimizer of the objective function. As we will discuss, the mechanism often involves random components that make, together with the high number of dependencies, the system difficult to analyze mathematically.

The Consensus-Based Optimization method represents an element of novelty in this regard. As a matter of fact, the dynamics is investigated under the framework of the individual-based models where techniques from kinetic theory are employed to study the large time behavior of the system. More specifically, the particles motion is determined by a system of N SDEs. The solution to such system is then approximated by mean-field limit by a PDE, whose solution represents the particles' density. Hence, the consensus mechanism is investigated on the continuous PDE level rather than on the discrete particle system.

In order to underline the properties and the innovation aspects of the Consensus-Based Optimization, we first present in Chapter 1 the basic concept of Swarm Intelligence Optimization and describe a well-studied method,

the Particle Swarm Optimization (PSO). Then, we introduce the central topic of the Thesis in detail; that is, the CBO method for constrained optimization on a generic hypersurface Γ.

Chapter 2 will focus on the well-posedness of the model and the derivation of the mean-field approximation for large particle limit whose proof consists of a generalization of the results contained in [16].

Furthermore, in Chapter 3, we will analyze the consensus mechanism by considering the evolution of the solution of the mean-field PDE. In particular, we will present the main techniques that have been employed to study the CBO method for unconstrained optimization and optimization on the sphere. Where possible, we will attempt to employ these techniques to generalize the results obtained in these cases. Moreover, we will discuss how the geometry of Γ can influence the consensus mechanism and the decay of the system variance.

In Chapter 4, we finally investigate, from a computational point of view, the behavior of the method on benchmark problems, specifically the optimization of the Ackley and the Rastrigin functions constrained on the three-dimensional torus.

Chapter 1

Swarm Intelligence Optimization

Notable algorithms within this class include the Ant Colony Optimization (AOC) [12], the Artificial Bee Colony optimization (ABC) [25] and the Particle Swarm Optimization (PSO) [27,35]. As the names suggest, these methods attempt to create a system of simple agents, or particles, showing an "intelligent" collective behavior capable of solving an optimization task [27, 35]. A vast number of these algorithms has been suggested in literature and the variants differ with respect to memory effects, stochasticity, time discretization and other features. In order to provide a well-studied example of this class, the next section illustrates the mechanism of PSO. This will allow us to make a comparison with the Consensus-Based Optimization methods in Section 1.2. To conclude the chapter, we introduce the main topic of the Thesis, the Consensus-Based Optimization method for constrained optimization on hypersurfaces.

1.1 Particle Swarm Optimization

The original PSO has been proposed by Kennedy et al. [26] as a method for the optimization of nonlinear functions, i.e. to solve the problem

$$\min_{v \in \mathbb{R}^d} \mathcal{E}(V)$$

given a certain objective function $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$, which we assume, without loss of generalization, to be a non-negative function.

A set of N particles is considered. Each individual of the particle swarm is described at every time t by a triplet $(V_t^i, W_t^i, P_t^i) \in \mathbb{R}^{N\times 3}$ of d-dimensional vectors of the search space $\mathbb{R}^d$. These are: the current position V_t^i, the velocity W_t^i and the previous best position P_t^i, which is defined as the location where the particle i attained its smaller value of the objective function, formally

$$P_t^i = \arg\min_{\{V_s^i : s\leq t\}} \mathcal{E}(V_t^i)\,.$$

The current position V_t^i can be considered as a set of coordinates describing a point in space. At each iteration of the algorithm, the current position is evaluated as a problem solution. If that position is better than any that has been found so far, then the coordinates are stored in the vector P_t^i. The objective is to keep finding better positions and updating P_t^i. New points are chosen by adding W_t^i coordinates to V_t^i, and the algorithm operates by adjusting the velocity W_t^i which can effectively be seen as a step size.

The particles are organized according to some sort of communication structure or topology. In view of the comparison between CBO and PSO, we consider the topology where every particle interacts with the rest of the swarm; namely, the topology of a fully-connected graph. We refer to [35] for examples where more complex topologies are taken into account. According to this topology, the global best P_t^g is defined as the optimal value between the personal best P_t^i, i.e.

$$g = \arg\min_{i=1,\dots,N} \mathcal{E}(P_t^i)\,.$$

In the PSO process, the velocity of each particle is iteratively adjusted so that the particle stochastically oscillates around the P_t^i and P_t^g locations. The PSO's system (as proposed in [37]) reads as follow:

$$\begin{cases} W_{t+1}^i = \omega W_t^i + U(\phi_1) \circ (P_t^i - V_t^i) + U(\phi_2) \circ (P_t^g - V_t^i) \\ V_{t+1}^i = V_t^i + W_{t+1}^i \,, \end{cases} \tag{1.1}$$

where ω is the "inertia weight", $U(\phi)$ is a random variable which is uniformly distributed on $[0, \phi]$, $\phi > 0$ and $\circ$ is the Hadamard product.

The parameters ϕ_1 and ϕ_2 determine the magnitude of the random forces in the direction of personal best P_t^i and global best P_t^g. Moreover, these forces depend on $|P_t^i - V_t^i|$ and $|P_t^g - V_t^i|$ and, therefore, the step size of a particle i is large if V_t^i is far from the personal best P_t^i or the global best P_t^g. The inertia weight ω is capable of regulating the exploration behavior, but large values of ω may make the swarm unstable. We refer to [35] for a complete discussion about the role of the parameters in the PSO dynamics.

In the update equations (1.1) we can recognize the two main features of the particles behavior in Swarm Intelligence algorithms:

1. particles share knowledge in order to move towards regions of the objective function domain where a minimizer is likely to be found;

2. a stochastic component is introduced in the step choice to partially explore the search space independently of the knowledge of the system.

Despite of its usefulness, a rigorous convergence analysis of such swarm intelligence algorithms is often missing: the high number of dependencies and the random components make the asymptotic analysis of these mechanisms extremely hard, especially when long-term dependencies through memory mechanisms are encoded.

In order to overcome this deficiency, the CBO method was first proposed in [33]. At the expense of a simpler metaheuristic mechanism with respect to PSO, CBO implements these features also allowing for a rigorous asymptotic analysis in the framework of statistical physics.

1.2 Consensus-Based Optimization

Individual-based models have been widely used in the investigation of complex systems that manifest self-organization or collective behavior. Examples of such complex systems include the already-mentioned swarming behavior, but also crowd dynamics, opinion formation, synchronization, and many more, that are present in the field of mathematical biology, ecology and social dynamics, see for instance [1, 4, 6, 7, 39].

CBO has been introduced in [33] and consists of a stochastic Swarm Intelligence algorithm that bears a particularly strong resemblance to opinion dynamics. In general, opinion dynamics within an interacting population can lead to either consensus, polarization or even fragmentation. A thorough understanding of such phenomena would, initially, require the formulation of mathematical models which describe the evolution of opinions in the population under investigation. CBO can be considered as one of these models, the stochastic Kuramoto-Vicsek model, introduced in [41] to study the cooperative behavior of animals.

In the context of global optimization, the model focuses on instantaneous stochastic and deterministic decisions in order to establish a consensus among particles or agents, on the location of a global minimizers within the domain. The particles are described only by their current position V_t^i at the time t. Thanks to the instantaneous nature of the dynamics, the evolution can be interpreted as a system of first-order stochastic differential equations (SDEs) defined as:

$$dV_t^i = -\lambda(V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N)) + \sigma|V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N)|dB_t^i\,, \tag{1.2}$$

where $v_\alpha^{\mathcal{E}}$ is the weighted average

$$v_{\alpha,\mathcal{E}}(\rho_t^N) = \frac{\sum_{j=1}^N V_t^j e^{-\alpha\mathcal{E}(V_t^j)}}{\sum_{j=1}^N e^{-\alpha\mathcal{E}(V_t^j)}}\,. \tag{1.3}$$

The positional change of a particle is given by two components. The

first component is an attraction component towards $v_{\alpha,\mathcal{E}}$ whose magnitude is given by the distance $|V_t^i - v_{\alpha,\mathcal{E}}|$ and a drift parameter λ. We note that $v_{\alpha,\mathcal{E}}$ is weighted with respect to the Gibbs distribution corresponding to $\mathcal{E}$; hence, it promotes the consensus among regions of the domain where $\mathcal{E}$ attains the minimum value. We will discuss in detail the reason for such distribution in the next section.

The second component is a random search term, which is modeled as independent Brownian motions σB_t^i with a uniform diffusion parameter σ. The individual variances are scaled with the distance towards $v_{\alpha,\mathcal{E}}$, enabling them to explore their current area, while agents near $v_{\alpha,\mathcal{E}}$ display no randomness, emphasizing their current position.

Even though we recognize the two main features of Swarm Intelligence Optimization in the system (1.2), knowledge sharing and random exploration, CBO presents important differences compared to PSO. Namely,

- $\rightarrow$ the dynamics is defined for every $t \in \mathbb{R}_{\geq 0}$;
- $\rightarrow$ the model does not use the evaluation of the global best, $\arg\min_{i=1,\cdot,N;s\leq t} \mathcal{E}(V_s^i)$, or personal best, $\arg\min_{s\leq t} \mathcal{E}(V_s^i)$ by employing the weighted average (1.3);
- $\rightarrow$ the model neglects memory effects and the inertia of the particles.

These characteristics certainly make the metaheuristic mechanism much simpler with respect to PSO, but this representation in the context of individual-based models allows us to perform a rigorous mathematical analysis of the method convergence. Indeed, the dynamics is approximated by its mean-field mono-particle process whose distribution is the solution of the corresponding Fokker-Planck equation. The aim is to acquire a deeper understanding of the performance of the particle-based algorithm through the mean-field perspective, especially regarding convergence properties. In the next section, we will present the mean-field process and the mean-field PDE for the CBO method on hypersurfaces in detail.

It is worth noting that, despite the simplicity of the mechanism, the algorithm seems to be powerful and robust enough to tackle many interesting non-convex optimizations of practical relevance, showing great scalability even for $d \gg 1$. We refer to [9, 17] for examples where the algorithm has been successfully employed to solve high dimensional tasks such as the Robust Subspace Detection problem and the training of a two-layer Neural Network. We refer to [40], instead, for a comparison between CBO and PSO on the optimization of benchmark functions.

1.3 CBO method on hypersurfaces

We now present the main topic of the Thesis, a new CBO method designed for global optimization on hypersurfaces. The setting of constrained optimization is motivated by the fact that several applications in Machine Learning can be seen as a constrained optimization task on manifolds [16,17]. Thus, the analysis of CBO on hypersurfaces should be seen as a first step towards the analysis of a wider class of CBO methods for constrained optimization on manifolds.

The metaheuristic technique we present, thus, attempts to solve the following constrained optimization problem

$$v^* \in \operatorname*{argmin}_{v \in \Gamma} \mathcal{E}(v)\,, \tag{1.4}$$

where $0 \leq \mathcal{E} : \mathbb{R}^d \to \mathbb{R}$ is a given continuous cost function, which we wish to minimize over a hypersurface Γ. The settings of Γ, as used in [11], are the following:

Definition 1.1. *Γ is a connected $\mathcal{C}^2$ compact hypersurface embedded in $\mathbb{R}^d$, which is represented as the 0-level set of a signed distance function γ with $|\gamma(v)| = dist(v, \Gamma)$. This means that:*

$$\Gamma = \left\{ v \in \mathbb{R}^d \middle|\ \gamma(v) = 0 \right\}\,.$$

The gradient $\nabla\gamma$ *is, then, the outward unit normal on* Γ *where* γ *is defined,*

$$|\nabla\gamma(v)| = 1 \quad \forall\, v \in \Gamma, \quad \textit{while} \quad P(v) = I - \nabla\gamma(v)\nabla\gamma(v)^t$$

is the linear projection operator $P(\cdot)$*. Both the operator norm* $\|.\|_2$ *of the Hessian matrix, defined as* $\|A\|_2 := \sup_{v\in\mathbb{R}^m} |Av|/|v|$ *for* $A \in \mathbb{R}^{n\times m}$*, and the* L^2*-norm of the Laplacian will also be bounded by a constant* c_γ*:*

$$\|\nabla^2\gamma(v)\|_2, |\Delta\gamma(v)| \leq c_\gamma \quad \forall v \ \in \Gamma$$

where c_γ *could, in general, depend on the dimension* d*.*

Moreover, we assume that there exists an open neighborhood $\overline{\Gamma}$ *of* Γ *such that* $\gamma \in \mathcal{C}^3(\overline{\Gamma})$ *and that, if* $\partial\Gamma = \emptyset$*, then* $\gamma < 0$ *in the interior of* Γ *and* $\gamma > 0$ *at the exterior.*

Example 1.1. *Examples of hypersurfaces* Γ *in this setting are*

- *the unit sphere* $\mathbb{S}^{d-1}$*, in which case* $\gamma(v) = |v| - 1$*,* $\nabla\gamma(v) = \frac{v}{|v|}$ *and* $\Delta\gamma(v) = \frac{d-1}{|v|}$*;*
- *a torus radially symmetric about the* v^d*-axis and of inner radius* $r > 0$ *and external radius* $R > 0$ *that is expressed in Cartesian coordinates as the* 0*-level set of the signed distance function*

$$\gamma(v) = \sqrt{(\sqrt{|v|^2 - (v^d)^2} - R)^2 + (v^d)^2} - r,$$

where $v = (v^1, \ldots, v^d)$*.*

The system constituting the method is a system of N interacting particles $((V_t^i)_{t\geq 0})_{i=1,\ldots,N}$ satisfying the following stochastic differential dynamics expressed in Itô's form

$$dV_t^i = -\lambda P(V_t^i)(V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N))dt + \sigma|V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N)|P(V_t^i)dB_t^i \\ - \frac{\sigma^2}{2}(V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N))^2\Delta\gamma(V_t^i)\nabla\gamma(V_t^i)dt\,, \quad (1.5)$$

where $\lambda > 0$ is a suitable drift parameter, $\sigma > 0$ a diffusion parameter,

$$\rho_t^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{V_t^i} \tag{1.6}$$

is the empirical measure of the particles (δ_v is the Dirac measure at $v \in \mathbb{R}^d$), while

$$v_{\alpha,\mathcal{E}}(\rho_t^N) = \frac{\sum_{j=1}^N V_t^j e^{-\alpha\mathcal{E}(V_t^j)}}{\sum_{j=1}^N e^{-\alpha\mathcal{E}(V_t^j)}} = \frac{\int_{\mathbb{R}^d} v\omega_\alpha^{\mathcal{E}}(v)d\rho_t^N}{\int_{\mathbb{R}^d}\omega_\alpha^{\mathcal{E}}(v)d\rho_t^N} \quad \text{with} \quad \omega_\alpha^{\mathcal{E}}(v) := e^{-\alpha\mathcal{E}(v)}\,. \tag{1.7}$$

This stochastic system is considered complemented with independent and identically distributed (i.i.d.) initial data $V_0^i \in \Gamma$ with $i = 1, \cdots, N$, and the common law is denoted by $\rho_0 \in \mathcal{P}(\Gamma)$. The trajectories $((B_t^i)_{t\geq 0})_{i=1,\dots N}$ denote N independent standard Brownian motions in $\mathbb{R}^d$.

As already mentioned, $e^{-\alpha\mathcal{E}(v)}$ is the Gibbs distribution corresponding to $\mathcal{E}(v)$. This choice comes from the well-known Laplace principle [10, 30, 34], a classical asymptotic method for integrals, which states that for any probability measure $\rho \in \mathcal{P}(\Gamma)$, it holds

$$\lim_{\alpha\to\infty}\left(-\frac{1}{\alpha}\log\left(\int_\Gamma e^{-\alpha\mathcal{E}(v)}d\rho(v)\right)\right) = \inf_{v\in \text{supp }(\rho)} \mathcal{E}(v)\,. \tag{1.8}$$

The right-hand side of equation (1.5) is made of three terms, all of which play a different role in the mechanism of the dynamics. The first deterministic term $-\lambda P(V_t^i)(V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N))dt$ imposes a drift to the dynamics towards $v_{\alpha,\mathcal{E}}$, which is the current consensus point at time t as approximation to the global minimizer. The second stochastic term $\sigma|V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N)|P(V_t^i)dB_t^i$ introduces a random decision to favor exploration, whose variance is a function of the distance of the particles to the current consensus points. The last term $-\frac{\sigma^2}{2}(V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N))^2\Delta\gamma(V_t^i)\nabla\gamma(V_t^i)dt$ combined with $P(\cdot)$ is needed to ensure that the dynamics stays on the hypersurface despite the Brownian motion component.

We further notice that the dynamics does not make use of any derivative

of $\mathcal{E}$, but only of its pointwise evaluations. We will require below regularity assumptions on $\mathcal{E}$ exclusively to ensure formal well-posedness of the evolution.

Through the same approach used in [16], in Chapter 2 we will show the well-posedness of (1.5) and its rigorous mean-field limit to the following nonlocal, nonlinear Fokker-Planck equation

$$\partial_t \rho_t = \lambda \nabla_\Gamma \cdot \big((P(v)(v - v_{\alpha,\mathcal{E}}(\rho_t)))\rho_t \big) + \frac{\sigma^2}{2} \Delta_\Gamma (|v - v_{\alpha,\mathcal{E}}(\rho_t)|^2 \rho_t), \quad t > 0,\ v \in \Gamma\,, \tag{1.9}$$

with the initial data $\rho_0 \in \mathcal{P}(\Gamma)$ and where $\rho = \rho(t, v) \in \mathcal{P}(\Gamma)$ is a Borel probability measure on Γ, while the operators $\nabla_\Gamma \cdot$ and Δ_Γ denote the divergence and Laplace-Beltrami operator on the hypersurface Γ, respectively.

The mean-field limit will be achieved through the coupling method [5, 15, 22, 38] by introducing an auxiliary mono-particle process, satisfying the self-consistent nonlinear SDE

$$\begin{aligned} d\overline{V}_t = &-\lambda P(\overline{V}_t)(V_t - v_{\alpha,\mathcal{E}}(\rho_t))dt + \sigma |\overline{V}_t - v_{\alpha,\mathcal{E}}(\rho_t)| P(\overline{V}_t) dB_t \\ &- \frac{\sigma^2}{2} (\overline{V}_t - v_{\alpha,\mathcal{E}}(\rho_t))^2 \Delta\gamma(\overline{V}_t) \nabla\gamma(\overline{V}_t) dt\,, \end{aligned} \tag{1.10}$$

with the initial data $\overline{V}_0$ distributed according to $\rho_0 \in \mathcal{P}(\Gamma)$. Here, we require ρ to be the law of the random process $(V_t)_{t\geq 0}$, $\rho_t = \text{law}(V_t)$. Formally, $(V_t)_{t\geq 0}$ is considered to be a continuous stochastic process on the probability space $(\Omega, \mathcal{F}, \overline{P})$ and it induces a function

$$\begin{aligned} \Phi_V : \Omega &\to \mathbb{R}_{\geq 0} \times \mathbb{R}^d \\ (\Phi_V(\omega))(t) &:= V_t(\omega). \end{aligned}$$

The law ρ of the process is, then, defined as the pushforward measure:

$$\rho := (\overline{P}) \circ \Phi_V^{-1}\,.$$

In the next chapter, we will show that $\rho(t, \cdot)$, as a measure on Γ, solves the PDE (1.9).

We call the SDE (1.10) "mean-field dynamics", and the PDE (1.9) "mean-field PDE".

Chapter 2

Well-Posedness and Mean-Field Limit

In this chapter, we will focus on the CBO method designed for constrained optimization on a hypersurface Γ of the objective function $\mathcal{E}$. In particular, we analyze the well-posedness of the equations involved and derive the rigorous mean-field limit. We remark that the rigorous derivation of the mean-field limit is an open issue for unconstrained CBO [8], due to the difficulties in establishing bounds on the moments of the particles probability distribution. This theoretical gap, indeed, was one of the reasons why the CBO method for constrained optimization on the sphere $\mathbb{S}^{d-1}$ was first introduced in [16, 17] and for which the mean-field limit can be rigorously proven. Following the approach of [16], we will generalize these results to generic hypersurfaces Γ.

