Consensus-Based Optimization for Saddle Point Problems

Hui Huang*[1], Jinniao Qiu*[2] and Konstantin Riedl*[3,4]

[1] *University of Graz, Institute of Mathematics and Scientific Computing, Graz, Austria*
[2] *University of Calgary, Department of Mathematics and Statistics, Calgary, Canada*
[3] *Technical University of Munich, School of Computation, Information and Technology, Department of Mathematics, Munich, Germany*
[4] *Munich Center for Machine Learning, Munich, Germany*

Abstract

In this paper we propose consensus-based optimization for saddle point problems (CBO-SP), a novel multi-particle metaheuristic derivative-free optimization method capable of provably finding global Nash equilibria. Following the idea of swarm intelligence, the method employs two groups of interacting particles, one of which performs a minimization over one variable while the other performs a maximization over the other variable. The two groups constantly exchange information through a suitably weighted average. This paradigm permits for a passage to the mean-field limit, which makes the method amenable to theoretical analysis and it allows to obtain rigorous convergence guarantees under reasonable assumptions about the initialization and the objective function, which most notably include nonconvex-nonconcave objectives. We further provide numerical evidence for the success of the algorithm.

Keywords: saddle point problems, Nash equilibria, nonconvex-nonconcave, derivative-free optimization, metaheuristics, consensus-based optimization, Fokker-Planck equations

AMS subject classifications: 90C47, 65C35, 65K05, 90C56, 35Q90, 35Q83

1 Introduction

Optimization problems where the goal is to find the best possible objective value for the worst-case scenario can be formulated as minimax optimization problems of the form

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \mathcal{E}(x,y).$$

To be more specific, given a class of objective functions $\{\mathcal{E}(\,\cdot\,,y), y\in\mathcal{Y}\}$, the aim is to determine the argument $x^*\in\mathcal{X}$ that leads to the smallest objective value even for the worst-case function parametrized by $y^*\in\mathcal{Y}$. Such type of problems were originally formulated in two-player zero-sum game theory [47] but now arise in many areas in mathematics, biology, the social sciences and especially in economics [34]. Diverse applications may be found in engineering, operational research, biology, ecology, finance, economics, energy industry, environmental sciences and so on. In the last few years, minimax optimization has also experienced substantial attention from the signal processing community, due to its connection to distributed processing [8], robust transceiver design [26], and communication in the presence of jammers [16]. Moreover, in modern machine learning, several problems are formulated as minimax optimization, such as the training of

*Email: hui.huang@uni-graz.at, jinniao.qiu@ucalgary.ca, konstantin.riedl@ma.tum.de

generative adversarial networks (GANs) [17], multi-agent reinforcement learning [38], fair machine learning [29], and adversarial training [30]. For example, when training GANs, x models the parameter of a generator, usually a neural network, whose aim is to generate synthetic data with the same statistics as of a given training set, while y represents the parameters of a competing discriminator, who has to distinguish generated data by the generator from data of the true distribution. Relatedly, in adversarial machine learning, one aims at learning the parameters x of a model in a robust manner by exposing it during training to possible adversarial attacks modeled by y. Both examples can be interpreted as a game between two neural networks trained in an adversarial manner until some kind of equilibrium is reached.