In order to do so, we require the following smoothness assumption on $\mathcal{E}$ throughout the chapter:

Assumption 2.1. *The objective function $0 \leq \mathcal{E} : \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz continuous.*

We recall the model comprises a system of N interacting particles $((V_t^i)_{t\geq 0})_{i=1,\dots,N}$ satisfying the following stochastic differential dynamics expressed in Itô's

form

$$dV_t^i = -\lambda P(V_t^i)\left(V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N)\right)dt + \sigma|V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N)|P(V_t^i)dB_t^i \\ - \frac{\sigma^2}{2}(V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N))^2\Delta\gamma(V_t^i)\nabla\gamma(V_t^i)dt\,, \quad (2.1)$$

where $\lambda > 0$ is a suitable drift parameter, $\sigma > 0$ a diffusion parameter, ρ_t^N is the empirical measure of the particles, and

$$v_{\alpha,\mathcal{E}}(\rho_t^N) = \frac{\int_{\mathbb{R}^d} v\omega_\alpha^{\mathcal{E}}(v)d\rho_t^N}{\int_{\mathbb{R}^d}\omega_\alpha^{\mathcal{E}}(v)d\rho_t^N} \quad \text{with} \quad \omega_\alpha^{\mathcal{E}}(v) := e^{-\alpha\mathcal{E}(v)}\,. \quad (2.2)$$

The mean-field limit of (2.1) is the mean-field PDE

$$\partial_t\rho_t = \lambda\nabla_\Gamma\cdot((P(v)(v - v_{\alpha,\mathcal{E}}(\rho_t))\rho_t) + \frac{\sigma^2}{2}\Delta_\Gamma(|v - v_{\alpha,\mathcal{E}}(\rho_t)|^2\rho_t), \quad t > 0,\ v \in \Gamma\,, \quad (2.3)$$

with the initial data $\rho_0 \in \mathcal{P}(\Gamma)$. Here $\rho = \rho(t,v) \in \mathcal{P}(\Gamma)$ is a Borel probability measure on Γ and $v_{\alpha,\mathcal{E}}(\rho_t)$ is defined as in equation (2.2).

The mean-field dynamics is the following self-consistent nonlinear SDE

$$d\overline{V}_t = -\lambda P(\overline{V}_t)(\overline{V}_t - v_{\alpha,\mathcal{E}}(\rho_t)dt + \sigma|\overline{V}_t - v_{\alpha,\mathcal{E}}(\rho_t))|P(\overline{V}_t)dB_t \\ - \frac{\sigma^2}{2}(\overline{V}_t - v_{\alpha,\mathcal{E}}(\rho_t))^2\Delta\gamma(\overline{V}_t)\nabla\gamma(\overline{V}_t)dt\,, \quad (2.4)$$

with the initial data $\overline{V}_0$ distributed according to $\rho_0 \in \mathcal{P}(\Gamma)$ and $\rho_t = \text{law}(\overline{V}_t)$.

As already mentioned, the results we present in this chapter can be considered as a generalization of the analysis in [16] for the CBO method for the constrained optimization on the sphere where $\Gamma = \mathbb{S}^{d-1}$. For this reason, only the idea of the proof will be included and we will focus on the differences led by the general setting.

We will start from the well-posedness results for the particle system (2.1) in Section 2.1. Then, we will show the well-posedness of the mean-field dynamics (2.4) and the mean-field PDE (2.3) in Sections 2.2 and 2.3 respectively. Finally, we conclude the chapter by proving the mean-field limit in

Section 2.4. For this purpose, we first recall a crucial tool for our analysis, the Itô's formula.

Theorem 2.1 (Multidimensional Itô's formula)**.** *Let*

$$dX_t = \boldsymbol{u}(t)dt + \boldsymbol{v}(t)dB_t \tag{2.5}$$

be a d-dimensional Itô process, where $X_t \in \mathbb{R}^d$, $\boldsymbol{u}(t) \in \mathbb{R}^d$, $\boldsymbol{v}(t) \in \mathbb{R}^{d\times d'}$, *and* $B_t = (B_{t,1}, \dots, B_{t,d'})$ *is a* d'*-dimensional Brownian motion.*

Assume $\varphi(x)$ *to be a* $\mathcal{C}^2$ *map from* $\mathbb{R}^d$ *to* $\mathbb{R}$*, then it holds*

$$\varphi(X_t) = \varphi(X_0) + \int_0^t \left(\nabla\varphi(X_s)\cdot \boldsymbol{u}(s) + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2\varphi}{\partial x_i \partial x_j}(X_s) v_i(s) v_j(s)^t \right) ds + \int_0^t \nabla\varphi(X_s)\cdot \boldsymbol{v}(s) dB_s \tag{2.6}$$

with $v_i(s)$ *being the i-th row of the matrix* $\boldsymbol{v}(s)$*.*

Remark 2.1. *Equation* (2.5) *should be read in integral form;* $\int \boldsymbol{v}(t)dB_t$ *is a d-dimensional vector whose components are defined as:*

$$\sum_{j=1}^{d'} \int v_{k,j}(t) dB_{t,j} \quad \forall\, k = 1, \dots, d\,,$$

where $v_{k,j}(t) = (\boldsymbol{v}(t))_{k,j}$*.*

2.1 Well-posedness for the interacting particle system

Similarly to [16], we note that the system is embedded in $\mathbb{R}^d$ instead of being defined on the hypersurface Γ directly. This setting has been chosen because it provides an explicit and computable representation of the system and it allows for a global description.

The difficulty in showing the well-posedness of (2.1) in the ambient space $\mathbb{R}^d$ is that the projection $P(V_t^i)$, $\Delta\gamma(V_t^i)$ and $\nabla\gamma(V_t^i)$ may not be well defined outside Γ. In the case of the sphere $\mathbb{S}^{d-1}$, this complication appears only on one single point, the origin. Indeed, when $V_t^i = 0$

$$P(V_t^i) = I - \frac{V_t^i (V_t^i)^t}{|V_t^i|^2} \quad \text{and} \quad \Delta\gamma(V_t^i)\nabla\gamma(V_t^i) = (d-1)\frac{V_t^i}{|V_t^i|^2}$$

are not defined. By simple computations, it is also possible to show that, when γ defines the torus $\mathbb{T}^{d-1}$, the gradient $\nabla\gamma(v)$ is not well-defined for $v \in \{v \in \mathbb{R}^d \,|\, |v|^2 - v_d^2 = R^2 \wedge v_d = 0\} \cup \{0\}$.

For a general hypersurface Γ, we consider the neighborhood $\overline{\Gamma}$ of Γ in which $\gamma \in \mathcal{C}^3(\overline{\Gamma})$ and, in order to overcome this problem, we regularize the diffusion and drift coefficients outside $\overline{\Gamma}$.

We replace them with some appropriate functions P_1, P_2 and P_3 respectively: let P_1 be a $d \times d$ matrix valued map on $\mathbb{R}^d$ with bounded derivatives such that $P_1(v) = P(v)$ for all $v \in \overline{\Gamma}$, P_2 be a $\mathbb{R}^d$ valued map on $\mathbb{R}$ with bounded derivatives such that $P_2(v) = \Delta\gamma(v)$ if $v \in \overline{\Gamma}$, and P_3 be a $\mathbb{R}^d$ valued map on $\mathbb{R}^d$, again with bounded derivatives such that $P_3(v) = \nabla\gamma(v)$ if $v \in \overline{\Gamma}$.

It is also useful to mention that, since for $v \in \Gamma$

$$\begin{aligned} P(v)\nabla\gamma(v) &= (I - \nabla\gamma(v)\nabla\gamma(v)^t)\nabla\gamma(v) && (2.7)\\ &= \nabla\gamma(v) - \nabla\gamma(v)|\nabla\gamma(v)|^2 = 0 && (2.8) \end{aligned}$$

it holds for any $y \in \mathbb{R}^d$

$$\nabla\gamma(v)^t P(v) y = 0. \tag{2.9}$$

Additionally, we further regularize the locally Lipschitz function $\mathcal{E}$: let us introduce $\tilde{\mathcal{E}}(v)$ satisfying the following assumptions.

Assumption 2.2. *The regularized extension function $\tilde{\mathcal{E}} : \mathbb{R}^d \to \mathbb{R}$ is globally Lipschitz continuous and satisfies the properties*

1. $\tilde{\mathcal{E}}(v) = \mathcal{E}(v)$ *when* $v \in \overline{\Gamma}$;

2. $\tilde{\mathcal{E}}(v) - \tilde{\mathcal{E}}(u) \leq L|v-u|$ *for all* $u, v \in \mathbb{R}^d$ *for a suitable global Lipschitz constant* $L > 0$;

3. $-\infty < \underline{\tilde{\mathcal{E}}} := \inf \tilde{\mathcal{E}} \leq \tilde{\mathcal{E}} \leq \sup \tilde{\mathcal{E}} =: \overline{\tilde{\mathcal{E}}} < +\infty$.

We stress the fact that $\tilde{\mathcal{E}}$ is introduced as an auxiliary function for the proof of well-posedness and mean-field limit only, and it does not play any role in the actual optimization problem, which is defined on Γ. Indeed, as we can see in Theorem 2.3 below, particles stay on the hypersurface Γ all the time, which means that certainly $v \in \overline{\Gamma}$, so one has $\tilde{\mathcal{E}}(v) \equiv \mathcal{E}(v)$. From this point on, $\mathcal{E}$ and $\tilde{\mathcal{E}}$ are always expected to satisfy Assumptions 2.1 and 2.2.

Given such P_1, P_2, P_3 and $\tilde{\mathcal{E}}$, we introduce the following regularized particle system

$$dV_t^i = -\lambda P_1(V_t^i)(V_t^i - v_{\alpha,\tilde{\mathcal{E}}}(\rho_t^N))dt + \sigma|V_t^i - v_{\alpha,\tilde{\mathcal{E}}}(\rho_t^N)|P_1(V_t^i)dB_t^i \\ - \frac{\sigma^2}{2}(V_t^i - v_{\alpha,\tilde{\mathcal{E}}}(\rho_t^N))^2 P_2(V_t^i)P_3(V_t^i)dt\,, \quad (2.10)$$

for $i \in \{1, \dots, N\} =: [N]$, where

$$v_{\alpha,\tilde{\mathcal{E}}}(\rho_t^N) = \frac{\int_{\mathbb{R}^d} v\omega_\alpha^{\tilde{\mathcal{E}}}(v)d\rho_t^N}{\int_{\mathbb{R}^d} \omega_\alpha^{\tilde{\mathcal{E}}}(v)d\rho_t^N}\,, \qquad \omega_\alpha^{\tilde{\mathcal{E}}}(v) = e^{-\alpha\tilde{\mathcal{E}}(v)}\,. \quad (2.11)$$

We, thus, study the existence of a unique process $(\mathbf{V}_t^N)_t$ with $\mathbf{V}^N := (V^{1,N}, \dots, V^{N,N})^t \in \mathbb{R}^{Nd}$ satisfying the regularized particle system (2.10) which we can rewrite, for an arbitrary but fixed $N \in \mathbb{N}$, as

$$d\mathbf{V}_t^N = -\mathbf{F}_n(\mathbf{V}_t^N)dt + \mathbf{M}_n(\mathbf{V}_t^N)d\mathbf{B}_t^N, \quad (2.12)$$

where $\mathbf{B} = (B^{1,N}, \dots, B^{N,N})^t$ is the standard Wiener process in $\mathbb{R}^{Nd}$, and

$$\begin{aligned}
\mathbf{F}_N &= (F_N^1(\mathbf{V}), \dots, F_N^N(\mathbf{V}))^t \in \mathbb{R}^{Nd}\,, \\
F_N^i(\mathbf{V}) &= \lambda P_1(V^i)(V^i - v_{\alpha,\tilde{\mathcal{E}}}(\rho^N)) + \frac{\sigma^2}{2}(V^i - v_{\alpha,\tilde{\mathcal{E}}}(\rho^N))^2 P_2(V^i) P_3(V^i)\,, \\
\mathbf{M}_N &= \mathrm{diag}(M_N^1(\mathbf{V}), \dots, M_N^N(\mathbf{V}))^t \in \mathbb{R}^{Nd\times Nd}\,, \\
M_N^i(\mathbf{V}) &= \sigma|V^i - v_{\alpha,\tilde{\mathcal{E}}}(\rho^N)|P_1(V^i).
\end{aligned}$$

We will show that (2.12), and consequently (2.10), admits a pathwise strong solution by employing the following standard SDE well-posedness result [13, Chap. 5, Theorem 3.1]:

Theorem 2.2. *If $\mathbf{F}_N$ and $\mathbf{M}_N$ are locally Lipschitz continuous and have linear growth,* (2.12) *admits a pathwise unique local strong solution.*

Thanks to the regularity Assumption 2.2, we can apply Theorem 2.2 for a fixed N. More precisely, the following result allows us to check the Lipschitz continuity.

Lemma 2.1. *Let $N \in \mathbb{N}$, $\alpha > 0$ be arbitrary and $\tilde{\mathcal{E}}$ satisfy Assumption 2.2. Then for any $\mathbf{V}^N, \widehat{\mathbf{V}}^N \in \mathbb{R}^{Nd}$, and corresponding empirical measures $\rho^N = \frac{1}{N}\sum_{i=1}^N \delta_{V^i}$, and $\widehat{\rho}^N = \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{V}^i}$, it holds*

$$|v_{\alpha,\tilde{\mathcal{E}}}(\rho^N)| \le \frac{1}{N} C_{\alpha,\tilde{\mathcal{E}}} \|\mathbf{V}^N\|_1 \tag{2.13}$$

and

$$|v_{\alpha,\tilde{\mathcal{E}}}(\widehat{\rho}^N) - v_{\alpha,\tilde{\mathcal{E}}}(\rho^N)| \le \left(\frac{C_{\alpha,\tilde{\mathcal{E}}}}{N} + \frac{2\alpha L C_{\alpha,\tilde{\mathcal{E}}}}{N}\|\widehat{\mathbf{V}}^N\|_\infty\right)\|\mathbf{V}^N - \widehat{\mathbf{V}}^N\|_1\,, \tag{2.14}$$

where $C_{\alpha,\tilde{\mathcal{E}}} = e^{\alpha(\overline{\tilde{\mathcal{E}}} - \underline{\tilde{\mathcal{E}}})}$. Here we used the notations for norms of vectors $\|\mathbf{V}\|_\infty = \max_{i\in[N]}|V^i|$ and $\|\mathbf{V}\|_1 = \sum_{i=1}^N |V^i|$.

Idea of the proof. The boundedness of $\tilde{\mathcal{E}}$ plays a key role in allowing us to have lower and upper estimates of the Gibbs distribution $\omega_\alpha^{\tilde{\mathcal{E}}}(V^j)$:

$$e^{-\alpha\overline{\tilde{\mathcal{E}}}} \leq \omega_\alpha^{\tilde{\mathcal{E}}}(V^j) = e^{-\alpha\tilde{\mathcal{E}}(V^j)} \leq e^{-\alpha\underline{\tilde{\mathcal{E}}}}. \tag{2.15}$$

This gives an estimate of the difference $v_{\alpha,\tilde{\mathcal{E}}}(\widehat{\rho}^N) - v_{\alpha,\tilde{\mathcal{E}}}(\rho^N)$ in terms of $|V^j - \widehat{V}^j|$

$$\begin{aligned}|\omega_\alpha^{\tilde{\mathcal{E}}}(V^j) - \omega_\alpha^{\tilde{\mathcal{E}}}(\widehat{V}^j)| = |e^{-\alpha\tilde{\mathcal{E}}(V^j)} - e^{-\alpha\tilde{\mathcal{E}}(\widehat{V}^j)}| &\leq \alpha e^{-\alpha\underline{\tilde{\mathcal{E}}}}|\tilde{\mathcal{E}}(V^j) - \tilde{\mathcal{E}}(\widehat{V}^j)| \\ &\leq \alpha L e^{-\alpha\underline{\tilde{\mathcal{E}}}}|V^j - \widehat{V}^j|\end{aligned}$$

through the derivative of $\omega_\alpha^{\tilde{\mathcal{E}}}$. □

Theorem 2.3. *Under Assumptions 2.1 and 2.2, let ρ_0 be a probability measure on Γ and, for every $N \in \mathbb{N}$, $(V_0^i)_{i\in[N]}$ be N i.i.d. random variables with the common law ρ_0.*

For every $N \in \mathbb{N}$, there exists a pathwise unique strong solution $((V_t^i)_{t\geq 0})_{i\in[N]}$ to the particle system (2.1) with the initial data $(V_0^i)_{i\in[N]}$. Moreover, it holds that $V_t^i \in \Gamma$ for all $i \in [N]$ and any $t > 0$.

Idea of the proof. Given P_1, P_2, P_3 and $\tilde{\mathcal{E}}$, Lemma 2.1 shows that the SDE (2.12) has locally Lipschitz coefficients, so it admits a local, pathwise-unique, strong solution by Theorem 2.2.

Moreover, we apply Itô's formula (2.6) with $\varphi(x) = \gamma(x)$ to show that $V_t^i \in \Gamma$ for all $i \in [N]$ and any $t > 0$. The process is indeed continuous and we can consider a smooth extension of γ outside the neighborhood $\overline{\Gamma}$ of Γ. If the first deterministic λ-dependent terms immediately vanish thanks to the orthogonality property (2.9), we understand here the role played by the correction term on the stochastic equation (2.1).

Simple computations where we employ that $|\nabla\gamma| = 1$ lead to

$$\frac{d\gamma(V_t^i)}{dt} = 0 \tag{2.16}$$

and hence $\gamma(V_t^i) = \gamma(V_0^i) = 0$ for all $t > 0$, which ensures that the solution is bounded at finite times, hence we have a global solution. Since all $V_t^i \in \Gamma$, the solution to the regularized system (2.10) is a solution to (2.1), which

provides the global existence of a solution to (2.1).

□

2.2 Well-posedness for the mean-field dynamics

From this section on, we will be working with Borel probability measures on $\mathbb{R}^d$ with finite second moment, namely

$$\mathcal{P}_2(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) \text{ such that } \int_{\mathbb{R}^d} |z|^2 \mu(dz) < \infty \right\}$$

that we equip with the 2-Wasserstein metric. From [3], we recall the definition of the p-th Wasserstein distance for $p \geq 1$.

Definition 2.1 (Wasserstein Metric)**.** *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-th moment. We equip this space with the Wasserstein distance*

$$W_p^p(\mu, \nu) := \inf \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} |z - \hat{z}|^p \, d\pi(\mu, \nu) \;\middle|\; \pi \in \Pi(\mu, \nu) \right\} \tag{2.17}$$

where $\Pi(\mu, \nu)$ denotes 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 components respectively. The Wasserstein distance can also be expressed as

$$W_p^p(\mu, \nu) = \inf \left\{ \mathbb{E}[|Z - \overline{Z}|^p] \right\} \tag{2.18}$$

where the infimum is taken over all joint distributions of the random variables Z, $\overline{Z}$ with marginals μ, ν respectively.

For $\mathbb{R}^d$ and $p \in [1, \infty)$, the Wasserstein distance W_p is compatible with the weak topology in $\mathcal{P}_p(\mathbb{R}^d)$ [3, Chapter 11]. Therefore, W_2 metrizes the weak convergence in $\mathcal{P}_2(\mathbb{R}^d)$ and convergence in W_2 implies convergence of the first two moments (see [42, Chapter 6] for more details).

Notice that, for any $\rho \in \mathcal{P}_2(\mathbb{R}^d)$

$$\frac{\omega_\alpha^{\tilde{\mathcal{E}}}(v)}{\|\omega_\alpha^{\tilde{\mathcal{E}}}\|_{L^1(\rho)}} \leq \frac{e^{-\alpha\underline{\tilde{\mathcal{E}}}}}{\|e^{-\alpha\tilde{\mathcal{E}}}\|_{L^1(\rho)}} \leq e^{\alpha(\overline{\tilde{\mathcal{E}}}-\underline{\tilde{\mathcal{E}}})} =: C_{\alpha,\tilde{\mathcal{E}}} \quad \forall v \in \mathbb{R}^d\,. \tag{2.19}$$

A direct application of the above leads to

$$v_{\alpha,\tilde{\mathcal{E}}}(\rho) := \frac{\int_{\mathbb{R}^d} v\omega_\alpha^{\tilde{\mathcal{E}}}(v)\,d\rho}{\int_{\mathbb{R}^d} \omega_\alpha^{\tilde{\mathcal{E}}}(v)\,d\rho} = \frac{\int_{\mathbb{R}^d} ve^{-\alpha\tilde{\mathcal{E}}(v)}\,d\rho}{\int_{\mathbb{R}^d} e^{-\alpha\tilde{\mathcal{E}}(v)}\,d\rho} \leq C_{\alpha,\tilde{\mathcal{E}}}\int_{\mathbb{R}^d}|v|d\rho \leq \frac{C_{\alpha,\tilde{\mathcal{E}}}(1+m_2)}{2}, \tag{2.20}$$

with $m_2 := m_2(\rho) := \int_{\mathbb{R}^d}|v|^2 d\rho(v)$.

The existence of a unique process satisfying the mean-field dynamics (2.4) is shown through the well-known Leray-Schauder fixed point theorem for infinite dimensional spaces, see for instance [18, Chapter 10].

Theorem 2.4. *Let $\mathcal{T}$ be a compact mapping of a Banach space $\mathcal{B}$ into itself, and suppose there exists a constant C such that*

$$\|x\|_{\mathcal{B}} < C$$

$\forall x \in \mathcal{B}$ and $\vartheta \in [0,1]$ satisfying $x = \vartheta\mathcal{T}x$. Then $\mathcal{T}$ has a fixed point.

The proof of the well-posedness follows closely the calculations carried out both for the unconstrained CBO method in [8] and for $\Gamma = \mathbb{S}^{d-1}$ in [16]. Before we state the theorem, let us start with the following stability estimate:

Lemma 2.2. *Assume that $\rho, \widehat{\rho} \in \mathcal{P}_c(\mathbb{R}^d)$ (with compact support), it holds*

$$|v_{\alpha,\tilde{\mathcal{E}}}(\rho) - v_{\alpha,\tilde{\mathcal{E}}}(\widehat{\rho})| \leq CW_p(\rho,\widehat{\rho})\,, \tag{2.21}$$

for any $1 \leq p < \infty$, where $C = C(C_{\alpha,\tilde{\mathcal{E}}}, \alpha, L) > 0$.

Idea of the proof. First of all, we notice that the difference can be rewritten

as

$$v_{\alpha,\tilde{\mathcal{E}}}(\rho) - v_{\alpha,\tilde{\mathcal{E}}}(\widehat{\rho}) = \frac{\int_{\mathbb{R}^d} v e^{-\alpha\tilde{\mathcal{E}}(v)}\, d\rho(v)}{\|e^{-\alpha\tilde{\mathcal{E}}}\|_{L^1(\rho)}} - \frac{\int_{\mathbb{R}^d} \widehat{v} e^{-\alpha\tilde{\mathcal{E}}(\widehat{v})}\, d\widehat{\rho}(\widehat{v})}{\|e^{-\alpha\tilde{\mathcal{E}}}\|_{L^1(\widehat{\rho})}}$$
$$= \iint_{\mathbb{R}^d\times\mathbb{R}^d} \frac{v e^{-\alpha\tilde{\mathcal{E}}(v)}}{\|e^{-\alpha\tilde{\mathcal{E}}}\|_{L^1(\rho)}} - \frac{\widehat{v} e^{-\alpha\tilde{\mathcal{E}}(\widehat{v})}}{\|e^{-\alpha\tilde{\mathcal{E}}}\|_{L^1(\widehat{\rho})}} d\pi(v,\widehat{v})$$

where $\pi \in \Pi(\rho,\widehat{\rho})$ is an arbitrary coupling of ρ and $\widehat{\rho}$. Using standard estimates of the kind of (2.15), we can bound the integrand norm in terms of $|v-\widehat{v}|^p$ and obtain

$$|v_{\alpha,\tilde{\mathcal{E}}}(\rho) - v_{\alpha,\tilde{\mathcal{E}}}(\widehat{\rho})| \leq C \left(\iint_{\mathbb{R}^d\times\mathbb{R}^d} |v-\widehat{v}|^p d\pi(v,\widehat{v}) \right)^{\frac{1}{p}}, \tag{2.22}$$

where C depends only on $C_{\alpha,\tilde{\mathcal{E}}}$ and α, L. Lastly, we need to optimize the last expression over all couplings π, which yields (2.21). □

The following theorem states the well-posedness for the mean-field dynamics (2.4).

Theorem 2.5. *Let $\mathcal{E}$ and $\tilde{\mathcal{E}}$ satisfy Assumptions 2.1 and 2.2. Then, there exists a unique process $\overline{V} \in \mathcal{C}([0,T],\mathbb{R}^d)$, $T>0$, satisfying the nonlinear SDE* (2.4)

$$d\overline{V}_t = \lambda P(\overline{V}_t) v_{\alpha,\mathcal{E}}(\rho_t) dt + \sigma |\overline{V}_t - v_{\alpha,\mathcal{E}}(\rho_t)| P(\overline{V}_t) dB_t -$$
$$\frac{\sigma^2}{2} (\overline{V}_t - v_{\alpha,\mathcal{E}}(\rho_t))^2 \Delta\gamma(\overline{V}_t) \nabla\gamma(\overline{V}_t) dt\,,$$

in strong sense for any initial data $\overline{V}_0 \in \Gamma$ distributed according to $\rho_0 \in \mathcal{P}(\Gamma)$, where

$$v_{\alpha,\mathcal{E}}(\rho_t) = \frac{\int_{\mathbb{R}^d} v e^{-\alpha\mathcal{E}(v)}\, d\rho_t}{\int_{\mathbb{R}^d} e^{-\alpha\mathcal{E}(v)}\, d\rho_t},$$

and $\rho_t = law(\overline{V}_t)$ for all $t \in [0,T]$. Moreover $\overline{V}_t \in \Gamma$ for all $t \in [0,T]$.

Idea of the proof. As already mentioned, the proof is based on Theorem 2.4 and it is carried out in several steps.

Primarly, we underline that for some given $\xi \in \mathcal{C}([0,T],\mathbb{R}^d)$, a distribution ρ_0 on Γ and $\overline{V}_0$ with law ρ_0, we can uniquely solve the SDE

$$d\overline{V}_t = \lambda P_1(\overline{V}_t)\xi_t dt + \sigma|\overline{V}_t - \xi_t|P_1(\overline{V}_t)dB_t - \frac{\sigma^2}{2}(\overline{V}_t - \xi_t)^2 P_2(\overline{V}_t)P_3(\overline{V}_t)dt\,. \tag{2.23}$$

We note that the same argument as before, see equation (2.16), shows that $\gamma(\overline{V}_t) = 0$ for all times t. This introduces $\rho_t = \text{law}(\overline{V}_t)$ and $\rho \in \mathcal{C}([0,T],\mathcal{P}_2(\mathbb{R}^d))$. By setting $\mathcal{T}\xi := v_{\alpha,\tilde{\mathcal{E}}}(\rho) \in \mathcal{C}([0,T],\mathbb{R}^d)$, we define the map

$$\mathcal{T} : \mathcal{C}([0,T],\mathbb{R}^d) \to \mathcal{C}([0,T],\mathbb{R}^d), \quad \xi \mapsto \mathcal{T}(\xi) = v_{\alpha,\tilde{\mathcal{E}}}(\rho)\,, \tag{2.24}$$

which we prove to be compact. In order to verify the compactness of $\mathcal{T}$, we first notice that, by Itô's isometry and by definition of Wasserstein distance we have

$$W_2(\rho_t,\rho_s) \le C|t-s|^{\frac{1}{2}} \tag{2.25}$$

and, afterwards, we apply Lemma 2.2 obtaining

$$|v_{\alpha,\tilde{\mathcal{E}}}(\rho_t) - v_{\alpha,\tilde{\mathcal{E}}}(\rho_s)| \le C|t-s|^{\frac{1}{2}}\,. \tag{2.26}$$

This provides the Hölder continuity of $t \to v_{\alpha,\tilde{\mathcal{E}}}(\rho_t)$. Thus, one has $\mathcal{T}(\mathcal{C}([0,T],\mathbb{R}^d)) \subset \mathcal{C}^{0,\frac{1}{2}}([0,T],\mathbb{R}^d) \hookrightarrow \mathcal{C}([0,T],\mathbb{R}^d)$, which implies the compactness of the map $\mathcal{T}$.

After checking the boundedness of the set

$$\mathcal{A} := \left\{\xi \in \mathcal{C}([0,T],\mathbb{R}^d) : \xi = \vartheta\mathcal{T}\xi \text{ for some } 0 \le \vartheta \le 1\right\}\,. \tag{2.27}$$

For $\xi \in \mathcal{A}$, there exists some $\overline{V}_t$ satisfying (2.23) with law $\rho \in \mathcal{C}([0,T],\mathcal{P}_2(\mathbb{R}^d))$ such that $\xi = \vartheta v_{\alpha,\tilde{\mathcal{E}}}(\rho)$. Due to (2.20), for any $t \in [0,T]$

$$|\xi_t|^2 = \vartheta^2|v_{\alpha,\tilde{\mathcal{E}}}(\rho_t)|^2 \le \vartheta^2\left(\frac{\int_{\mathbb{R}^d} v e^{-\alpha\tilde{\mathcal{E}}(v)}d\rho_t(v)}{\|e^{-\alpha\tilde{\mathcal{E}}}\|_{L^1(\rho_t)}}\right)^2 \le C \tag{2.28}$$

and we apply the Leray-Schauder fixed point theorem. Hence, there exists a fixed point ξ for the mapping $\mathcal{T}$ and thereby a solution of

$$d\overline{V}_t = \lambda P_1(\overline{V}_t)v_{\alpha,\tilde{\mathcal{E}}}(\rho_t)dt + \sigma|\overline{V}_t - v_{\alpha,\tilde{\mathcal{E}}}(\rho_t)|P_1(\overline{V}_t)dB_t \\ - \frac{\sigma^2}{2}(\overline{V}_t - v_{\alpha,\tilde{\mathcal{E}}}(\rho_t))^2 P_2(\overline{V}_t)P_3(\overline{V}_t)dt \quad (2.29)$$

with $\text{law}(\overline{V}_t) = \rho_t$.

In order to prove the uniqueness of the solution of (2.29), we consider two fixed points ξ^1 and ξ^2, and their corresponding processes $\overline{V}_t^1$, $\overline{V}_t^2$. In particular, applying the Itô's isometry and standard estimates on $\mathbb{E}[|\overline{Z}_t|^2]$ we obtain the inequality

$$\mathbb{E}[|\overline{Z}_t|^2] \leq C\mathbb{E}[|\overline{Z}_0|^2] + C\int_0^t \mathbb{E}[|\overline{Z}_s|^2]ds\,. \quad (2.30)$$

Therefore, thanks to the Grönwall's inequality with $\mathbb{E}[|\overline{Z}_0|^2] = 0$, we can conclude $\mathbb{E}[|\overline{Z}_t|^2] = 0$ for all $t \in [0,T]$ and hence $\xi^1 \equiv \xi^2$ by Lemma 2.2. Finally, similar to the argument in Theorem 2.3, the unique solution to the regularized SDE (2.29) is also the unique solution to the mean-field dynamics (2.4) due to the fact that $\gamma(\overline{V}_t) = 0$ for all $t \in [0,T]$. □

2.3 Well-posedness for the mean-field PDE

In this section we will briefly show how to obtain a weak solution to the mean-field PDE (2.3) by construction.

We start from the solution $\{\overline{V}_t : t \geq 0\}$ to (2.4) obtained in the last section, with the initial data $\overline{V}_0$ distributed according to ρ_0. For any $\varphi \in$

$C_c^\infty(\mathbb{R}^d)$, it follows from Itô's formula (2.6) that

$$\begin{aligned} d\varphi(\overline{V}_t) = \nabla\varphi(\overline{V}_t)\cdot\Big(\lambda P(\overline{V}_t)v_{\alpha,\mathcal{E}}(\rho_t) - \frac{\sigma^2}{2}(\overline{V}_t - v_{\alpha,\mathcal{E}}(\rho_t))^2\Delta\gamma(\overline{V}_t)\nabla\gamma(\overline{V}_t)\Big)\,dt \\ + \sigma|\overline{V}_t - v_{\alpha,\mathcal{E}}(\rho_t)|\nabla\varphi(\overline{V}_t)\cdot P(\overline{V}_t)dB_t \\ + \frac{\sigma^2}{2}\left(\overline{V}_t - v_{\alpha,\mathcal{E}}(\rho_t)\right)^2\nabla^2\varphi(\overline{V}_t) : \left(I - \nabla\gamma(\overline{V}_t)\nabla\gamma(\overline{V}_t)^t\right)dt\,, \quad (2.31) \end{aligned}$$

where $\nabla^2\varphi$ is the Hessian and $A : B := \mathrm{Tr}(A^T B)$. Taking the expectation on both sides of (2.31), the law ρ_t of $\overline{V}_t$ as a measure on $\mathbb{R}^d$ satisfies

$$\begin{aligned} \frac{d}{dt}\int_{\mathbb{R}^d}\varphi(v)d\rho_t(v) = \int_{\mathbb{R}^d}\nabla\varphi(v)\cdot\Big(\lambda(I - \nabla\gamma(v)\nabla\gamma(v)^t)v_{\alpha,\mathcal{E}}(\rho_t) \\ - \frac{\sigma^2}{2}(v - v_{\alpha,\mathcal{E}}(\rho_t))^2\Delta\gamma(v)\nabla\gamma(v)\Big)d\rho_t(v) \\ + \int_{\mathbb{R}^d}\frac{\sigma^2}{2}\left(v - v_{\alpha,\mathcal{E}}(\rho_t)\right)^2\nabla^2\varphi(v) : \left(I - \nabla\gamma(v)\nabla\gamma(v)^t\right)d\rho_t(v)\,. \quad (2.32) \end{aligned}$$

As we have proved that $\overline{V}_t \in \Gamma$ almost surely, that is, the density ρ_t is concentrated on Γ for any t, we have $\mathrm{supp}(\rho_t) \subset \Gamma$. Let us now define the restriction μ_t of ρ_t on Γ by

$$\int_\Gamma \Phi(v)d\mu_t(v) = \int_{\mathbb{R}^d}\varphi(v)d\rho_t(v) \quad (2.33)$$

for all continuous maps $\Phi \in \mathcal{C}(\Gamma)$, where $\varphi \in \mathcal{C}_b(\mathbb{R}^d)$ equals Φ on Γ.

Next, we define the projection

$$\Pi_\Gamma(v) = v - \gamma(v)\nabla\gamma(v) \in \Gamma, \quad \text{for } v \in \mathbb{R}^d\,.$$

In the case of $\Gamma = \mathbb{S}^{d-1}$ a unique projection can be defined on $\mathbb{R}^d \setminus \{0\}$, but for generic hypersurfaces we need to take into account a strip of width $\delta > 0$ about Γ, $\Gamma_\delta \subset \mathbb{R}^d$, where $\delta > 0$ is small enough to ensure that the decomposition

$$v = \Pi_\Gamma(v) + \gamma(v)\nabla\gamma(v) \quad (2.34)$$

is unique for $v \in \Gamma_\delta$. We know such δ exists since $\gamma \in \mathcal{C}^2(\overline{\Gamma})$, see for example [11, Section 2.1]. Let now $\Phi \in \mathcal{C}^\infty(\Gamma)$ and define a function $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ such that

$$\varphi(v) = \Phi\left(\Pi_\Gamma(v)\right) \qquad \text{for all } v \in \Gamma_\delta\,. \tag{2.35}$$

Then, φ defined above is 0-homogeneous in v in the strip Γ_δ, so that $\nabla\varphi(v)\cdot\nabla\gamma(v) = 0$ for all v in the support Γ of ρ_t, which leads to $\nabla^2\varphi(v) : \nabla\gamma(v)\nabla\gamma(v)^t = 0$. Hence,

$$\begin{aligned}\frac{d}{dt}\int_\Gamma \Phi(v)d\mu_t(v) &= \frac{d}{dt}\int_{\mathbb{R}^d}\varphi(v)d\rho_t(v)\\ &= \lambda\int_{\mathbb{R}^d}\nabla\varphi(v)\cdot\left((I-\nabla\gamma(v)\nabla\gamma(v)^t)v_{\alpha,\mathcal{E}}(\rho_t)\right)d\rho_t(v)\\ &\quad+\int_{\mathbb{R}^d}\frac{\sigma^2}{2}\left(v-v_{\alpha,\mathcal{E}}(\rho_t)\right)^2\Delta\varphi(v)d\rho_t(v)\,.\end{aligned}$$

Let us now relate the Euclidean differential operators to corresponding operators on Γ, so that for $v \in \Gamma$ it holds $\nabla_\Gamma\Phi(v) = \nabla\varphi(v)$ and $\Delta_\Gamma\Phi(v) = \Delta\varphi(v)$. Therefore, we obtain

$$\begin{aligned}\frac{d}{dt}\int_\Gamma \Phi(v)d\mu_t(v) = \lambda\int_\Gamma\nabla_\Gamma\Phi(v)\cdot\left((I-\nabla\gamma(v)\nabla\gamma(v)^t)v_{\alpha,\mathcal{E}}(\mu_t)\right)d\mu_t(v)\\ +\int_\Gamma\frac{\sigma^2}{2}\left(v-v_{\alpha,\mathcal{E}}(\mu_t)\right)^2\Delta_\Gamma\Phi(v)d\mu_t(v)\,,\end{aligned} \tag{2.36}$$

where

$$v_{\alpha,\mathcal{E}}(\mu_t) = \frac{\int_\Gamma v e^{-\alpha\mathcal{E}(v)}\,d\mu_t}{\int_\Gamma e^{-\alpha\mathcal{E}(v)}\,d\mu_t}. \tag{2.37}$$

Thus, by this construction, we obtain a weak solution μ_t to the PDE (2.3).

Next, we prove the uniqueness of solutions to (2.3). We assume that ρ_t^1 and ρ_t^2 are two solutions to (2.3) with the same initial data ρ_0, and that at each time t we treat them as measures on $\mathbb{R}^d$ concentrated on the hypersurface

Γ. Then, we construct two linear processes $(\overline{V}_t^i)_{t\geq 0}$ $(i=1,2)$ satisfying

$$d\overline{V}_t^i = \lambda P_1(\overline{V}_t^i)v_{\alpha,\mathcal{E}}(\rho_t^i)dt + \sigma|\overline{V}_t^i - v_{\alpha,\mathcal{E}}(\rho_t^i)|P_1(\overline{V}_t^i)dB_t \\ - \frac{\sigma^2}{2}(\overline{V}_t^i - v_{\alpha,\mathcal{E}}(\rho_t^i))^2 P_2(\overline{V}_t^i)P_3(\overline{V}_t^i)dt\,, \quad (2.38)$$

with the common initial data $\overline{V}_0$ distributed according to ρ_0. Let us denote $\mathrm{law}(\overline{V}_t^i) = \bar{\rho}_t^i$ $(i=1,2)$ as measures on $\mathbb{R}^d$, which are solutions to the following linear PDE

$$\partial_t \bar{\rho}_t^i = \nabla\cdot\left(\bar{\rho}_t^i\left(-\lambda P_1(v)v_{\alpha,\mathcal{E}}(\rho_t^i) + \frac{\sigma^2}{2}(v - v_{\alpha,\mathcal{E}}(\rho_t^i))^2 P_2(v)P_3(v)\right)\right) \\ + \frac{\sigma^2}{2}\sum_{k,\ell=1}^{d}\frac{\partial^2}{\partial v_k \partial v_\ell}\left(|v - v_{\alpha,\mathcal{E}}(\rho_t^i)|^2 (P_1 P_1^T)_{k\ell}\bar{\rho}_t^i\right)\,. \quad (2.39)$$

Since the uniqueness for the above linear PDE holds and ρ_t^i is also a solution to the above PDE on $\mathbb{R}^d$ (see (2.32)), it follows that $\bar{\rho}_t^i = \rho_t^i$ $(i = 1,2)$. Consequently, the processes $(\overline{V}_t^i)_{(t\geq 0)}$ are solutions to the nonlinear SDE (2.4), for which the uniqueness has been obtained. Hence, $(\overline{V}_t^1)_{(t\geq 0)}$ and $(\overline{V}_t^2)_{(t\geq 0)}$ are equal, which implies $\rho_t^1 = \bar{\rho}_t^1 = \bar{\rho}_t^2 = \rho_t^2$. Thus, the uniqueness is obtained.