In a two-player zero-sum game, the joint payoff function $\mathcal{E}(x, y)$ encodes the gain of the maximization player whose action is to choose $y \in \mathcal{Y}$, as well as the loss of the minimization player controlling the action $x \in \mathcal{X}$. In simultaneous games, each player chooses its action without the knowledge of the action chosen by the other player, so both players act simultaneously. Conversely, in sequential games there is an intrinsic order according to which the players take their actions, meaning that the ordering of the minimization and maximization matters, i.e., it plays a priorly a role whether is $\min_x \max_y$ or $\max_y \min_x$. GANs and adversarial training, for instance, are in fact sequential games in their standard formulations. In the classical case, where the payoff function $\mathcal{E}$ is *convex-concave* (i.e., $\mathcal{E}(\,\cdot\,, y)$ is convex for all $y \in \mathcal{Y}$ and $\mathcal{E}(x, \,\cdot\,)$ is concave for all $x \in \mathcal{X}$), the intrinsic order of sequential games does not matter under an additional compactness assumption on either $\mathcal{X}$ or $\mathcal{Y}$ by the well-known minimax theorems of von Neumann and Sion [45,46]. However, nowadays, most modern applications in signal processing and machine learning entail the setting of *nonconvex-nonconcave* minimax problems, where the minimization and maximization problems are potentially nonconvex and nonconcave. This is significantly more complicated and available tool sets and theories are very limited; see the review paper [42].

A well-known notion of optimality originating from game theory is the one of Nash equilibria (also referred to as saddle points) [35], where neither of the players has anything to gain by changing only his own strategy. This concept is formalized within the following definition.

Definition 1. *A point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is called Nash equilibrium or saddle point of a function $\mathcal{E}$ if it holds*

$$\mathcal{E}(x^*, y) \leq \mathcal{E}(x^*, y^*) \leq \mathcal{E}(x, y^*) \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}$$

or, equivalently, if

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathcal{E}(x, y) = \mathcal{E}(x^*, y^*) = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \mathcal{E}(x, y).$$

To keep the notation concise we write $\mathcal{E}^*$ for $\mathcal{E}(x^*, y^*)$ in what follows.

In the *convex-concave* setting an approximate Nash equilibrium can be found efficiently by variants of gradient descent-ascent (GDA) algorithms [4,18], which alternate between one or more gradient decent steps in the x variable and gradient ascent steps in the y coordinate. Indeed, even if $\mathcal{E}(x, y)$ is either concave in y or convex in x, there are available some multi-step GDA algorithms; see [36, 42] for instance. However, as soon as the payoff function becomes *nonconvex-nonconcave*, finding a global equilibrium is in general an NP-hard problem [33]. For this reason, some recent works such as [9,31] consider a local version of equilibria. More precisely, a point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is called local Nash equilibrium if there exists some $\delta > 0$ such that (x^*, y^*) satisfies Definition 1 in a δ-neighborhood of (x^*, y^*). Local Nash equilibria can be characterized in terms of the so-called quasi-Nash equilibrium condition [39] or the first-order Nash equilibrium condition [36]. Even so, we mention two recent works where special classes of nonconvex-nonconcave payoff functions are concerned. When $\mathcal{E}(x, y)$ is weakly convex in x and weakly concave in y and the associated Minty variational inequality admits a solution, Liu et al. [25] employ the inexact proximal point method and prove the first-order convergence, while under the so-called "sufficiently bilinear" condition, the stochastic Hamiltonian method is investigated by Loizou et al. [28]. In this work, we shall drop such restrictions and the gradient-dependence in the algorithms and consider a *zero-order* (derivative-free) method with rigorous convergence guarantees. Note that the family of

population-based algorithms, such as Particle Swarm Optimization (PSO) [22], has been adapted to solve min-max problems as done for instance in [23, 24, 44]. One straightforward approach is to treat the min-max problem as a minimization problem and embed the maximization part in the calculation of the objective values [24]. Alternatively, a multi-PSO strategy [23, 44] may be employed, where the min-max problem is converted into two optimization problems, one being a maximization problem, and the other a minimization problem. Two PSO algorithms are then used to solve these two optimization problems, respectively, and they are run independently. Each PSO is treated as a changing environment of the other PSO, allowing them to cooperate through the calculation of the objective. Both approaches cannot avoid the necessity of nested loops/circles of optimization algorithms, which significantly increases the time complexity.