2.4 Mean-field limit

The well-posedness of (2.1), (2.3) and (2.4) obtained in the last section provides all the tools we need for the mean-field limit. Let $((\overline{V}_t^i)_{t\geq 0})_{i\in[N]}$ be N independent copies of solutions to the mean-field dynamics (2.4). They are i.i.d. with the same distribution ρ_t and we assume that $((V_t^i)_{t\geq 0})_{i\in[N]}$ is the solution to the particle system (2.1). Since $\overline{V}_t^i, V_t^i \in \Gamma$ for all i and t, $((\overline{V}_t^i)_{t\geq 0})_{i\in[N]}$ and $((V_t^i)_{t\geq 0})_{i\in[N]}$ are solutions to the corresponding regularized systems, (2.29) and (2.10) respectively. We denote below $\overline{\rho}_t^N = \frac{1}{N}\sum_{j=1}^N \delta_{\overline{V}_t^j}$ and $\rho_t = \mathrm{law}(\overline{V}_t)$.

Before stating our theorem on the mean-field limit, let us introduce the

following lemma on a large deviation bound.

Lemma 2.3. *Let $\mathcal{E}$ and $\tilde{\mathcal{E}}$ satisfy Assumptions 2.1 and 2.2. Let $((\overline{V}_t^i)_{t\geq 0})_{i\in[N]}$ be the solution to the mean-field dynamics (2.29), which are i.i.d. with common distribution $\rho \in \mathcal{C}([0,T],\mathcal{P}_2(\mathbb{R}^d))$. Then, there exists a constant C depending only on $C_{\alpha,\tilde{\mathcal{E}}}$ and $M = \mathrm{diam}(\Gamma)$ such that*

$$\sup_{t\in[0,T]} \mathbb{E}\left[|v_{\alpha,\tilde{\mathcal{E}}}(\overline{\rho}_t^N) - v_{\alpha,\tilde{\mathcal{E}}}(\rho_t)|^2\right] \leq CN^{-1}\,. \tag{2.40}$$

Idea of the proof. The calculations can be carried out exactly as in [16]: we bound quantities

$$\overline{Z}_t^j := \overline{V}_t^j e^{-\alpha\tilde{\mathcal{E}}(\overline{V}_t^j)} - \int_{\mathbb{R}^d} v e^{-\alpha\tilde{\mathcal{E}}(v)} d\rho_t\,, \tag{2.41}$$

thanks to the existence of a supremum and an infimum of $\tilde{\mathcal{E}}$. □

We note that $C \propto C^3_{\alpha,\mathcal{E}}$ and it goes to infinity exponentially in α as $\alpha \to \infty$ and that C here depends on M. We can now present the mean-field limit result, which relates the empirical measure ρ_t^N to ρ_t, the solution of the mean-field PDE.

Theorem 2.6. *Under the Assumptions 2.1 and 2.2, for any $T > 0$, let $((V_t^i)_{t\in[0,T]})_{i\in[N]}$ and $((\overline{V}_t^i)_{t\in[0,T]})_{i\in[N]}$ be respective solutions to (2.1) and (2.4) up to time T with the same initial data $V_0^i = \overline{V}_0^i$ and the same Brownian motions $((B_t^i)_{t\in[0,T]})_{i\in[N]}$. Then, there exists a constant $C > 0$ depending only on α, σ, $\|\nabla P_1\|_\infty$, $\|P_1\|_\infty$, $\|\nabla P_2\|_\infty$, $\|P_2\|_\infty$, $\|\nabla P_3\|_\infty$, $\|P_3\|_\infty$, L, M and $C_{\alpha,\tilde{\mathcal{E}}}$, such that*

$$\sup_{i=1,\cdots,N} \mathbb{E}[|V_t^i - \overline{V}_t^i|^2] \leq CT\left(1 + CTe^{CT}\right)\frac{1}{N}\,, \tag{2.42}$$

holds for all $0 \leq t \leq T$.

We remark that the estimate above guarantees the weak convergence of

the empirical measure ρ_t^N towards ρ_t, in the following sense

$$\sup_{t\in[0,T]} \mathbb{E}\left[|\langle \rho_t^N, \phi\rangle - \langle \rho_t, \phi\rangle|^2\right] \to 0 \text{ as } N \to \infty \tag{2.43}$$

for any test function $\phi \in \mathcal{C}_b^1(\mathbb{R}^d)$.

Indeed, one has

$$\mathbb{E}\left[|\langle \rho_t^N, \phi\rangle - \langle \rho_t, \phi\rangle|^2\right] = \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N \phi(V_t^i) - \int_{\mathbb{R}^d}\phi(v)d\rho_t(v)\right|^2\right]$$
$$\leq 2\mathbb{E}\left[|\phi(V_t^1) - \phi(\overline{V}_t^1)|^2\right] + 2\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N \phi(\overline{V}_t^i) - \int_{\mathbb{R}^d}\phi(v)d\rho_t(v)\right|^2\right] \leq \frac{C}{N}\|\phi\|_{\mathcal{C}_1}^2 .$$

Idea of the proof. Given that $((\overline{V}_t^i)_{t\geq 0})_{i\in[N]}$ and $((V_t^i)_{t\geq 0})_{i\in[N]}$ are also solutions to the corresponding regularized systems (2.29) and (2.10) respectively, the following holds

$$\begin{aligned} d(V_t^i - \overline{V}_t^i) &= \lambda\left(P_1(V_t^i)v_{\alpha,\tilde{\mathcal{E}}}(\rho_t^N) - P_1(\overline{V}_t^i)v_{\alpha,\tilde{\mathcal{E}}}(\rho_t)\right)dt \\ &\quad + \sigma\left(|V_t^i - v_{\alpha,\tilde{\mathcal{E}}}(\rho_t^N)|P_1(V_t^i) - |\overline{V}_t^i - v_{\alpha,\tilde{\mathcal{E}}}(\rho_t)|P_1(\overline{V}_t^i)\right)dB_t^i \\ &- \frac{\sigma^2}{2}\Big((V_t^i - v_{\alpha,\tilde{\mathcal{E}}}(\rho_t^N))^2 P_2(V_t^i)P_3(V_t^i) - (\overline{V}_t^i - v_{\alpha,\tilde{\mathcal{E}}}(\rho_t))^2 P_2(\overline{V}_t^i)P_3(\overline{V}_t^i)\Big)dt . \end{aligned}$$

Then, we then Itô's formula to $d(V_t^i - \overline{V}_t^i)^2$ and carry out the calculations in order to obtain an estimate in terms of $|V_t^i - \overline{V}_t^i|^2$ itself.

If we examine the expectation afterwards, it holds:

$$\begin{aligned} \mathbb{E}[|V_t^i - \overline{V}_t^i|^2] &\leq \mathbb{E}[|V_0^i - \overline{V}_0^i|^2] \\ &\quad + C\int_0^t \frac{\sum_{i=1}^N \mathbb{E}[|V_s^i - \overline{V}_s^i|^2]}{N}ds + C\int_0^t \mathbb{E}[|V_s^i - \overline{V}_s^i|^2]ds \\ &\quad + C\int_0^t \mathbb{E}\left[|v_{\alpha,\tilde{\mathcal{E}}}(\overline{\rho}_s^N) - v_{\alpha,\tilde{\mathcal{E}}}(\rho_s)|^2\right]ds \end{aligned} \tag{2.44}$$

We use Lemma 2.3 and Grönwall's inequality together with $\mathbb{E}[|V_0^i - \overline{V}_0^i|^2] = 0$ to get

$$\sup_{i=1,\cdots,N} \mathbb{E}[|V_t^i - \overline{V}_t^i|^2] \leq CT\left(1 + CTe^{CT}\right)\frac{1}{N}, \tag{2.45}$$

for all $t \in [0,T]$, which completes the proof. □

Let us draw the attention to the constant $C > 0$ appearing in the estimate above. C may depend on the dimension through the norm of P_2 or ∇P_2. Nevertheless, we expect this dependency to scale at most linearly as $d-1$. In fact, for the case $\Gamma = \mathbb{S}^{d-1}$, we have $P_2(v) = \Delta\gamma(v) = \frac{d-1}{|v|}$. Fornasier et al. suggest in [16] that, in general, there is no curse of dimensionality involved in estimates of the type of (2.45).

Chapter 3

Convergence Estimates

In this chapter we will attempt to analyze the large time behavior of the solution $\rho_t \in \mathcal{P}_2(\Gamma)$ of the mean-field PDE, the Fokker-Planck equation

$$\partial_t \rho_t = \lambda \nabla_\Gamma \cdot ((P(v)(v - v_{\alpha,\mathcal{E}}(\rho_t))\rho_t) + \frac{\sigma^2}{2}\Delta_\Gamma \left(|v - v_{\alpha,\mathcal{E}}(\rho_t)|^2 \rho_t\right), \tag{3.1}$$

for $t > 0$, $v \in \Gamma$ and with initial data $\rho_0 \in \mathcal{P}_2(\Gamma)$. As in [8] and [17], we focus on the moments

$$E(\rho_t) = \int_\Gamma v \, d\rho_t(v) \quad \text{and} \quad V(\rho_t) = \frac{1}{2}\int_\Gamma |v - E(\rho_t)|^2 d\rho_t(v)$$

in order to study the evolution of ρ_t.

Ideally, we would like to provide sufficient conditions on $\mathcal{E}$, the parameters $\{\lambda, \sigma, \alpha\}$ and ρ_0 such that a *uniform consensus formation* among a minimizer v^* happens, more precisely, such that

$$\rho_t \longrightarrow \delta_{v^*} \quad \text{as} \quad t \to \infty \,. \tag{3.2}$$

In practice, this has been shown to be a challenging task both for the CBO method for unconstrained optimization and for the method developed for constrained optimization on the sphere, see [8] and [17]. The main difficulties lie on the fact that, if on one hand we need to choose α to be large in order to

apply the Laplace principle (1.8), on the other hand this makes the moments estimation much harder.

What it is possible to prove, both when the domain Ω of the objective function $\mathcal{E}$ is $\mathbb{R}^d$ and when $\Omega = \mathbb{S}^{d-1}$ is that, for any $\varepsilon > 0$, there exists a choice of the parameters α, λ, σ and specific conditions on the initial datum ρ_0, such that at some time $T > 0$ it holds

$$|E(\rho_T) - v^*| \leq \varepsilon\,, \tag{3.3}$$

where v^* is a minimizer of $\mathcal{E}$. Even though this is a different result from (3.2), the analysis sheds some light on the evolution of ρ_t and, hence, on the particles dynamics of the CBO method.

Section 3.1, illustrates the main techniques that have been employed in [8,17]. We start by studying the consensus formation of CBO systems and by focusing on how to prove the exponential decay of the variance. In Section 3.2, we attempt to adapt these techniques to the solution ρ_t of the mean-field PDE (3.1) defined on the hypersurface Γ and discuss the main difficulties that the constraint involves. We will computationally investigate the presented results with an example of constrained optimization on the three-dimensional torus, Section 3.3, and provide the proof of the auxiliary lemmas, Section 3.4.

Before we start, we present the class of functions we will consider throughout this chapter. The objective function $\mathcal{E}$ is a $\mathcal{C}^2(\Omega)$ function within its domain Ω. Depending on the context, Ω will be the whole space $\mathbb{R}^d$, or a neighborhood of $\mathbb{S}^{d-1}$ or of Γ, a generic hypersurface. Moreover, we assume $\mathcal{E}$ to satisfy the following properties.

Assumption 3.1.

1. *$\mathcal{E}$ is bounded and* $0 \leq \underline{\mathcal{E}} := \inf \mathcal{E} \leq \sup \mathcal{E} =: \overline{\mathcal{E}} < \infty$*;*

2. $\|\nabla \mathcal{E}\|_\infty \leq c_1$*;*

3. $\max\{\|\nabla^2 \mathcal{E}\|_\infty, \|\Delta \mathcal{E}\|_\infty\} \leq c_2$*;*

4. *For any $v \in \Omega$, there exists a minimizer $v^* \in \Omega$ of $\mathcal{E}$ (which may depend on v) such that it holds*

$$|v - v^*| \leq C_0 |\mathcal{E}(v) - \underline{\mathcal{E}}|^\beta ,$$

where β, C_0 are some positive constants.

The boundedness of $\mathcal{E}$, as we will see, is a key factor in the analysis of the consensus. This is automatically fulfilled (as soon as smoothness is provided) when we consider a compact Ω. The inverse continuity assumption 4., which needs to be verified depending on the specific application, is more technical and it is another key assumption for the proofs since it will allow us to use the Laplace principle.

In the next section, we show how we can obtain convergence guarantees for the CBO method designed for the unconstrained optimization method introduced in [33].

3.1 Convergence guarantees for unconstrained CBO

In this section, we briefly present the CBO method introduced in [33] to solve the problem

$$\min_{v \in \mathbb{R}^d} \mathcal{E}(v)$$

and its convergence guarantees. We remark that, even if the proof strategy is similar, the results we illustrate differ from the original analysis made in [8] since some of the arguments presented important issues. In particular, we will observe that, in order to have convergence in sense of (3.3), the initial datum ρ_0 needs to be already *well-concentrated*, in the sense that $V(\rho_0)$ needs to be sufficiently small.

In these simple settings, the particles $(V_t^i)_{i=1,\dots,N}$ satisfy the system of

SDEs

$$dV_t^i = -\lambda(V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N))dt + \sigma|V_t^i - v_{\alpha,\mathcal{E}}(\rho_t^N)|dB_t^i \tag{3.4}$$

where, as before, ρ_t^N denotes the empirical measure at time t. We then consider $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ which describes the evolution of the one-particle process resulting from the mean-field limit. The Fokker-Planck equation in this case reads

$$\partial_t \rho_t = \lambda \nabla \cdot ((v - v_{\alpha,\mathcal{E}}(\rho_t))\rho_t) + \frac{\sigma^2}{2}\Delta\left(|v - v_{\alpha,\mathcal{E}}(\rho_t)|^2 \rho_t\right), \tag{3.5}$$

for $t > 0$, $v \in \mathbb{R}^d$ and with initial datum ρ_0.

We first enunciate two auxiliary lemmas and, then, the statement of the convergence guarantees, together with the main results the proof is based on.

Lemma 3.1. *Let $v_{\alpha,\mathcal{E}}$ be defined as the expectation with respect to the measure $\omega_{\mathcal{E}}^\alpha / \|\omega_{\mathcal{E}}^\alpha\|_{L^1(\rho_t)} d\rho_t$, it holds*

$$\int_\Omega |v - v_{\alpha,\mathcal{E}}|^2 \, d\rho_t(v) \leq 2C_{\alpha,\mathcal{E}} V(\rho_t), \quad \text{where} \quad C_{\alpha,\mathcal{E}} := e^{-\alpha(\overline{\mathcal{E}} - \underline{\mathcal{E}})}. \tag{3.6}$$

Lemma 3.2. *The derivatives of $\omega_{\mathcal{E}}^\alpha(v) = e^{-\alpha\mathcal{E}(v)}$ are:*

$$\begin{aligned}
\nabla \omega_{\mathcal{E}}^\alpha &= -\alpha e^{-\alpha\mathcal{E}} \nabla\mathcal{E} \in \mathbb{R}^d; \\
\nabla^2 \omega_{\mathcal{E}}^\alpha &= -\alpha(-\alpha \nabla\mathcal{E} \otimes \nabla\mathcal{E} + \nabla^2\mathcal{E}) \in \mathbb{R}^{d\times d}; \\
\Delta \omega_{\mathcal{E}}^\alpha &= \alpha^2 e^{-\alpha\mathcal{E}} |\nabla\mathcal{E}|^2 - \alpha e^{-\alpha\mathcal{E}} \Delta\mathcal{E} \in \mathbb{R}.
\end{aligned}$$

Theorem 3.1. *Let ρ_t be the solution of (3.5). For any $\varepsilon > 0$ there exists a choice of parameters α, λ, σ such that, if the initial datum ρ_0 satisfies the condition*

$$V(\rho_0) \leq \|\omega_{\mathcal{E}}^\alpha\|^2_{L^1(\rho_0)} \tag{3.7}$$

for some $T > 0$, it holds

$$|E(\rho_t) - v^*| \leq \varepsilon \tag{3.8}$$

for some minimizer v^* of $\mathcal{E}$.

Remark 3.1. *The reason why we need to assume (3.7) will become clear in*

the proof of Proposition 3.1. The condition characterizes the result as local. The initial distribution ρ_0 (and so the particles) needs to be concentrated in order to have a convergent behavior at later times $t > 0$. Moreover, we recall that since

$$\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)} = \int e^{-\alpha\mathcal{E}(v)} d\rho_0(v)$$

vanishes as $\alpha \to \infty$, the condition is more restrictive if α is large.

Remark 3.2. *The original analysis attempted to show that, for any given $0 < \varepsilon \ll 1$ there exist some parameters $\{\alpha, \lambda, \sigma\}$ such that $\rho_t \longrightarrow \tilde{v}$ for $t \to \infty$, with $|\tilde{v} - v^*| \leq \varepsilon$ (see [8, Theorem 4.2]). It is certainly possible to show that, if the parameters are carefully chosen, ρ_t concentrates around a point $\tilde{v}$. However, limiting the evolution to a time horizon T seems to be a necessary condition for proving the concentration of the solution ρ_t around a minimizer (informally, $\tilde{v} \in B_\varepsilon(v^*)$). Indeed, the reason why we have to consider t to be bounded by the time horizon T will become clear in the proof of Proposition 3.1.*

Idea of the proof of Theorem 3.1. As in the sphere case, the proof of the theorem is based on the inverse continuity assumption, which assumes that there exists a minimizer v^* such that

$$|E(\rho_t) - v^*| \leq C_0 |\mathcal{E}(E(\rho_t)) - \underline{\mathcal{E}}|^{\beta} . \tag{3.9}$$

An estimate of the RHS is then given by triangular inequality as follows:

$$|\mathcal{E}(E(\rho_t)) - \underline{\mathcal{E}}| \leq \left| \mathcal{E}(E(\rho_t)) - \frac{-1}{\alpha} \log \|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} \right| + \left| \frac{-1}{\alpha} \log \|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} - \underline{\mathcal{E}} \right| .$$

The first term can be bounded by showing the variance decay, whereas for the second term one needs to make use of the Laplace principle. In view of our purpose, we focus now on how we can demonstrate the variance decay. Indeed, this will be interesting for later analysis when we consider the dynamics on hypersurfaces.

By the dual representation of 1-Wasserstein distance W_1, we know that

$$\begin{aligned}\left|\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} - \omega_{\mathcal{E}}^{\alpha}(E(\rho_t))\right| &= \left|\int_{\mathbb{R}^d} e^{-\alpha\underline{\mathcal{E}}} d(\rho_t(v) - \delta_{E(\rho_t)}(v))\right| \\ &\leq \alpha e^{-\alpha\underline{\mathcal{E}}} \|\nabla\mathcal{E}\|_{\infty} W_1(\rho_t, \delta_{E(\rho_t)}) \\ &\leq \alpha c_1 e^{-\alpha\underline{\mathcal{E}}} W_2(\rho_t, \delta_{E(\rho_t)}) \leq \sqrt{2}\alpha c_1 e^{-\alpha\underline{\mathcal{E}}} V(\rho_t)^{\frac{1}{2}},\end{aligned}$$

which implies

$$\begin{aligned}\left|\mathcal{E}(E(\rho_t)) - \frac{-1}{\alpha}\log\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)}\right| &= \left|\frac{1}{\alpha}\log\omega_{\mathcal{E}}^{\alpha}(E(\rho_t)) - \frac{-1}{\alpha}\log\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)}\right| \\ &\leq \frac{e^{\alpha\overline{\mathcal{E}}}}{\alpha}\left|\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} - \omega_{\mathcal{E}}^{\alpha}(E(\rho_t))\right| \\ &\leq \sqrt{2}c_1 C_{\alpha,\mathcal{E}} V(\rho_t)^{\frac{1}{2}}.\end{aligned}$$

We show now that, if the condition

$$2\lambda > d\sigma^2 C_{\alpha,\mathcal{E}} \tag{3.10}$$

holds, then the variance $V(\rho_t)$ decays exponentially. A simple computation of the evolution of the variance through the corresponding Itô's formula gives

$$\frac{d}{dt}V(\rho_t) = -\lambda\int_{\mathbb{R}^d}(v - E(\rho_t))^t(v - v_{\alpha,\mathcal{E}})d\rho_t(v) + \frac{d\sigma^2}{2}\int_{\mathbb{R}^d}|v - v_{\alpha,\mathcal{E}}|^2 d\rho_t(v) \tag{3.11}$$

$$\begin{aligned}&= -2\lambda V(\rho_t) + \frac{d\sigma^2}{2}\int_{\mathbb{R}^d}|v - v_{\alpha,\mathcal{E}}|^2 d\rho_t(v) \\ &\leq -\left(2\lambda - d\sigma^2 C_{\alpha,\mathcal{E}}\right)V(\rho_t)\end{aligned} \tag{3.12}$$

and, thus, we can conclude

$$V(\rho_t) \leq V(\rho_0)e^{-\left(2\lambda - d\sigma^2 C_{\alpha,\mathcal{E}}\right)t} \tag{3.13}$$

by the Grönwall's inequality. □

As we will see in Section 3.3, this simple argument is extremely difficult to generalize in the case of generic hypersurfaces Γ, mainly because the operator $P(v)$ breaks the symmetry we have in (3.11).

3.2 Converge estimate on hypersurfaces

In this section, we will study the evolution of ρ_t, the solution of the (3.1) defined on the hypersurface Γ. We will attempt to adapt the techniques presented in Section 3.1 and in [17] and discuss what the major complications caused by considering the dynamics to be constrained on Γ are. Indeed, even though we are able to formally derive the mean-field limit process, which is still an open issue for unconstrained CBO, the drawback of these settings is that the geometry of Γ makes it extremely hard to prove the variance decay through arguments like the ones employed in Theorem 3.1.

Given that the inverse continuity property holds only for $v \in \Gamma$, we slightly modify the previous approach, inequality (3.9), and we bound $|E(\rho_t) - v^*|$ in the following way

$$|E(\rho_t) - v^*| \leq |E(\rho_t) - \Pi(E(\rho_t))| + |\Pi(E(\rho_t)) - v^*|\,, \tag{3.14}$$

where $\Pi(\cdot)$ is the projection defined on a neighborhood of Γ such that $\Pi(w) = \operatorname{argmin}_{v\in\Gamma} |v - w|$. We assume here that $\Pi(E(\rho_t))$ is well-defined for any t we consider. Similarly to (3.9), we obtain

$$|\Pi(E(\rho_t)) - v^*| \leq C_0 |\mathcal{E}(\Pi(E(\rho_t))) - \underline{\mathcal{E}}|^\beta\,. \tag{3.15}$$

The main result of this section is how we can further bound these estimates in terms of the variance $V(\rho_t)$ and through the use of the Laplace principle. The results we present consist of a generalization of the analysis carried out in [17] where Γ is considered to be $\mathbb{S}^{d-1}$.

Theorem 3.2. *For any $\varepsilon > 0$, fixed parameters λ, σ and time horizon T,*

there exists $\alpha \gg 1$ *such that, if* ρ_0 *satisfies*

$$\overline{\mathcal{V}}_T = \sup_{t\in[0,T]} V(\rho_t) \leq \|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)}^4, \tag{3.16}$$

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

$$|E(\rho_t) - v^*| \leq \sqrt{2}V(\rho_t)^{\frac{1}{2}} + C(C_0, c_1, \beta)\left(C_{\alpha,\mathcal{E}}^{\beta} V(\rho_t)^{\frac{\beta}{2}} + \varepsilon^{\beta}\right). \tag{3.17}$$

Before we provide the proof of the Theorem, let us first discuss how we can adapt the same technique that has been used in Section 3.1 to fit our case.

Firstly, we need to estimate the RHS in equation (3.15) as follows:

$$\begin{aligned}|\mathcal{E}(\Pi(E(\rho_t))) - \underline{\mathcal{E}}| \leq &\left|\mathcal{E}(\Pi(E(\rho_t))) - \frac{-1}{\alpha}\log\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)}\right| \\ &+ \left|\frac{-1}{\alpha}\log\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} - \underline{\mathcal{E}}\right|. \end{aligned}\tag{3.18}$$

We note that, in order to estimate the second term above, we will make use of the Laplace principle. Proposition 3.1 will allow us to investigate what the relation between the Laplace principle and the evolution of the distribution ρ_t is. In particular, we show that if

$$-\frac{1}{\alpha}\log\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)} - \underline{\mathcal{E}} \leq \varepsilon$$

and if α is sufficiently large, we expect this bound to hold also for ρ_t for any $t \in [0,T]$, where $T > 0$ is a fixed time horizon.

In order to do so, we will require $V(\rho_t)$ to be bounded by $\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)}^4$ for all $t \in [0,T]$. Even though this seems a strong assumption, especially because $\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)}^4$ goes to 0 as $\alpha \to \infty$, we claim it is a reasonable condition given the exponential decay of the variance we showed in Section 3.1 for the unconstrained settings and the results obtained in [17]. We will computationally investigate the variance decay for our CBO method on hypersurfaces

in Section 3.3.

In view of the proof of Proposition 3.1, we present the following auxiliary Lemma which provides us with a lower bound on the norm of the weights $\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)}$.

Lemma 3.3. *Let c_1, c_2 be the bounds on the derivatives of $\mathcal{E}$ and c_γ the bound on the second derivatives of γ (see Definition (1.1)). Then we have:*

$$\frac{d}{dt}\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)}^2 \geq -\sigma^2 b_1(c_\gamma, \alpha, c_1, c_2, \underline{\mathcal{E}})V(\rho_t) - \lambda b_2(\alpha, c_1, \underline{\mathcal{E}})V(\rho_t)^{\frac{1}{2}} \tag{3.19}$$

with $b_1, b_2 \geq 0$ and $b_1, b_2 \to 0$ as $\alpha \to \infty$.

Proposition 3.1. *For any $\varepsilon > 0$, fixed parameters λ, σ and time horizon T, there exists $\alpha \gg 1$ such that, if ρ_0 satisfies*

$$\overline{\mathcal{V}}_T := \sup_{t\in[0,T]} V(\rho_t) \leq \|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)}^4, \tag{3.20}$$

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

$$-\frac{1}{\alpha}\log\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} - \underline{\mathcal{E}} \leq \varepsilon. \tag{3.21}$$

Proof. From Lemma 3.3,

$$\begin{aligned}
\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)}^2 &\geq \|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)}^2 - \sigma^2 b_1(\alpha)\int_0^t V(\rho_s)ds - \lambda b_2(\alpha)\int_0^t V(\rho_t)^{\frac{1}{2}} \\
&\geq \|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)}^2 - \sigma^2 b_1(\alpha)\overline{\mathcal{V}}_T\, t - \lambda b_2(\alpha)\overline{\mathcal{V}}_T^{\frac{1}{2}}\, t \geq \\
&\geq \|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)}^2 - \sigma^2 b_1(\alpha)\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)}^2 t - \lambda b_2(\alpha)\|\omega_{\alpha,\mathcal{E}}\|_{L^1(\rho_0)}^2 t,
\end{aligned}$$

where we used that condition (3.20) implies:

$$\overline{\mathcal{V}}_T, \overline{\mathcal{V}}_T^{\frac{1}{2}} \leq \|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)}^2.$$

We then have the following estimate for any $t \in [0,T]$:

$$-\frac{1}{\alpha}\log\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} \leq -\frac{1}{\alpha}\log\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)} - \frac{1}{2\alpha}\log(1-\sigma^2 b_1(\alpha)t - \lambda b_2(\alpha)t).$$

By the Laplace principle, (1.8), it holds that for any α greater than a certain α_0,

$$-\frac{1}{\alpha}\log\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_0)} - \underline{\mathcal{E}} \leq \frac{\varepsilon}{2}.$$

Moreover, as $b_1, b_2 \longrightarrow 0$ as $\alpha \to \infty$, see Lemma 3.3,

$$-\frac{1}{2\alpha}\log(1-\sigma^2 b_1(\alpha)t - \lambda b_2(\alpha)t) \leq \frac{\varepsilon}{2}$$

if $\alpha > \alpha_1$ for a certain $\alpha_1 > \alpha_0$ sufficiently large. We can, thus, conclude that

$$-\frac{1}{\alpha}\log\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} - \underline{\mathcal{E}} \leq \varepsilon \quad \forall\, t \in [0,T]\,.$$

□

We now estimate the first term in (3.18), by adapting the technique used in the proof of Theorem 3.1.

Proposition 3.2. *The following inequality holds for any $\rho_t \in \mathcal{P}_2(\Gamma)$:*

$$\left| -\frac{1}{\alpha}\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} - \mathcal{E}\big(\Pi(E(\rho_t))\big)\right| \leq 2c_1 C_{\alpha,\mathcal{E}} V(\rho_t)^{\frac{1}{2}}. \tag{3.22}$$

Proof. The dual representation of 1-Wasserstein distance W_1, see for instance [36, Theorem 4.12, Chapter 1], states that if $\mu, \nu \in \mathcal{P}$ have bounded support, then the 1-Wasserstein distance can be equivalently expressed in terms of the dual formulation

$$W_1(\mu,\nu) := \sup\left\{ \int_{\mathbb{R}^d} f(v)d(\mu-\nu)(v) \,\middle|\, f \in \mathrm{Lip}(\mathbb{R}^d), \mathrm{Lip}(f) \leq 1 \right\}.$$

We can make an estimate as the following

$$\begin{aligned}\Big|\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} - \omega_{\mathcal{E}}^{\alpha}\big(\Pi(E(\rho_t))\big)\Big| &= \left|\int e^{-\alpha\mathcal{E}(v)} d\big(\rho_t(v) - \delta\big(\Pi(E(\rho_t))\big)(v)\big)\right| \leq \\ &\leq \alpha e^{-\alpha\underline{\mathcal{E}}}\|\nabla\mathcal{E}\|_{\infty} W_1\big(\rho_t, \delta\big(\Pi(E(\rho_t))\big)\big) \leq \\ &\leq \alpha c_1 e^{-\alpha\underline{\mathcal{E}}} W_2\big(\rho_t, \delta\big(\Pi(E(\rho_t))\big)\big) \leq \\ &\leq 2\alpha c_1 e^{-\alpha\underline{\mathcal{E}}} V(\rho_t)^{\frac{1}{2}}.\end{aligned}$$

Where the last inequality follows from:

$$\begin{aligned}W_2(\rho_t, \delta_{\Pi(E(\rho_t))})^2 &\leq \int |v - \Pi(E(\rho_t))|^2 d\rho_t = \\ &= \int |v - E(\rho_t) + E(\rho_t) - \Pi(E(\rho_t))|^2 d\rho_t \\ &\leq 2V(\rho_t) + \gamma(E(\rho_t))^2 \leq 4V(\rho_t).\end{aligned}$$

Above we used the definition of γ as the signed distance, $|\gamma(w)| = \mathrm{dist}(w, \Gamma)$ for all $w \in \mathbb{R}^d$, which implies, by considering $w = E(\rho_t)$,

$$|\gamma(E(\rho_t))| \leq |E(\rho_t) - v|$$

for any $v \in \Gamma$. Therefore, we get

$$\gamma(E(\rho_t))^2 \leq \int |E(\rho_t) - v|^2 d\rho_t(v) = 2V(\rho_t)\,.$$

Finally, we obtain the desired estimate in terms of $V(\rho_t)$.

$$\begin{aligned}\left|-\frac{1}{\alpha}\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} - \mathcal{E}\big(\Pi(E(\rho_t))\big)\right| &= \left|-\frac{1}{\alpha}\Big(\log\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} - \log\omega_{\mathcal{E}}^{\alpha}\big(\Pi(E(\rho_t))\big)\Big)\right| \\ &\leq \frac{e^{\alpha\overline{\mathcal{E}}}}{\alpha}\left|\|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} - \omega_{\mathcal{E}}^{\alpha}\big(\Pi(E(\rho_t))\big)\right| \\ &\leq 2c_1 C_{\alpha,\mathcal{E}} V(\rho_t)^{\frac{1}{2}}.\end{aligned}$$

□

We will now prove Theorem 3.2.

Proof of Theorem 3.2. As above, $|E(\rho_t) - \Pi(E(\rho_t))| = |\gamma(E(\rho_t)|$ implies

$$\begin{aligned} |E(\rho_t) - \Pi(E(\rho_t))| \leq & \int |E(\rho_t) - v| d\rho_t \\ \leq & \sqrt{2} V(\rho_t)^{\frac{1}{2}} \end{aligned}$$

by Hölder's inequality.

Moreover, by collecting the results from Propositions 3.1 and 3.2, there exists α large enough such that

$$\begin{aligned} |\mathcal{E}(\Pi(E(\rho_t))) - \underline{\mathcal{E}}| \leq & \left| \mathcal{E}(\Pi(E(\rho_t))) - \frac{-1}{\alpha} \log \|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} \right| + \left| \frac{-1}{\alpha} \log \|\omega_{\mathcal{E}}^{\alpha}\|_{L^1(\rho_t)} - \underline{\mathcal{E}} \right| \\ \leq & \, 2 c_1 C_{\alpha, \mathcal{E}} V(\rho_t)^{\frac{1}{2}} + \varepsilon. \end{aligned}$$

We can, then, conclude that there exists a constant C that depends only on C_0, c_1, β for which it holds

$$\begin{aligned} |E(\rho_t) - v^*| \leq & |E(\rho_t) - \Pi(E(\rho_t))| + |\Pi(E(\rho_t)) - v^*| \\ \leq & |E(\rho_t) - \Pi(E(\rho_t))| + C_0 |\mathcal{E}(\Pi(E(\rho_t))) - \underline{\mathcal{E}}|^{\beta} \\ \leq & \sqrt{2} V(\rho_t)^{\frac{1}{2}} + C(C_0, c_1, \beta) \left(C_{\alpha, \mathcal{E}}^{\beta} V(\rho_t)^{\frac{\beta}{2}} + \varepsilon^{\beta} \right). \end{aligned}$$

□

Theorem 3.2 states that, provided the parameter α is sufficiently large, if the distribution ρ_t concentrates, it concentrates around a minimizer v^* of $\mathcal{E}$. Therefore, in the next section, we analyze the behavior of the variance as ρ_t evolves.

3.3 Variance decay

As Theorem 3.2 shows, the variance $V(\rho_t)$ not only gives a measure on how concentrated ρ_t is, but also on how well $E(\rho_t)$ approximates a global

minimizer v^*. For this reason, being able to bound and to show the decay of $V(\rho_t)$ is a key aspect if our goal is to provide convergence guarantees for the CBO method around a minimizer v^*.

In this section, we will try to understand why the geometry of Γ makes this task extremely hard and we will computationally show how the choice of the parameters α, λ, σ influences the behavior of ρ_t. We remark that, in the specific case where $\Gamma = \mathbb{S}^{d-1}$, it is possible to show an exponential decay of the $V(\rho_t)$ thanks to the simple geometry of the sphere [17].

We start by demonstrating the following estimate on the derivative of $V(\rho_t)$.

Proposition 3.3. *Let ρ_t be the solution of (3.1) with initial datum ρ_0, it holds for any $t > 0$*

$$\frac{d}{dt}V(\rho_t) \leq -\lambda \int_\Gamma |P(v)(v - E(\rho_t))|^2 d\rho_t(v) + 8\lambda c_\gamma V(\rho_t)^{\frac{3}{2}} + 8\sigma^2 V(\rho_t) C_{\alpha,\mathcal{E}}(c_\gamma + d - 1)\,. \quad (3.23)$$

Remark 3.3. *Even if Proposition 3.3 does not imply the variance decay, it gives interesting insights into the evolution of $V(\rho_t)$. First of all, it shows analytically that, in order to have consensus, we need $\lambda \gg \sigma$. Indeed, only one of the three terms comprising the RHS of (3.23) is negative and depends on λ. The term that depends on σ is positive and, hence, does not contribute to the variance decay.*

Similarly, it suggests that a large α inhibits the variance zeroing. We will attempt to computationally investigate these aspects at the end of this section.

Lemma 3.4. *Let P be defined as $P(v) = I - \nabla\gamma(v)\nabla\gamma(v)^t$. For all $u, v \in \Gamma$ and $w \in \mathbb{R}^d$, it holds*

$$\|P(u) - P(v)\|_2 \leq 2c_\gamma |u - v| \quad (3.24)$$

where with $\|.\|_2$ we denote the operator norm.