In the present paper we propose a zero-order consensus-based optimization method for finding the global Nash equilibrium (x^*, y^*) of a smooth objective function $\mathcal{E} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}^{d_2}$, which is designed to be amenable to a rigorous theoretical convergence analysis, missing so far in the literature on population-based methods for min-max problems. The dynamics of the algorithm is inspired by consensus-based optimization, a paradigm for global nonconvex minimizations, which was introduced by the authors of [40]. Their method employs a system of interacting particles which explore the energy landscape in order to form a global consensus about the global minimizer of the objective function as time passes. Taking inspiration from this concept, let us consider two sets of particles $(X^i)_{i=1}^{N_1}$ and $(Y^i)_{i=1}^{N_2}$ of potentially different size, one for minimization, the other for maximization. Each individual particle of either set is formally described by a stochastic process. In order to achieve consensus about the equilibrium point of $\mathcal{E}$, the particles interact through a system of stochastic differential equations (SDEs) of the form

$$dX_t^i = -\lambda_1 \left(X_t^i - x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1})\right) dt + \sigma_1 D\Big(X_t^i - x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1})\Big)\, dB_t^{X,i}, \qquad \widehat{\rho}_{X,t}^{N_1} = \tfrac{1}{N_1}\textstyle\sum_{i=1}^{N_1} \delta_{X_t^i}, \tag{1a}$$

$$dY_t^i = -\lambda_2 \left(Y_t^i - y_\beta^X(\widehat{\rho}_{Y,t}^{N_2})\right) dt + \sigma_2 D\Big(Y_t^i - y_\beta^X(\widehat{\rho}_{Y,t}^{N_2})\Big)\, dB_t^{Y,i}, \qquad \widehat{\rho}_{Y,t}^{N_2} = \tfrac{1}{N_2}\textstyle\sum_{i=1}^{N_2} \delta_{Y_t^i}, \tag{1b}$$

which is complemented by suitable initial conditions $X_0^i \sim \rho_{X,0} \in \mathcal{P}(\mathbb{R}^{d_1})$ for $i = 1, \dots, N_1$ and $Y_0^i \sim \rho_{Y,0} \in \mathcal{P}(\mathbb{R}^{d_2})$ for $i = 1, \dots, N_2$ and where $\big((B_t^{X,i})_{t\geq 0}\big)_{i=1,\dots,N_1}$ and $\big((B_t^{Y,i})_{t\geq 0}\big)_{i=1,\dots,N_2}$ are independent standard Brownian motions in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Moreover, $\widehat{\rho}_{X,t}^{N_1}$ and $\widehat{\rho}_{Y,t}^{N_2}$ denote the empirical measures of the particles' x- and y-positions, respectively. While the dynamics (1a) performs minimization in the x-variable, (1b) performs maximization in the y-coordinate. This is encoded in the computation of the so-called consensus point $\big(x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1}), y_\beta^X(\widehat{\rho}_{Y,t}^{N_2})\big)$, whose components are given by

$$x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1}) = \int x \, \frac{\omega_\alpha\big(x, \int y\, d\widehat{\rho}_{Y,t}^{N_2}(y)\big)}{\big\|\omega_\alpha\big(\,\cdot\,, \int y\, d\widehat{\rho}_{Y,t}^{N_2}(y)\big)\big\|_{L_1(\widehat{\rho}_{X,t}^{N_1})}}\, d\widehat{\rho}_{X,t}^{N_1}(x), \qquad \text{with } \omega_\alpha(x,y) := \exp(-\alpha \mathcal{E}(x,y)), \tag{2a}$$

$$y_\beta^X(\widehat{\rho}_{Y,t}^{N_2}) = \int y \, \frac{\omega_{-\beta}\big(\int x\, d\widehat{\rho}_{X,t}^{N_1}(x), y\big)}{\big\|\omega_{-\beta}\big(\int x\, d\widehat{\rho}_{X,t}^{N_1}(x), \,\cdot\,\big)\big\|_{L_1(\widehat{\rho}_{Y,t}^{N_2})}}\, d\widehat{\rho}_{Y,t}^{N_2}(y), \quad \text{with } \omega_{-\beta}(x,y) := \exp(\beta \mathcal{E}(x,y)). \tag{2b}$$

Attributed to the Laplace principle [32], $x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1})$ can be interpreted as an approximation of $\arg\min_{i=1,\dots,N_1} \mathcal{E}(X_t^i, \int y\, d\widehat{\rho}_{Y,t}^{N_2}(y))$ as $\alpha \to \infty$ while $y_\beta^X(\widehat{\rho}_{Y,t}^{N_2}) \approx \operatorname{argmax}_{i=1,\dots,N_2} \mathcal{E}(\int x\, d\widehat{\rho}_{X,t}^{N_1}(x), Y_t^i)$ as $\beta \to \infty$, see, e.g., [14, Equation (7)]. The dynamics of each of the particles in (1) is governed by two terms. A drift term drags the particles towards the respective component of the instantaneous consensus point $\big(x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1}), y_\beta^X(\widehat{\rho}_{Y,t}^{N_2})\big)$ and thereby expectedly improves the position of the particles. The second term injects stochasticity into the dynamics by diffusing the particles according to a scaled Brownian motion, which features the exploration of the landscape of the objective. In what follows we use anisotropic noise, i.e., $D(\,\cdot\,) = \operatorname{diag}(\,\cdot\,)$, which is typically more competitive in high dimensions compared to isotropic noise $D(\,\cdot\,) = \|\,\cdot\,\|_2$, see, e.g., [6, 15]. The theoretical results of this paper, however, can be obtained mutatis mutandis also in the isotropic setting.

An implementable scheme for a numerical algorithm can be obtained from (1) by a simple Euler-Maruyama time discretization [19, 41]. For details about the implementation we refer to Algorithm 1 in Section 5.1.

Remark 2. *While the definition of the consensus point in (2) is a natural option, there are two equally reasonable alternatives. The first possibility is to replace the mean* $\int y \, d\widehat{\rho}_{Y,t}^{N_2}(y)$ *in (2a) simply by* y *and integrate w.r.t. the joint measure* $\widehat{\rho}_t^N$*. This case would require* $N_1 = N_2$*. Analogously,* $\int x \, d\widehat{\rho}_{X,t}^{N_1}(x)$ *is substituted by* x *in (2b). The second option is to use the other component of the consensus point instead of the respective mean, i.e.,* $y_\beta^X(\widehat{\rho}_{Y,t}^{N_2})$ *replaces* $\int y \, d\widehat{\rho}_{Y,t}^{N_2}(y)$ *in (2a) and* $x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1})$ *substitutes* $\int x \, d\widehat{\rho}_{X,t}^{N_1}(x)$ *in (2b).*

The main motivation for using the variant as in (2) is of theoretical nature. Using either of the other two alternatives significantly complicates the convergence analysis in Sections 3 and 4.

Understanding the convergence properties of the dynamics (1) can take place either by investigating the long time behavior of the interacting particle system itself, or by analyzing the macroscopic behavior of the agent density associated with (1) through a mean-field limit. This theoretical approach proved successful in [2, 3, 5–7, 12–15, 43] for proving global convergence for several variants of consensus-based optimization in the setting of minimization. It is moreover theoretically justified by the mean-field approximation which shows that $(\widehat{\rho}_{X,t}^{N_1}, \widehat{\rho}_{Y,t}^{N_2})$ converges in some sense to a mean field law $(\rho_{X,t}, \rho_{Y,t})$ as $N_1, N_2 \to \infty$. Again, for consensus-based optimization there exist by now several results in this direction such as [11, 14, 20], which may be extended to CBO-SP in an immediate manner. In the setting of saddle point problems, the mean-field dynamics associated with (1) can be described by the self-consistent mono-particle dynamics