Proof of Proposition 3.3. Let us first calculate the derivative of $V(\rho_t)$ through the Itô's formula (2.6):

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}\int v^2 d\rho_t(v) &= -\lambda\int v\cdot P(v)(v-v_{\alpha,\mathcal{E}})d\rho_t + \sigma^2\int v\cdot|v-v_{\alpha,\mathcal{E}}|^2\Delta\gamma(v)\nabla\gamma(v)d\rho_t \\ &\quad + \sigma^2(d-1)\int|v-v_{\alpha,\mathcal{E}}|^2 d\rho_t \\ \frac{d}{dt}\frac{1}{2}E(\rho_t)^2 &= E(\rho_t)\frac{d}{dt}E(\rho_t) \\ &= E(\rho_t)\bigg(-\lambda\int P(v)(v-v_{\alpha,\mathcal{E}})d\rho_t + \sigma^2\int|v-v_{\alpha,\mathcal{E}}|^2\Delta\gamma(v)\nabla\gamma(v)\rho_t\bigg)\end{aligned}$$

which implies

$$\begin{aligned}\frac{d}{dt}V(\rho_t) &= \frac{d}{dt}\frac{1}{2}\int v^2 - E(\rho_t)^2 d\rho_t \\ &= -\lambda\int(v-E(\rho_t))P(v)(v-v_{\alpha,\mathcal{E}})d\rho_t + \sigma^2\int(v-E(\rho_t))\nabla\gamma(v)|v-v_{\alpha,\mathcal{E}}|^2\Delta\gamma(v)d\rho_t \\ &\quad + \sigma^2(d-1)\int|v-v_{\alpha,\mathcal{E}}|^2 d\rho_t =: \mathrm{I}_\lambda + \mathrm{I}_\sigma\,.\end{aligned}$$

Let us first consider I_λ, which we rewrite in the following way:

$$\begin{aligned}\mathrm{I}_\lambda &= -\lambda\int(v-E(\rho_t))^t P(v)(v-v_{\alpha,\mathcal{E}})d\rho_t \\ &= -\lambda\int(v-E(\rho_t))^t P(v)(v-E(\rho_t))d\rho_t - \lambda\int(v-E(\rho_t))^t P(v)(E(\rho_t)-v_{\alpha,\mathcal{E}})d\rho_t \\ &\le -\lambda\int|P(v)(v-E(\rho_t))|^2 d\rho_t - \lambda(E(\rho_t)-v_{\alpha,\mathcal{E}})\int P(v)(v-E(\rho_t))d\rho_t\,.\end{aligned}$$

where we used that $P(v) = P(v)P(v)$. In fact, keeping in mind that $|\nabla\gamma(v)| = 1$ for all $v\in\mathbb{R}^d$,

$$\begin{aligned}P(v)P(v) &= (I-\nabla\gamma(v)\nabla\gamma(v)^t)(I-\nabla\gamma(v)\nabla\gamma(v)^t) \\ &= I - 2\nabla\gamma(v)\nabla\gamma(v)^2 + \nabla\gamma\nabla\gamma^t\nabla\gamma\nabla\gamma^t \\ &= I - \nabla\gamma(v)\nabla\gamma(v)^t = P(v)\,.\end{aligned}$$

Now, we know that

$$\int P(E(\rho_t))(v - E(\rho_t))d\rho_t = 0$$

and, hence, by Lemma 3.4 and the Cauchy-Schwarz inequality it holds

$$\begin{aligned} -\int P(v)(v - E(\rho_t))d\rho_t &= -\int \left(P(v) - P(E(\rho_t))\right)(v - E(\rho_t))d\rho_t \\ &\leq \int \|P(v) - P(E(\rho_t))\|_2 |v - E(\rho_t)|d\rho_t \\ &\leq 2c_\gamma \int |v - E(\rho_t)|^2 d\rho_t = 4c_\gamma V(\rho_t)\,. \end{aligned}$$

Moreover, by Jensen's inequality and Lemma 3.1 we can bound $|E(\rho_t) - v_{\alpha,\mathcal{E}}|$ as follows

$$|E(\rho_t) - v_{\alpha,\mathcal{E}}| \leq 2C_{\alpha,\mathcal{E}} V(\rho_t)^{\frac{1}{2}}\,.$$

Finally, we obtain an upper bound for I_λ

$$\mathrm{I}_\lambda \leq -\lambda \int |P(v)(v - E(\rho_t))|^2 d\rho_t + 8\lambda c_\gamma C_{\alpha,\mathcal{E}} V(\rho_t)^{\frac{3}{2}}\,. \tag{3.25}$$

The second integral depends on σ^2 and can be estimated through Cauchy-Schwarz inequality and Lemma 3.1

$$\begin{aligned} \mathrm{I}_\sigma &= \sigma^2 \int (v - E(\rho_t))\nabla\gamma(v)|v - v_{\alpha,\mathcal{E}}|^2 \Delta\gamma(v)d\rho_t + \sigma^2(d-1)\int |v - v_{\alpha,\mathcal{E}}|^2 d\rho_t \\ &\leq \sigma^2 \int |v - E(\rho_t)||v - v_{\alpha,\mathcal{E}}|^2 c_\gamma d\rho_t + \sigma^2(d-1)\int |v - v_{\alpha,\mathcal{E}}|^2 d\rho_t \\ &\leq \sigma^2 \left(8C_{\alpha,\mathcal{E}} V(\rho_t)^{\frac{3}{2}} + 4(d-1)C_{\alpha,\mathcal{E}} V(\rho_t)\right) \\ &\leq 8\sigma^2 C_{\alpha,\mathcal{E}}(c_\gamma + d - 1). \end{aligned}$$

We note that I_σ goes to zero as $\sigma \to 0$. This concludes the proof. □

Proposition 3.3 illustrates how the analysis is made more complicated by considering the dynamics to be constrained on a hypersurface Γ. In detail, this is attributed to the singularity of the projection matrix $P(v)$.

Indeed, let us focus on a simple example, where Γ is the Torus. In these settings, projection $\Pi(w)$ is well-defined for any $w \in \mathbb{R}^d$ and, by simple geometric arguments, it is possible to show that

$$P(\Pi(E(\rho_t)))\left(\Pi(E(\rho_t)) - E(\rho_t)\right) = 0. \tag{3.26}$$

We remark that $\Pi(E(\rho_t)) \in \Gamma$ and so it is *not* possible to directly claim that there exists $\delta > 0$ such that

$$|P(v)(v - E(\rho_t))|^2 \geq \delta |v - E(\rho_t)|^2$$

for all $v \in \Gamma$ and consequently that

$$-\int |P(v)(v - E(\rho_t))|^2 d\rho_t \leq -\delta \int |v - E(\rho_t)|^2 d\rho_t. \tag{3.27}$$

In practice, this means that, as the system evolves, if a particle belongs to a region of Γ such that $(v - E(\rho_t))$ is close to be a singular value of $P(v)$, then the step size of the particle will be particularly small. We conjecture that, in pathological situations, a particle could also be captured in this region and, hence, prevent the mechanism from creating a complete consensus around a minimizer.

In Figure 3.3 we plot the quantity

$$\delta(v) = \frac{|P(v)(v - E(\rho_t))|^2}{|v - E(\rho_t)|^2} \tag{3.28}$$

for $v \in \Gamma = \mathbb{T}^2$ and $E(\rho_t) = (0, 1, 0.5)^t$. The figure shows that in large regions of the Torus, $\delta(v)$ is small and, hence, $P(v)$ has a big impact on the dynamics.

Remark 3.4. *As already mentioned, [17] shows that, when considering* $\Gamma = \mathbb{S}^{d-1}$*, it is possible to prove the exponential decay of the Variance, at the price of a correction term of the type* $\mathcal{O}(\delta^{\frac{d-2}{4}})$*. More formally, if* $\delta > 0$ *and the parameters* $\{\lambda, \alpha, \sigma\}$ *are carefully chosen, there exist* $\theta > 0$ *(*$\theta \sim \lambda\delta$*) and*

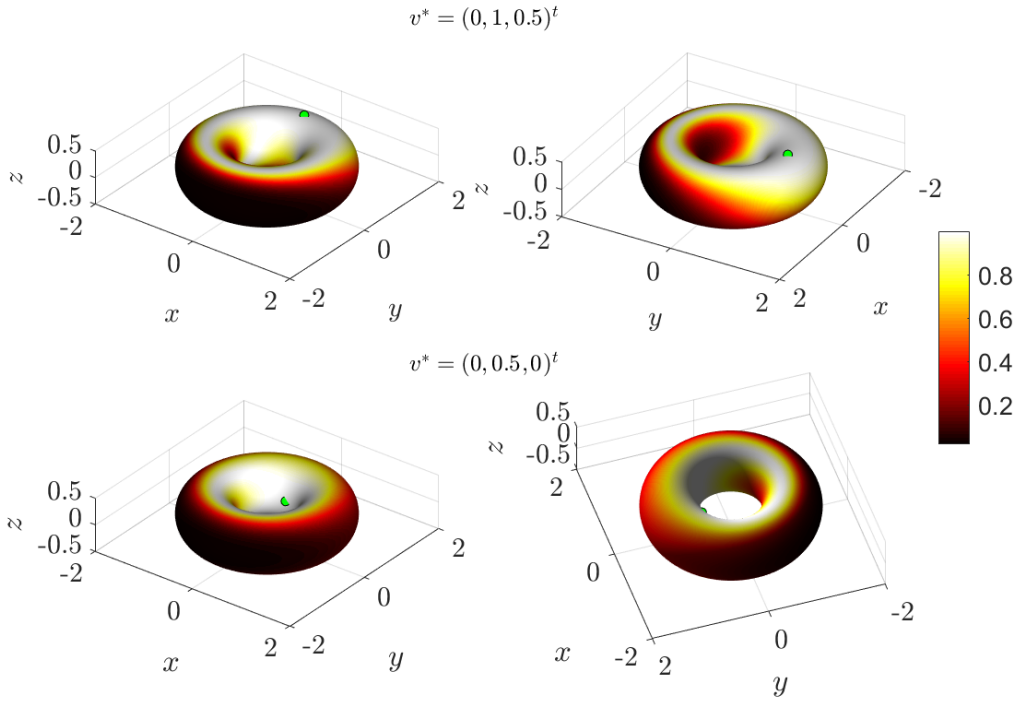

Figure 3.1: Plot of the quantity $\delta(v) = |P(v)(v - E(\rho_t))|^2/|v - E(\rho_t)|^2$, which gives a measure of the impact of the operator $P(\cdot)$ on the dynamics. Indeed, if a particle belongs to a region where $\delta(v)$ is small, $P(\cdot)$ reduces the particle motion. $E(\rho_t)$ is considered to be the point $(0, 1, 0.5)^t$.

$C > 0$ such that:

$$\frac{d}{dt}V(\rho_t) \leq -\theta V(\rho_t) + C\delta^{\frac{d-2}{4}}$$

for $t \in [0, T]$. Hence, by Grönwall's inequality one can conclude that

$$V(\rho_t) \leq V(\rho_0)e^{-\theta t} + \frac{C\delta^{\frac{d-2}{4}}}{\theta}$$

for any t up to the time horizon $T > 0$.

We conclude the chapter by briefly analyzing the influence of the parameters α, λ, σ on the variance decay. As we have already discussed, Proposition 3.3 suggests that, in order to create consensus, we need

$$\lambda \gg \sigma^2 C_{\alpha,\mathcal{E}}. \tag{3.29}$$

Naturally, we would expect that a large drift parameter λ boosts the consensus mechanism, while a large stochastic component, i.e. σ large, inhibits it. In figure 3.3 we recognize this behavior. Namely, large values of λ drastically improve the variance decay. At the same time, when large values of σ are considered, we note several oscillations in the evolution of the variance, indicative of an unstable system.

On the other hand, condition (3.29) suggests that large values of α have a negative impact on the variance decay. In Figure 3.3, we consider different values of α and we plot both the variance evolution (left) and approximation of the quantity

$$V(\rho_t)^* := \int |v - v^*| d\rho_t(v)$$

which gives as a measure of the concentration of ρ_t around the minimizer.

We note that even large values of α do not impede a fast variance decay. A possible explanation could be that estimates of the type of Lemma 3.1,

$$\int |v - v_{\alpha,\mathcal{E}}|^2 d\rho_t(v) \leq 2C_{\alpha,\mathcal{E}}V(\rho_t)$$

are inaccurate, since the constant $C_{\alpha,\mathcal{E}}$ increases exponentially as $\alpha \to \infty$.

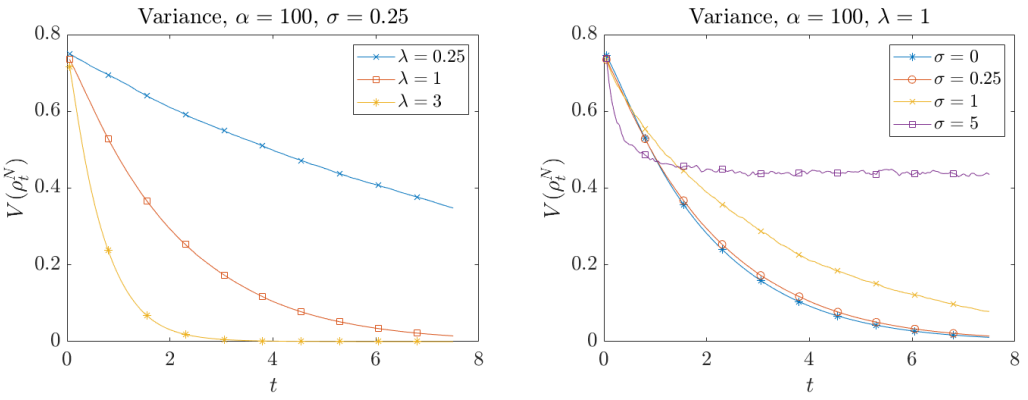

Figure 3.2: Plots of the variance decay for different values of λ (left), σ (right). The solution of the mean-field PDE is approximated by considering a stochastic system of 10^4 particle uniformly distributed on $\mathbb{T}^2$ at the time $t = 0$, with $\Delta t = 0.05$ up to $T = 7.5$. The variance is defined as $V(\rho_t^N)$ where ρ_t^N is the empirical measure. We plot the result of a single simulation in order to underline the presence of variance oscillations.

Nevertheless, large values of α slightly improve the convergence around the minimizer. As we will see in Chapter 4, this behavior is stronger when we consider the microscopic system and not its mean-field approximation.

3.4 Proofs of auxiliary lemmas

Proof of Lemma 3.1. Thanks to the Jensen's inequality, we obtain

$$\begin{aligned}\int_\Omega |v - v_{\alpha,\mathcal{E}}|^2 \, d\rho_t(v) &= \int_\Omega \left| v - \int_\Omega u \frac{e^{-\alpha\mathcal{E}(u)}}{\|\omega_{\mathcal{E}}^\alpha\|_{L^1(\rho_t)}} d\rho_t(u) \right|^2 d\rho_t(v) \\ &\leq \int_\Omega |v-u|^2 \frac{e^{-\alpha\mathcal{E}(u)}}{\|\omega_{\mathcal{E}}^\alpha\|_{L^1(\rho_t)}} d\rho_t(u) d\rho_t(v).\end{aligned}$$

We now employ the inequalities $e^{-\alpha\mathcal{E}(v)} \leq e^{-\underline{\mathcal{E}}}$ and $\|\omega_{\mathcal{E}}^\alpha\|_{L^1(\rho_t)} \geq e^{-\alpha\underline{\mathcal{E}}}$ and conclude that

$$\int_\Omega |v - v_{\alpha,\mathcal{E}}|^2 \, d\rho_t(v) \leq C_{\alpha,\mathcal{E}} \int_\Omega |u-v|^2 d\rho_t(u) d\rho_t(v) = 2C_{\alpha,\mathcal{E}} V(\rho_t). \quad (3.30)$$

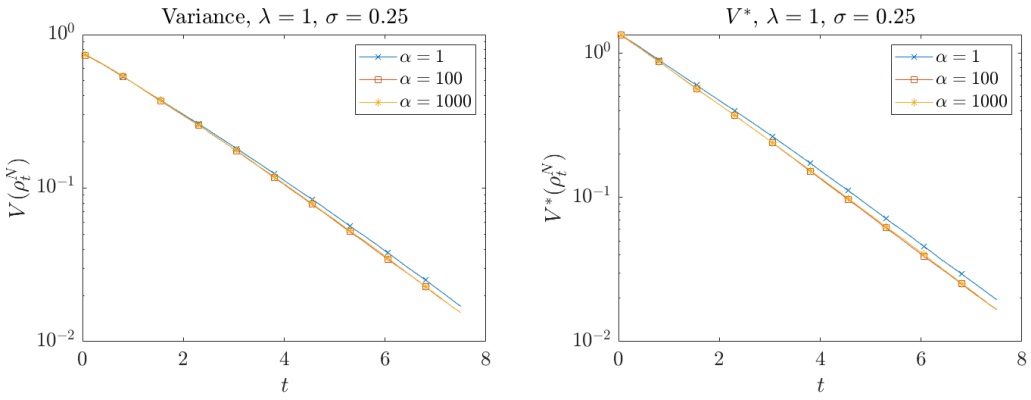

Figure 3.3: Plots of the variance decay and of $V^*(\rho_t)$ for different values of α. The solution of the mean-field PDE is approximated by considering a stochastic system of 10^4 particle, with $\Delta t = 0.05, T = 7.5$. $V(\rho_t)$ and $V^*(\rho_t)$ are approximated by considering ρ_t^N, the empirical measure.The results are the average of 100 simulations.

□

We remark that from this estimate, it directly follows by Hölder's inequality

$$\int_\Omega |v - v_{\alpha,\mathcal{E}}| \, d\rho_t(v) \leq 2C_{\alpha,\mathcal{E}} V(\rho_t)^{\frac{1}{2}}. \tag{3.31}$$

Proof of Lemma 3.4. Let $w \in \mathbb{R}^d$, we have that

$$\begin{aligned}
\big|(P(v) - P(u))w\big| &= \big|\big(\nabla\gamma(u)\nabla\gamma(u)^t - \nabla\gamma(v)\nabla\gamma(v)^t\big)w\big| \\
&= \big|\big(\nabla\gamma(u)\nabla\gamma(u)^t - \nabla\gamma(v)\nabla\gamma(u)^t\big)w\big| \\
&\quad + \big|\big(\nabla\gamma(v)\nabla\gamma(u)^t - \nabla\gamma(v)\nabla\gamma(v)^t\big)w\big| \\
&= \big|\big(\nabla\gamma(u) - \nabla\gamma(v)\big)\nabla\gamma(u)^t w\big| + \big|\nabla\gamma(v)\big(\nabla\gamma(u)^t - \nabla\gamma(v)^t\big)w\big| \\
&\leq 2\big|\big(\nabla\gamma(u) - \nabla\gamma(v)\big)^t w\big| \\
&\leq 2 \sup \|\nabla^2(\xi)\|_2 \, |u - v||w| \leq 2c_\gamma |u - v||w|.
\end{aligned}$$

□

Proof of Lemma 3.3. By Itô's formula (2.6), it holds

$$\begin{aligned}\frac{d}{dt}\|\omega_\varepsilon^\alpha\|_{L_1(\rho_t)} &= \int_{\mathbb{R}^d} \nabla\omega_\varepsilon^\alpha \cdot \Big(-\lambda P(v)(v - v_{\alpha,\mathcal{E}})(\rho_t) - \frac{\sigma^2}{2}(v - v_{\alpha,\mathcal{E}}(\rho_t))^2 \Delta\gamma(v)\nabla\gamma(v)\Big) d\rho_t(v) \\ &\qquad + \int_{\mathbb{R}^d} \frac{\sigma^2}{2}(v - v_{\alpha,\mathcal{E}}(\rho_t))^2 \nabla^2\omega_\varepsilon^\alpha : P(v) d\rho_t(v) \\ &= -\int_{\mathbb{R}^d} \nabla\omega_\varepsilon^\alpha \cdot (\lambda P(v)(v - v_{\alpha,\mathcal{E}})(\rho_t)) d\rho_t(v) \\ &\qquad + \int_{\mathbb{R}^d} -\nabla\omega_\varepsilon^\alpha \cdot \frac{\sigma^2}{2}(v - v_{\alpha,\mathcal{E}}(\rho_t))^2 \Delta\gamma(v)\nabla\gamma(v) \\ &\qquad\qquad + \frac{\sigma^2}{2}(v - v_{\alpha,\mathcal{E}}(\rho_t))^2 \nabla^2\omega_\varepsilon\alpha : P(v) d\rho_t(v) \;= \\ &=: \mathrm{I}_\lambda + \mathrm{I}_\sigma\,.\end{aligned}$$

We now separately estimate from below these two integrals. By Lemma 3.2, I_σ can be rewritten as the following

$$\begin{aligned}\mathrm{I}_\sigma &= \int_{\mathbb{R}^d} \frac{\sigma^2}{2}(v - v_{\alpha,\mathcal{E}})^2 \Big(-\nabla\omega_{\mathcal{E}}^\alpha \cdot \Delta\gamma\nabla\gamma + \nabla^2\omega_{\mathcal{E}}^\alpha : P(v)\Big) d\rho_t \\ &= \frac{\sigma^2}{2}\int_{\mathbb{R}^d} |v - v_{\alpha,\mathcal{E}}|^2 e^{-\alpha\mathcal{E}} \Big(\alpha\nabla\mathcal{E}\cdot\nabla\gamma\Delta\gamma + (\alpha^2|\nabla\mathcal{E}|^2 - \alpha\Delta\mathcal{E}) \\ &\qquad - \nabla\gamma(v)\otimes\nabla\gamma(v) : [\alpha\nabla\mathcal{E}\otimes\nabla\mathcal{E} + \nabla^2\mathcal{E}]\Big) d\rho_t \\ &\geq \frac{\sigma^2}{2} e^{-\alpha\underline{\mathcal{E}}} \int_{\mathbb{R}^d} |v - v_{\alpha,\mathcal{E}}|^2 \Big(-\alpha c_1|\Delta\gamma(v)| - \alpha c_2 - \alpha^2 c_1 - \alpha c_2\Big) d\rho_t \\ &\geq -\alpha\frac{\sigma^2}{2} e^{-\alpha\underline{\mathcal{E}}}(c_1 c_\gamma + 2c_2 + \alpha c_1) \int_{\mathbb{R}^d} |v - v_{\alpha,\mathcal{E}}|^2 d\rho_t\end{aligned}$$

where we used that $\|\nabla\mathcal{E}\|_\infty \leq c_1$, $|\Delta\gamma(v)| \leq c_\gamma$ and that $|\nabla\gamma(v)| = 1$ for all

v. By the Jensen's inequality, we obtain

$$\begin{aligned}
\mathrm{I}_\sigma &\geq -\alpha\frac{\sigma^2}{2}e^{-\alpha\underline{\mathcal{E}}}(c_1c_\gamma + 2c_2 + \alpha c_1)\int_{\mathbb{R}^d}|v - E(\rho_t)|^2\frac{e^{-\alpha\mathcal{E}(v)}}{\|\omega_{\alpha,\mathcal{E}}\|_{L^1(\rho_t)}}d\rho_t(v) \\
&\geq -\alpha\frac{\sigma^2}{2}e^{-2\alpha\underline{\mathcal{E}}}(c_1c_\gamma + 2c_2 + \alpha c_1)\frac{1}{\|\omega_{\alpha,\mathcal{E}}\|_{L^1(\rho_t)}}\int_{\mathbb{R}^d}|v - E(\rho_t)|^2 d\rho_t(v) \\
&\geq -\sigma^2 b_1(c_\gamma, \alpha, c_1, c_2, \underline{\mathcal{E}})\frac{V(\rho_t)}{\|\omega_{\alpha,\mathcal{E}}\|_{L^1(\rho_t)}} \qquad (3.32)
\end{aligned}$$

with $b_1 \longrightarrow 0$ as $\alpha \to 0$, $b_1 \geq 0$.