$$d\overline{X}_t = -\lambda_1 \left(\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right) dt + \sigma_1 D\big(\overline{X}_t - x_\alpha^Y(\rho_{X,t})\big)\, dB_t^X, \qquad \rho_{X,t} = \textstyle\int d\rho_t(\,\cdot\,, y), \tag{3a}$$

$$d\overline{Y}_t = -\lambda_2 \left(\overline{Y}_t - y_\beta^X(\rho_{Y,t})\right) dt + \sigma_2 D\big(\overline{Y}_t - y_\beta^X(\rho_{Y,t})\big)\, dB_t^Y, \qquad \rho_{Y,t} = \textstyle\int d\rho_t(x, \,\cdot\,), \tag{3b}$$

where $\rho_t = \rho(t) = \mathrm{Law}\left((\overline{X}_t, \overline{Y}_t)\right)$ with marginals $\rho_{X,t}$ and $\rho_{Y,t}$, respectively. In particular, the measure $\rho \in \mathcal{C}([0,T], \mathcal{P}(\mathbb{R}^{d_1+d_2}))$ weakly satisfies the nonlinear nonlocal Fokker-Planck equation

$$\begin{aligned}\partial_t \rho_t =& \lambda_1 \mathrm{div}_x\big(\big(x - x_\alpha^Y(\rho_t^X)\big)\,\rho_t\big) + \lambda_2 \mathrm{div}_y\big(\big(y - y_\beta^X(\rho_t^Y)\big)\,\rho_t\big) \\ &+ \frac{\sigma_1^2}{2} \sum_{k=1}^{d_1} \partial_{x_k x_k}^2 \big((x - x_\alpha^Y(\rho_t^X))_k^2 \rho_t\big) + \frac{\sigma_2^2}{2} \sum_{k=1}^{d_2} \partial_{y_k y_k}^2 \big((y - y_\beta^X(\rho_t^Y))_k^2 \rho_t\big).\end{aligned} \tag{4}$$

Contributions. Motivated by the fundamental importance of *nonconvex-nonconcave* saddle point problems in various applicational areas and the desire for having numerical algorithms with rigorous global convergence guarantees, we theoretically analyze in this work a novel consensus-based optimization method (CBO-SP) capable of tackling saddle point problems. Using mean-field analysis techniques, we rigorously prove that CBO-SP converges to saddle points as the number of interacting particles goes to infinity. Our results hold under reasonable assumptions about the objective function and under certain conditions of well-preparation of the hyperparameters and the initial data.

1.1 Organization

In Section 2 we first investigate the well-posedness of both the interacting particle system (1) of CBO-SP and its associated mean-field dynamics (3). Section 3 then presents and discusses the main theoretical statement of this work concerned with the convergence of the mean-field dynamics (3) towards saddle points of the objective function $\mathcal{E}$, which are proven in Section 4. Section 5 contains details about the implementation of the numerical algorithm as well as instructive numerical examples which illustrate how CBO-SP works, before we conclude the paper in Section 6. In the GitHub repository `https://github.com/KonstantinRiedl/CBOSaddlePoints` we provide the Matlab code implementing CBO-SP.

2 Well-Posedness of CBO-SP and its Mean-Field Dynamics

In the first part of this section we provide a well-posedness result about the interacting particle system (1) of CBO-SP, i.e., we show that a process obeying (1) exists and is unique. Afterwards we also prove the well-posedness of the nonlinear macroscopic SDE (3).

2.1 Well-Posedness of the Interacting Particle System

To keep the notation concise in what follows, let us denote the state vector of the entire particle system (1) by $\mathbf{Z} \in \mathcal{C}([0,\infty), \mathbb{R}^{N_1 d_1 + N_2 d_2})$ with $\mathbf{Z}(t) = \mathbf{Z}_t = \left((X_t^1)^T, \dots, (X_t^{N_1})^T, (Y_t^1)^T, \dots, (Y_t^{N_2})^T\right)^T$ for every $t \geq 0$. Equation (1) can then be reformulated as

$$d\mathbf{Z}_t = -\boldsymbol{\lambda}\mathbf{F}(\mathbf{Z}_t)\,dt + \boldsymbol{\sigma}\mathbf{M}(\mathbf{Z}_t)\,d\mathbf{B}_t \tag{5}$$

with $(\mathbf{B}_t)_{t\geq 0}$ being a standard Brownian motion in $\mathbb{R}^{N_1 d_1 + N_2 d_2}$ and definitions

$$\begin{aligned}
\mathbf{F}(\mathbf{Z}_t) &:= \left(F^{1,X}(\mathbf{Z}_t)^T, \dots, F^{N_1,X}(\mathbf{Z}_t)^T, F^{1,Y}(\mathbf{Z}_t)^T, \dots, F^{N_2,Y}(\mathbf{Z}_t)^T\right)^T \\
&\quad \text{with } F^{i,X}(\mathbf{Z}_t) = \left(X_t^i - x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1})\right) \text{ and } F^{i,Y}(\mathbf{Z}_t) = \left(Y_t^i - y_\beta^X(\widehat{\rho}_{Y,t}^{N_2})\right), \\
\mathbf{M}(\mathbf{Z}_t) &:= \operatorname{diag}\left(M^{1,X}(\mathbf{Z}_t), \dots, M^{N_1,X}(\mathbf{Z}_t), M^{1,Y}(\mathbf{Z}_t), \dots, M^{N_2,Y}(\mathbf{Z}_t)\right) \\
&\quad \text{with } M^{i,X}(\mathbf{Z}_t) = D\left(X_t^i - x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1})\right) \text{ and } M^{i,Y}(\mathbf{Z}_t) = D\left(Y_t^i - y_\beta^X(\widehat{\rho}_{Y,t}^{N_2})\right).
\end{aligned}$$

The diag operator in the definition of $\mathbf{M}$ maps the input matrices onto a block-diagonal matrix with them as its diagonal. $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}$ are $(N_1 d_1 + N_2 d_2) \times (N_1 d_1 + N_2 d_2)$-dimensional diagonal matrices, whose first $N_1 d_1$ entries are λ_1 and σ_1, and the remaining $N_2 d_2$ entires are λ_2 and σ_2, respectively.

Having fixed the notation, we have the following well-posedness result for the SDE system (5), respectively (1), which is proven towards the end of this section.

Theorem 3. *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^{d_1+d_2})$ be locally Lipschitz continuous. Then, for $N_1, N_2 \in \mathbb{N}$ fixed, the system of SDEs (1) admits a unique strong solution $(\mathbf{Z}_t)_{t\geq 0}$ for any initial condition $\mathbf{Z}_0$ satisfying $\mathbb{E}\|\mathbf{Z}_0\|_2^2 < \infty$.*

In order to employ the standard result [10, Chapter 5, Theorem 3.1] about the existence and uniqueness of solutions to SDEs, we need to verify that the coefficients of the SDE are locally Lipschitz continuous and of at most linear growth. This is inherited from the assumed local Lipschitz continuity of $\mathcal{E}$ as we make explicit in the subsequent lemma.

Lemma 4. *Let $N_1, N_2 \in \mathbb{N}$, $\alpha, \beta > 0$ and $R > 0$ be arbitrary. Let $\mathbf{z}, \widehat{\mathbf{z}} \in \mathbb{R}^{N_1 d_1 + N_2 d_2}$ be of the form $\mathbf{z} = (\mathbf{x}^T, \mathbf{y}^T)^T = \left((x^1)^T, \dots, (x^{N_1})^T, (y^1)^T, \dots, (y^{N_2})^T\right)^T$ and analogously for $\widehat{\mathbf{z}}$. Then, for any $\mathbf{z}, \widehat{\mathbf{z}}$ with $\|\mathbf{z}\|_2 \leq R$ and $\|\widehat{\mathbf{z}}\|_2 \leq R$ it holds for any i the bounds*