Furthermore, we can estimate I_λ as the following:

$$\begin{aligned}
\mathrm{I}_\lambda &= \lambda\alpha\int e^{-\alpha\mathcal{E}}\nabla\mathcal{E}\cdot P(v)(v - v_{\alpha,\mathcal{E}})d\rho_t \\
&= \lambda\alpha\iint\frac{e^{-\alpha(\mathcal{E}(u)+\mathcal{E}(v))}}{\|\omega_{\alpha,\mathcal{E}}\|_{L^1(\rho_t)}}\nabla\mathcal{E}\cdot P(v)(v-u)\, d\rho_t(u)d\rho_t(v) \\
&\geq\ -\lambda\alpha c_1\frac{e^{-2\alpha\underline{\mathcal{E}}}}{\|\omega_{\alpha,\mathcal{E}}\|_{L^1(\rho_t)}}\iint|v-u|d\rho_t d\rho_t \\
&\geq -\lambda\alpha c_1\frac{e^{-2\alpha\underline{\mathcal{E}}}}{\|\omega_{\alpha,\mathcal{E}}\|_{L^1(\rho_t)}}2V(\rho_t)^{\frac{1}{2}} = -\lambda b_2(\alpha, c_1, \underline{\mathcal{E}})\frac{V(\rho_t)^{\frac{1}{2}}}{\|\omega_{\alpha,\mathcal{E}}\|_{L^1(\rho_t)}} \qquad (3.33)
\end{aligned}$$

Now we combine the inequalities, (3.32) and (3.33), and conclude

$$\frac{1}{2}\frac{d}{dt}\|\omega_{\alpha,\mathcal{E}}\|^2_{L^1(\rho_t)} = \|\omega_{\alpha,\mathcal{E}}\|_{L^1(\rho_t)}\frac{d}{dt}\|\omega_{\alpha,\mathcal{E}}\|_{L^1(\rho_t)} \geq -\sigma^2 b_1 V(\rho_t) - \lambda b_2 V(\rho_t)^{\frac{1}{2}}.$$

□

Chapter 4

Implementation and Tests

In this chapter, we report an example application of the Consensus-Based Optimization method on the stochastic Kuramoto-Vicsek (sKV) model for constrained optimization on hypersurfaces.

Firstly, we present the discretization scheme of the sKV system and discuss some practical aspects of the implementation. Then, we present the algorithm and employ it for the optimization of two benchmark functions on the three-dimensional torus, namely the Rastrigin and the Ackley functions. We will examine the convergence rate of the method paying special attention to the the evolution of the empirical variance and the parameter choice.

To conclude, we present some practical implementations that have been proposed in [17] and [28] to speed up the algorithm in case of high dimensional functions optimization.

4.1 Discretization of the sKV system

We discuss the discretization of the sKV system in Itô's form

$$dV_t^i = -\lambda P(V_t^i)(V_t^i - V_t^{\alpha,\mathcal{E}})dt + \sigma|V_t^i - V_t^{\alpha,\mathcal{E}}|P(V_t^i)dB_t^i \\ - \frac{\sigma}{2}(V_t^i - V_t^{\alpha,\mathcal{E}})^2\Delta\gamma(V_t^i)\nabla\gamma(V_t^i)dt \quad (4.1)$$

with $V_t^i \in \Gamma$, $i = 1, \dots, N$ and

$$V_t^{\alpha,\mathcal{E}} = \sum_{j=1}^{N} \frac{V_t^i \omega_\alpha^{\mathcal{E}}(V_t^j)}{\sum_{i=1}^{N} \omega_\alpha^{\mathcal{E}}(V_t^i)} = v_{\alpha,\mathcal{E}}(\rho_t^N)\,. \tag{4.2}$$

A natural approach to the numerical solution of differential equations on manifold is by projection [20], hence, we consider a one-step size discretization of (4.1) followed by the projection operator Π. This class of schemes has the general form

$$\begin{cases} \tilde{V}_{n+1}^i = V_n^i + \Phi(\Delta t, V_n^i, V_{n+1}^i, \xi_n^i) \\ V_{n+1}^i = \Pi\left(\tilde{V}_{n+1}^i\right) \end{cases} \tag{4.3}$$

where the function $\Phi_\Gamma(\Delta, \cdot, \cdot, \xi_n^i) : \mathbb{R}^{2d} \to \mathbb{R}^d$ defines the method, $\Delta t > 0$ is the time step, $V_n^i \approx V_t^i|_{t=t^n}$, $t^n = n\Delta t$ and $\xi_n^i =$ are independent random variables. The operator Π is defined on a strip of width $\delta > 0$, Γ_δ, where δ is sufficiently small such that Π is well defined as:

$$\Pi : \Gamma_\delta \to \Gamma, \quad \Pi(\tilde{V}) = \operatorname*{argmin}_{V \in \Gamma} ||\tilde{V} - V||^2. \tag{4.4}$$

Hence, for the computation of $V_{n+1} = \Pi\left(\tilde{V}_{n+1}^i\right)$ we need to solve the constrained optimization problem

$$\min_{\mathbb{R}^d} |V_{n+1} - \tilde{V}_{n+1}| \quad \text{subject to} \quad \gamma(V_{n+1}) = 0 \tag{4.5}$$

which for an arbitrary hypersurface Γ could be a complex task, we refer to [20, Chapter 4] for more details regarding projection methods on manifolds. Nevertheless, $\Pi(\cdot)$ has a closed form definition in case of $\Gamma = \mathbb{S}^{d-1}$ or Γ being the torus $\mathbb{T}^2$:

$$\Pi_{\mathbb{S}^{d-1}} : \mathbb{R}^d \setminus \{0\} \to \mathbb{S}^{d-1}, \quad \Pi_{\mathbb{S}^{d-1}}(\tilde{V}) = \frac{\tilde{V}}{\|\tilde{V}\|}$$

and, respectively, for $\Gamma = \mathbb{T}^2$:

$$\Pi_{\mathbb{T}^2} : \mathbb{R}^3 \setminus \{z = 0\} \to \mathbb{T}^2,$$
$$\Pi_{\mathbb{T}^2}(\tilde{V}) = r\frac{\tilde{V} - RV_d}{\|\tilde{V} - RV_d\|} + RV_d, \quad \text{with} \quad V_d = \frac{\tilde{V} - \langle \tilde{V}, e_3 \rangle}{\|\tilde{V} - \langle \tilde{V}, e_3 \rangle\|}$$

where $r > 0$ is the inner radius, $R > 0$ is the outer radius, and $e_3 = (0, 0, 1)$.

In order to solve (4.1) on the torus, we will use the simple Euler-Maruyama scheme

$$\tilde{V}^i_{n+1} = V^i_n - \lambda \Delta t P(V^i_n)(V^i_n - V^{\alpha,\mathcal{E}}_n) + \sigma |V^i_n - V^{\alpha,\mathcal{E}}_n| P(V^i_n)\xi^i_n \\ - \Delta \frac{\sigma^2}{2}(V^i_n - V^{\alpha,\mathcal{E}}_n)^2 \nabla\gamma(V^i_n)\Delta\gamma(V^i_n). \quad (4.6)$$

In [17] it is shown that it is possible to construct implicit methods where the dynamics remains on the sphere without employing the projection $\Pi(\cdot)$, i.e.

$$V^i_{n+1} = V^i_n + \Phi(\Delta t, V^i_n, V^i_{n+1}, \xi^i_n), \quad \|V_{n+1}\| = \|V_n\| = 1. \quad (4.7)$$

This can be done by simply modifying the Euler-Maruyama method or by considering implicit methods of weak order higher than one, which preserves the solution norm. Due to the nonlinearity of the projection operator $P(\cdot)$, implicit methods require the solution of a large nonlinear system. This constitutes a serious problem because our aim is to design a scalable optimization algorithm. Therefore, a simple scheme like the one presented in (4.6) was considered in [17]. The scheme has to be followed by the projection $\Pi(\cdot)$.

We are now ready to present the algorithm and discuss some implementation aspects.

4.2 Algorithm and implementation

First, we highlight that the set of three computational parameters $\Delta t, \sigma, \lambda$ can be reduced by setting $\lambda = 1$ to obtain a scheme depending only on Δt

and σ^2. We define n_T to be the maximum number of iterations.

Starting from a set of parameters $\{\Delta t, \sigma, \alpha, N, n_T\}$, a given objective function $\mathcal{E}(\cdot)$ defined on Γ and the projection operator Π_Γ, the KV-CBO method is described in Algorithm 1.

Algorithm 1: KV-CBO on Γ

Input: $\Delta t, \sigma, \alpha, N, n_T$ and the functions $\mathcal{E}(\cdot)$, $\Pi_\Gamma(\cdot)$

Generate $V_0^i,\ i = 1, \dots, N$ sample vectors uniformly on Γ ;
for $n = 0$ **to** n_T **do**
  Generate ΔB_n^i, independent normal random vectors $\mathcal{N}(0, \Delta t)$;
  Compute $V_n^{\alpha,\mathcal{E}}$;
  if *consensus* **then**
    break
  end
  $\tilde{V}_{n+1}^i \leftarrow V_n^i - \lambda \Delta t P(V_n^i)(V_n^i - V_n^{\alpha,\mathcal{E}}) + \sigma |V_n^i - V_n^{\alpha,\mathcal{E}}| P(V_n^i) \Delta B_n^i -$
  $-\Delta \frac{\sigma^2}{2} (V_n^i - V_n^{\alpha,\mathcal{E}})^2 \nabla\gamma(V_n^i) \Delta\gamma(V_n^i)$;
  $V_n^i \leftarrow \Pi_\Gamma(\tilde{V}_{n+1}^i)$;
end

We note that the computational cost for a single time step of KV-CBO is $\mathcal{O}(N)$, which is the minimum cost to evolve a system of N particles, since $V_n^{\alpha,\mathcal{E}}$ is the same for all agents.

The stopping criterion depends on the way we define the consensus status. As proposed in [8, 17], for a given tolerance ε, a suitable condition is

$$\frac{1}{N} \sum_{i=1}^{N} |V_n^i - V_n^{\alpha,\mathcal{E}}| < \varepsilon, \tag{4.8}$$

or, alternatively, as in [28], for some a priori selected $p \geq 0$ we can check if

$$|V_{n+1}^{\alpha,\mathcal{E}} - V_{n-p}^{\alpha,\mathcal{E}}| \leq \varepsilon\,. \tag{4.9}$$

As we will discuss later, the computational parameters Δt, σ and α can in

practice be adaptively modified from step to step to improve the performance of the method.

We remark that the computation of $V_n^{\alpha,\mathcal{E}}$, point 4 of Algorithm 1, is crucial and that a straightforward evaluation using

$$V_n^{\alpha,\mathcal{E}} = \frac{1}{N_\alpha} \sum_{j=1}^{N} \omega_\alpha^{\mathcal{E}}(V_n^j) V_n^j \,, \quad N_\alpha := \sum_{j=1}^{N} \omega_\alpha^{\mathcal{E}}(V_n^j) \,, \tag{4.10}$$

where $\omega_\alpha^{\mathcal{E}}(V_n^j) = \exp(-\alpha \mathcal{E}(V_n^j))$, is generally unstable since for large values of $\alpha \gg 1$, the value of N_α is close to zero. On the other hand, the use of large values of α is essential for the performance of the method. A well-known way to overcome this issue is based on the following trick

$$\begin{aligned} \frac{\omega_\alpha^{\mathcal{E}}(V_n^i)}{N_\alpha} &= \frac{\omega_\alpha^{\mathcal{E}}(V_n^i)}{\sum_{j=1}^{N} \omega_\alpha^{\mathcal{E}}(V_n^j)} \cdot \frac{\omega_\alpha^{\mathcal{E}}(V_n^*)}{\omega_\alpha^{\mathcal{E}}(V_n^*)} \\ &= \frac{e^{-\alpha(\mathcal{E}(V_n^i) - \mathcal{E}(V_n^*))}}{\sum_{j=1}^{N} e^{-\alpha(\mathcal{E}(V_n^j) - \mathcal{E}(V_n^*))}} \end{aligned}$$

where

$$V_n^* := \operatorname*{argmin}_{V_n^i} \mathcal{E}(V) \tag{4.11}$$

is the location of the particle with the minimal function value in the current population. This ensures that for at least one particle $V_n^j = V_n^*$, we have $\mathcal{E}(V_n^j) - \mathcal{E}(V_n^*) = 0$ and hence, $\exp(-\alpha(\mathcal{E}(V_n^j) - \mathcal{E}(V_n^*))) = 1$. For the sum, this leads to $N_\alpha \geq 1$, so that the division does not induce a computational difficulty. In the simulations, we will always compute the weights by the above strategy. Note that the evaluation of (4.11) has a linear cost and does not have an impact on the asymptotic computational cost of the algorithm.

The computation of $V_n^{\alpha,\mathcal{E}}$ may be accelerated by using the random approach presented in [2]. The approach considers a random subset J_M of size $M < N$ of the indexes $\{1, \dots N\}$ and computes

$$V_n^{\alpha,\mathcal{E},J_M} = \frac{1}{N_\alpha^{J_M}} \sum_{j \in J_M} \omega_\alpha^{\mathcal{E}}(V_n^j) V_n^j \,, \quad N_\alpha^{J_M} := \sum_{j \in J_M} \omega_\alpha^{\mathcal{E}}(V_n^j) \,. \tag{4.12}$$

Similarly, the above computation can be stabilized by centering it to

$$V_n^{J_M,*} := \arg\min_{V_n^j, j\in J_M} \mathcal{E}(V)\,. \tag{4.13}$$

The random subset is typically chosen at each time step in the simulation.

As a further randomization variant, at each time step, we may partition particles into disjoint subsets $J_M^k, k = 1, \ldots, S$ of size M such that $SM = N$, and compute the evolution of each batch separately, see [24, 28] for more details. We note that, since the computational cost is already linear, these randomization techniques can accelerate the simulation process and, eventually, improve the particles exploration thanks to additional stochasticity, but cannot reduce the overall asymptotic cost $\mathcal{O}(N)$.

In the next section, we present computational experiments where we apply Algorithm 1 on a low dimensional optimization problem on the three-dimensional torus.

4.3 Computational experiments

We study the performance of the consensus-based optimization algorithm and investigate, in particular, how the choice of the parameters modifies the computational outcome. We employ two standard test cases from the optimization literature [23], namely the Ackley function:

$$\mathcal{E}_A(v) = -20\exp\left(-\frac{0.2}{\sqrt{d}}\|v - B\|\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(2\pi(v_i - B))\right) + 20 + e + C \tag{4.14}$$

and the Rastrigin function

$$\mathcal{E}_R(v) = \frac{1}{d}\sum_{i=1}^{d}\left[(x_i - B)^2 - 10\cos(2\pi(v_i - B)) + 10\right] + C, \tag{4.15}$$

where $d \in \mathbb{N}$ is the dimension of the search space and $B, C \in \mathbb{R}$ are constant shifts. Both functions attain multiple local minima but only one global

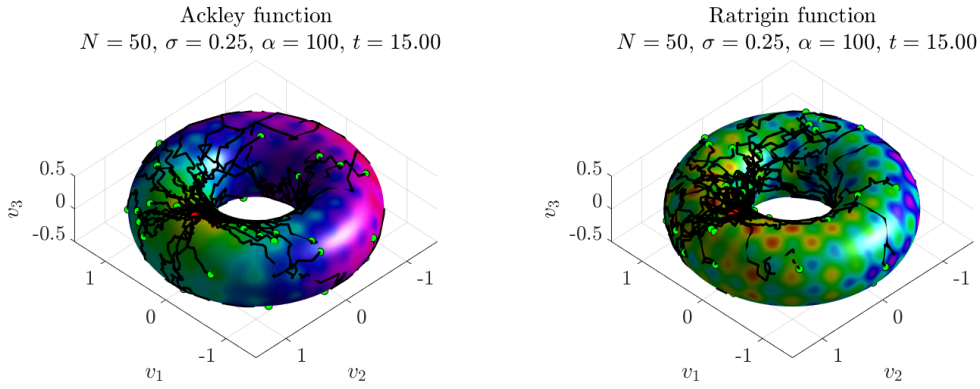

Figure 4.1: Particles trajectories along the simulation for the Ackley function (left) and the Rastrigin function (right) for $\alpha = 100$, $\sigma = 0.25$, $N = 50$, and $\Delta t = 0.05$. We notice that, compared to the Ackley function, the local minima of the Rastrign function on the torus are much closer to the global minimum, so it is harder to find its global minimum.

minimum.

We consider the constrained optimization problem on the torus

$$\min_{v \in \mathbb{T}^2} \mathcal{E}(v), \quad \mathbb{T}^2 = \{v \in \mathbb{R}^3 \,|\, \gamma(v) = 0\}$$

where, for $v = (v^1, \dots, v^d)$ and $R = 1$, $r = 0.5$:

$$\gamma(v) = \sqrt{(\sqrt{|v|^2 - (v^d)^2} - 1)^2 + (v^d)^2} - 0.5\,.$$

In all our simulations, we initialize the particles with a uniform distribution over the torus and we employ the simple Euler-Maruyama scheme with projection, by using Algorithm 1. We report in Figure 4.3 the particle trajectories during a simulation for $t \in [0, 15]$ in the case of $N = 50$, $\Delta t = 0.05$, $\sigma = 0.25$ and $\alpha = 100$. In both cases, the minimum is obtained at $v^* = (0, 1, 0.5)^t$.

Next, in Figure 4.3 we consider the convergence to consensus measured using two different values of α for the optimization of the Ackley function.