$$\left\|F^{i,X}(\mathbf{z})\right\|_2 \leq \|x^i\|_2 + \|\mathbf{x}\|_2 \quad \text{and} \quad \left\|F^{i,Y}(\mathbf{z})\right\|_2 \leq \|y^i\|_2 + \|\mathbf{y}\|_2,$$

and, abbreviating $c_R(\gamma) := 4\gamma e^{2\gamma \Delta_R \mathcal{E}} \left\|\|\nabla_{\mathbf{z}} \mathcal{E}\|_2\right\|_{L^\infty(B_R)}$ with $\Delta_R \mathcal{E} := \sup_{\mathbf{z}\in B_R} \mathcal{E}(\mathbf{z}) - \inf_{\mathbf{z}\in B_R} \mathcal{E}(\mathbf{z})$,

$$\left\|F^{i,X}(\mathbf{z}) - F^{i,X}(\widehat{\mathbf{z}})\right\|_2 \leq \|x^i - \widehat{x}^i\|_2 + \left(1 + \frac{c_R(\alpha)}{N_1}\sqrt{N_1 \|\widehat{x}^i\|_2^2 + \|\widehat{\mathbf{x}}\|_2^2}\right)\left(\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|_2\right),$$

$$\left\|F^{i,Y}(\mathbf{z}) - F^{i,Y}(\widehat{\mathbf{z}})\right\|_2 \leq \|y^i - \widehat{y}^i\|_2 + \left(1 + \frac{c_R(\beta)}{N_2}\sqrt{N_2 \|\widehat{y}^i\|_2^2 + \|\widehat{\mathbf{y}}\|_2^2}\right)\left(\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|_2\right).$$

Proof. To derive the first bound we note that

$$\left\|F^{i,X}(\mathbf{z})\right\|_2 = \left\| x^i - \sum_{j=1}^{N_1} x^j \frac{\omega_\alpha\left(x^j, \frac{1}{N_1}\sum_{k=1}^{N_1} y^k\right)}{\sum_{j=1}^{N_1} \omega_\alpha\left(x^j, \frac{1}{N_1}\sum_{k=1}^{N_1} y^k\right)} \right\|_2 \leq \|x^i\|_2 + \|\mathbf{x}\|_2.$$

Analogously the bound for $\left\|F^{i,Y}(\mathbf{z})\right\|_2$ is obtained. For the other estimates we first notice that

$$\begin{aligned} F^{i,X}(\mathbf{z}) - F^{i,X}(\widehat{\mathbf{z}}) &= \frac{\sum_{j=1}^{N_1}(x^i - x^j)\,\omega_\alpha\big(x^j, \frac{1}{N_1}\sum_{k=1}^{N_1} y^k\big)}{\sum_{j=1}^{N_1}\omega_\alpha\big(x^j, \frac{1}{N_1}\sum_{k=1}^{N_1} y^k\big)} - \frac{\sum_{j=1}^{N_1}(\widehat{x}^i - \widehat{x}^j)\,\omega_\alpha\big(\widehat{x}^j, \frac{1}{N_1}\sum_{k=1}^{N_1} \widehat{y}^k\big)}{\sum_{j=1}^{N_1}\omega_\alpha\big(\widehat{x}^j, \frac{1}{N_1}\sum_{k=1}^{N_1} \widehat{y}^k\big)} \\ &= I_1 + I_2 + I_3 \end{aligned}$$

with I_1, I_2 and I_3 being defined as in what follows. Firstly, for I_1 we have

$$\|I_1\|_2 := \left\| \frac{\sum_{j=1}^{N_1}\left((x^i - x^j) - (\widehat{x}^i - \widehat{x}^j)\right)\omega_\alpha\big(x^j, \frac{1}{N_1}\sum_{k=1}^{N_1} y^k\big)}{\sum_{j=1}^{N_1}\omega_\alpha\big(x^j, \frac{1}{N_1}\sum_{k=1}^{N_1} y^k\big)} \right\|_2 \leq \|x^i - \widehat{x}^i\|_2 + \|\mathbf{x} - \widehat{\mathbf{x}}\|_2 .$$