The results have been averaged 1000 times and in Table 4.3 we summarize the success rates. As considered in [17, 28, 33], we count as successful a run when at the final time it holds

$$\|V_{n_T}^{\alpha,\mathcal{E}} - v^*\|_\infty := \max_{k=1,\dots,d} |(V_{n_T}^{\alpha,\mathcal{E}})_k - (v^*)_k| \leq \frac{1}{4}\,. \tag{4.16}$$

We remark that condition (4.16) excludes $V_{n_T}^{\alpha,\mathcal{E}}$ from being any local minimizer in the benchmarks functions taken into account. We also compute the expected error in the computation of the minimum by considering time averages of $\|V^{\alpha,\mathcal{E}} - v^*\|_\infty$ and we report the quantity $\|V^{\alpha,\mathcal{E}} - v^*\|/d$ used in [17, 28, 33].

	$\alpha = 1$	$\alpha = 500$
Ackley	99.3%	100%
Rastrigin	73%	92.9%

Table 4.1: Success rates over 1000 runs.

As shown in Figure 4.2 and Table 4.1, where $\sigma = 0.25$, the accuracy of the computation of the minimum is higher when $\alpha = 500$ than when we consider $\alpha = 5$. Clearly, if $\alpha_1 > \alpha_2$, we expect to have $V^{\alpha_1,\mathcal{E}}(\rho^N)$ closer to v^* than $V^{\alpha_2,\mathcal{E}}(\rho^N)$, even for the same distribution ρ^N. The reason lies in the fact that

$$V^{\alpha,\mathcal{E}}(\rho^N) \longrightarrow \operatorname*{argmin}_{V^i} \mathcal{E}(V^i), \quad \text{as} \quad \alpha \to \infty.$$

Still, we also claim that the indicators show that a large α speeds up the concentration of all particles around the global minimizer and, consequently, the consensus mechanism.

Furthermore, as the minima of the Ackley function are more separated than the minima of the Rastrigin, we note that the converge is slower and less accurate when we try to optimize the latter.

We now investigate the influence of σ in the accuracy of the algorithm. For this purpose, we consider the Rastrigin function which presents several local minima close to the global minimum. In particular, we compare the success

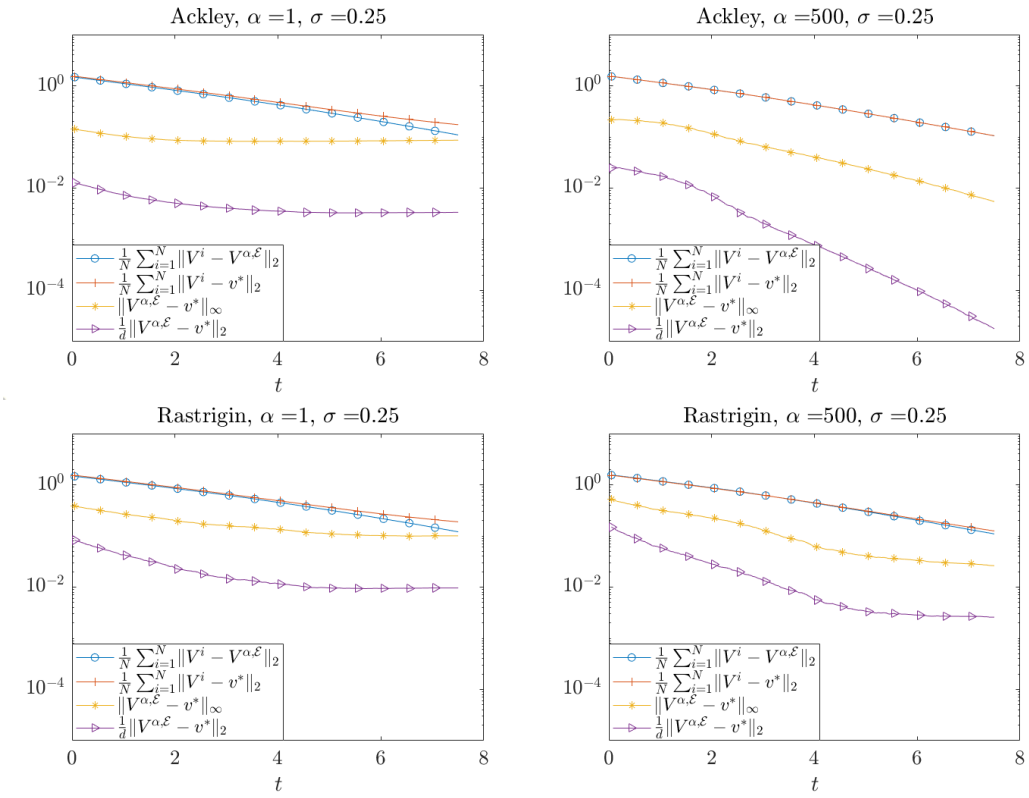

Figure 4.2: Behavior of various converge indicators in time for the Ackley and Rastrigin functions in the case of $d = 3$, $N = 50$, $\Delta t = 0.05$. The graphs show the accuracy of KV-CBO for $\sigma = 0.25$. We choose $\alpha = 1$ (left) and $\alpha = 500$ (right). The results have been averaged 1000 times, see Table 4.1 for the success rates.

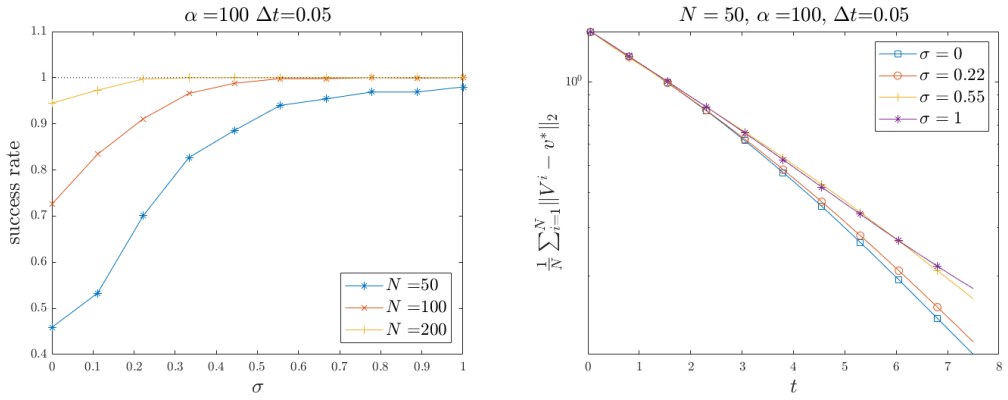

Figure 4.3: Optimization of the Rastrigin function, with parameters $\alpha = 100$, $\Delta t = 0.05$. On the left, the success rate as function of σ for three different values of N. The plot underlines how large values of σ are necessary in order to generate consensus among the global minimum in case of a small batch of particles. On the right, we plot the convergence rate for selected values of σ averaged only on successful runs. We notice that, as σ decreases, the variance decays faster when the run is successful.

rate of the algorithm for several values of σ when we consider three different numbers of particles N. Figure 4.3 (left) shows that having a sufficient number of particles is essential in order to create consensus around the global minimum. Nevertheless, if we consider larger values of σ, the success rate improves significantly even for a small particle batch. This suggests that boosting the stochastic component of the system could counterbalance the lack of particles.

A drawback of this approach is represented by a slower convergence rate when large values of σ are employed in the computation. Indeed, in Figure 4.3 we notice that, as σ decays, the system - in case of a successful run - generates consensus faster. This can be seen as evidence of the theoretical results we presented in Chapter 3, where the variance decay rate was shown to be larger for smaller values of σ.

To conclude, we present some corrections to SK-CBO that have been proposed in [17] to improve the performance of the algorithm.

4.4 Adaptive parameters

Our theoretical analysis of the mean-field approximation ρ_t, suggests that, once N is large, for σ small enough and α large enough, Algorithm 1 will converge near a global minimizer. One important aspect, therefore, concerns the choice of the parameters. In this section, we present some adaptive strategies that have been proposed in [17, 28] to improve the accuracy of the method and speed up the consensus generation.

The adaptation of hyperparameters in multi-particle optimization is a well-known problem, we refer to [14] for a complete discussion. In our case, we observed that large values of σ increase the success rate of the method, whereas small values of σ accelerate the consensus dynamic. One strategy, therefore, would be to start with a large σ and to progressively reduce it over time as a function of a suitable indicator of convergence, for example the average variance of the solution or the relative variation of $V^{\alpha,\mathcal{E}}$ over time. A simple adaptive strategy, proposed in [17], is to start from a value σ_0 and continue the computation while decreasing it as

$$\sigma_{n+1} = \frac{\sigma_n}{\tau}\,, \tag{4.17}$$

where τ is a constant.

Another technique that can be used to decrease σ is, for instance, the cooling strategy as in the Simulated Annealing approach [21]. [28] proposes to reduce σ independently of the solution behavior, as a function of the initial value σ_0 and the number of iterations. This corresponds to taking

$$\sigma_{n+1} = \frac{\sigma_n}{\sigma_0 \log(n+1)}\,.$$

As a result of these strategies, the noise level in the system will decrease in time but we are allowed to start with a larger σ which permits to explore the surrounding area well before entering the consensus regime.

Similarly, it might not be beneficial to start with a large α from the

beginning. As a matter of fact, in this case the weighted average $V^{\alpha,\mathcal{E}}$ would right away equal the particle with the lowest energy and all the other particles will be forced to move towards the first particle, with a lower impact on the initial exploration mechanism. Therefore, we can start with an initial value α_0 and gradually increase it to a maximum value $\alpha_{\max}$ according to an appropriate convergence indicator, or independently as a function of the number of iterations. In particular, large values of α at the end of the simulation process are essential in achieving high accuracy in the computation of the minimum.

The number of particles of the system can be considered a parameter that needs to be tuned during the computation as well. Since in the CBO methods the variance of the system tends to vanish because of the consensus dynamics (see Theorem 3.1), we may accelerate the simulation by discarding particles in time according to the variance of the system [2]. This also influences the computation of $V_n^{\alpha,\mathcal{E}}$, by increasing the randomness and reducing the possibilities to get trapped in local minima. We now illustrate a practical implementation of such a strategy as has been presented in [17].

For a set of N_n particle we define the empirical variance at time $T^n = n\Delta t$ as

$$\Sigma_n = \frac{1}{N_n}\sum_{j=1}^{N_n}(V_n^j - \overline{V}_n)^2\,, \quad \overline{V}_n = \frac{1}{N_n}\sum_{j=1}^{N_n} V_n^j\,.$$

When the trend consensus is monotone, that is $\Sigma_{n+1} \leq \Sigma_n$, we can discard particles uniformly in the next time step $t^{n+1} = (n+1)\Delta t$ according to the ratio $\Sigma_{n+1}/\Sigma_n \leq 1$, without affecting their theoretical distribution. One way to make this possible is to define the new number of particles as

$$N_{n+1} = \left\lfloor N_n \left(1 + \mu\left(\frac{\hat{\Sigma}_{n+1} - \Sigma_n}{\Sigma_n}\right)\right)\right\rfloor \tag{4.18}$$

where $\mu \in [0, 1]$ and

$$\hat{\Sigma}_{n+1} = \frac{1}{N_n} \sum_{j=1}^{N_n} (V_{n+1}^j - \hat{V}_{n+1})^2, \quad \hat{V}_{n+1} = \frac{1}{N_n} \sum_{j=1}^{N_n} V_{n+1}^j \,.$$

If $\mu = 0$, we have the standard algorithm where no particles are discarded, whereas for $\mu = 1$ we achieve the maximum speed up.

We conclude by briefly discussing another variant of the CBO method, proposed in [9], for the global optimization of high dimensional Machine Learning problems. The modification directly involves the definition of the stochastic system, by considering the component-wise geometric Brownian motion. Namely, it is suggested to replace the stochastic term in equation (4.1)

$$\sigma |V_t^i - V_t^{\alpha,\mathcal{E}}| P(V_t^i) dB_t^i$$

with the following

$$\sigma |V_t^i - V_t^{\alpha,\mathcal{E}}| P(V_t^i) \sum_{k=1}^{d} dB_t^{i,k}$$

where $(B_t^{i,k})t \geq 0$ are one-dimensional Brownian motions.

In case of unconstrained optimization, it is analytically proved in [9] that such a modification relaxes the convergence conditions, see inequality (3.10), by making them independent from the dimension d. This is a great achievement as the the scalability of the method is a key characteristic of the Consensus-Based Optimization. The new algorithm is then employed in [9] for the optimization of shallow two-layers Neural-Network showing encouraging results.

Conclusions and perspectives

In this work, we introduced and studied a Consensus-Based Optimization method for constrained optimization on compact, implicitly defined hypersurfaces. By employing kinetic theory techniques, we analyzed the evolution of the system through its mean-field approximation which we have been able to derive thanks to the compactness of the domain. In particular, we investigated the variance decay which describes the formation of consensus in the particles dynamics. We noted that choosing the parameters is crucial in order to guarantee that the consensus is generated around a minimizer of the function and that the geometry of the hypersurface could make this process extremely difficult. Moreover, the algorithm has been tested on benchmark functions showing capability of escaping from local minima and fast convergence.

The analysis of CBO methods shed some light on the promising characteristics of these algorithms. Nevertheless, the convergence guarantees, both in the constrained and unconstrained settings, are still restrictive and it is claimed [17] that further improvements could be done by studying the mean-field equation with different techniques. Moreover, after designing CBO methods for hypersurfaces, the next step consists of developing algorithms to solve constrained optimization problems on manifolds with a particular focus on matrix manifolds. To conclude, only the Kuramoto-Vicsek model has been considered so far and, hence, employing other (e.g. second-order) individual-based models is still left to be investigated.

Bibliography

[1] G. Albi, N. Bellomo, L. Fermo, S.-Y. Ha, J. Kim, L. Pareschi, D. Poyato, and J. Soler. Vehicular traffic, crowds, and swarms: From kinetic theory and multiscale methods to applications and research perspectives. *Mathematical Models and Methods in Applied Sciences*, 29(10):1901–2005, 2019.

[2] Giacomo Albi and Lorenzo Pareschi. Binary interaction algorithms for the simulation of flocking and swarming dynamics. *Multiscale Modeling & Simulation*, 11(1):1–29, 2013.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

[4] Andrea Bertozzi, Jesús Rosado, Martin Short, and Li Wang. Contagion shocks in one dimension. *Journal of Statistical Physics*, 158, 02 2015.

[5] François Bolley, José A Canizo, and José A Carrillo. Stochastic mean-field limit: non-Lipschitz forces and swarming. *Mathematical Models and Methods in Applied Sciences*, 21(11):2179–2210, 2011.

[6] J. Carrillo, Young-Pil Choi, and Maxime Hauray. The derivation of swarming models: Mean-field limit and wasserstein distances. *CISM International Centre for Mechanical Sciences, Courses and Lectures*, 553, 04 2013.

[7] J. Carrillo, M Fornasier, Jesús Rosado, and Giuseppe Toscani. Asymptotic flocking dynamics for the kinetic cucker–smale model. *SIAM Journal on Mathematical Analysis*, 42, 05 2009.

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

[9] José A Carrillo, Shi Jin, Lei Li, and Yuhua Zhu. A consensus-based global optimization method for high dimensional machine learning problems. *arXiv preprint arXiv:1909.09249*, 2019.

[10] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer-Verlag Berlin Heidelberg, 2010.

[11] Alan Demlow and Gerhard Dziuk. An adaptive finite element method for the Laplace-Beltrami operator on implicitly defined surfaces. *SIAM Journal on Numerical Analysis*, 45(1):421–442, 2007.

[12] Marco Dorigo and Christian Blum. Ant colony optimization theory: A survey. *Theoretical computer science*, 344(2-3):243–278, 2005.

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

[14] Hugo Jair Escalante, Manuel Montes, and Luis Enrique Sucar. Particle swarm model selection. *Journal of Machine Learning Research*, 10(15):405–440, 2009.

[15] Razvan C Fetecau, Hui Huang, and Weiran Sun. Propagation of chaos for the Keller–Segel equation over bounded domains. *Journal of Differential Equations*, 266(4):2142–2174, 2019.

[16] Massimo Fornasier, Hui Huang, Lorenzo Pareschi, and Philippe Sünnen. Consensus-based optimization on the sphere I: Well-posedness and mean-field limit. *Math. Mod. Meth. Appl. Scie*, 2020.

[17] Massimo Fornasier, Hui Huang, Lorenzo Pareschi, and Philippe Sünnen. Consensus-based optimization on the sphere II : Convergence to global minimizers and machine learning. *arXiv:2001.11988*, 2020.

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

[19] Seung-Yeal Ha, Shi Jin, and Doheon Kim. Convergence of a first-order consensus-based global optimization algorithm, 2019.

[20] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations. 2nd ed*, volume 31. 01 2006.

[21] Richard Holley and Daniel Stroock. Simulated annealing via Sobolev inequalities. *Communications in Mathematical Physics*, 115(4):553–569, 1988.

[22] Hui Huang and Jian-Guo Liu. Error estimate of a random particle blob method for the Keller–Segel equation. *Mathematics of Computation*, 86(308):2719–2744, 2017.

[23] Momin Jamil and Xin She Yang. A literature survey of benchmark functions for global optimisation problems. *International Journal of Mathematical Modelling and Numerical Optimisation*, 4(2):150, 2013.

[24] Shi Jin, Lei Li, and Jian-Guo Liu. Random batch methods (RBM) for interacting particle systems. *Journal of Computational Physics*, 400:108877, 2020.

[25] Dervis Karaboga, Beyza Gorkemli, Celal Ozturk, and Nurhan Karaboga. A comprehensive survey: Artificial bee colony (ABC) algorithm and applications. *Artificial Intelligence Review*, 42, 06 2012.

[26] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 1942–1948 vol.4, 1995.

[27] James Kennedy. Particle swarm optimization. *Encyclopedia of machine learning*, pages 760–766, 2010.

[28] José A. Carrillo Shi Jin Lei Li and Yuhua Zhu. A consensus-based global optimization method for high dimensional machine learning problems. *arXiv preprint arXiv:1909.09249*, 2019.

[29] James G. March. Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87, 1991.

[30] Peter David Miller. *Applied Asymptotic Analysis*, volume 75. American Mathematical Soc., 2006.

[31] Ibrahim Osman and Gilbert Laporte. Metaheuristics: A bibliography. *Annals of Operational Research*, 63:513–628, 10 1996.

[32] Konstantinos Parsopoulos and Michael Vrahatis. Recent approaches to global optimization problems through particle swarm optimization. *Natural Computing ACM Computing Classification*, 116:235–3063, 06 2002.

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

[34] René Pinnau, Claudia Totzeck, Oliver Tse, and Stephan Martin. A consensus-based model for global optimization and its mean-field limit.

Mathematical Models and Methods in Applied Sciences, 27(01):183–204, 2017.

[35] Riccardo Poli, James Kennedy, and Tim Blackwell. Particle swarm optimization. *Swarm intelligence*, 1(1):33–57, 2007.

[36] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling.* Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Basel, 2015.

[37] Y. Shi and R. Eberhart. A modified particle swarm optimizer. In *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*, pages 69–73, 1998.

[38] Alain-Sol Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer, 1991.

[39] Giuseppe Toscani. Kinetic models of opinion formation. *Commun. Math. Sci.*, 4, 09 2006.

[40] Claudia Totzeck, René Pinnau, Sebastian Blauth, and Steffen Schotthöfer. A numerical comparison of consensus-based global optimization to other particle-based global optimization schemes. *PAMM*, 18(1):e201800291, 2018.

[41] Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Sochet. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75, 12 1995.

[42] Cédric Villani. *Optimal transport: old and new.* Grundlehren der mathematischen Wissenschaften. Springer, 2008.

[43] David Wolpert and William Macready. No free lunch theorems for optimization. *Evolutionary Computation, IEEE*, 1:67–82, 01 1997.