For I_2 and I_3, on the other hand, let us first notice that it holds

$$\begin{aligned} &\left|\omega_\alpha\left(x^j, \frac{1}{N_1}\sum_{k=1}^{N_1} y^k\right) - \omega_\alpha\left(\widehat{x}^j, \frac{1}{N_1}\sum_{k=1}^{N_1} \widehat{y}^k\right)\right| \\ &\qquad \leq \alpha e^{-\alpha \inf_{\mathbf{z}\in B_R}\mathcal{E}(\mathbf{x},\mathbf{y})}\left\|\|\nabla_{\mathbf{z}}\mathcal{E}\|_2\right\|_{L^\infty(B_R)}\left(\left\|x^j - \widehat{x}^j\right\|_2 + \frac{1}{N_1}\sum_{k=1}^{N_1}\left\|y^k - \widehat{y}^k\right\|_2\right) \end{aligned}$$

and

$$\frac{1}{\sum_{j=1}^{N_1}\omega_\alpha\big(x^j, \frac{1}{N_1}\sum_{k=1}^{N_1} y^k\big)} \leq \frac{1}{N_1 \inf_{\mathbf{z}\in B_R}\exp(-\alpha\mathcal{E}(\mathbf{x},\mathbf{y}))} \leq \frac{1}{N_1 e^{-\alpha\sup_{\mathbf{z}\in B_R}\mathcal{E}(\mathbf{x},\mathbf{y})}}.$$

With this we immediately obtain for the norm of I_2 the upper bound

$$\begin{aligned} \|I_2\|_2 &:= \left\| \frac{\sum_{j=1}^{N_1}(\widehat{x}^i - \widehat{x}^j)\left(\omega_\alpha\big(x^j, \frac{1}{N_1}\sum_{k=1}^{N_1} y^k\big) - \omega_\alpha\big(\widehat{x}^j, \frac{1}{N_1}\sum_{k=1}^{N_1} \widehat{y}^k\big)\right)}{\sum_{j=1}^{N_1}\omega_\alpha\big(x^j, \frac{1}{N_1}\sum_{k=1}^{N_1} y^k\big)} \right\|_2 \\ &\leq \frac{2\alpha e^{\alpha\Delta_R\mathcal{E}}\left\|\|\nabla_{\mathbf{z}}\mathcal{E}\|_2\right\|_{L^\infty(B_R)}}{N_1}\sqrt{N_1\,\|\widehat{x}^i\|_2^2 + \|\widehat{\mathbf{x}}\|_2^2}\big(\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|_2\big), \end{aligned}$$

where we abbreviate $\Delta_R\mathcal{E} := \sup_{\mathbf{z}\in B_R}\mathcal{E}(\mathbf{x},\mathbf{y}) - \inf_{\mathbf{z}\in B_R}\mathcal{E}(\mathbf{x},\mathbf{y})$. Similarly, for I_3 we have

$$\begin{aligned} \|I_3\|_2 &:= \left\| \sum_{j=1}^{N_1}(\widehat{x}^i - \widehat{x}^j)\,\omega_\alpha\left(\widehat{x}^j, \frac{1}{N_1}\sum_{k=1}^{N_1}\widehat{y}^k\right)\frac{\left(\sum_{j=1}^{N_1}\omega_\alpha\big(\widehat{x}^j, \frac{1}{N_1}\sum_{k=1}^{N_1}\widehat{y}^k\big) - \omega_\alpha\big(x^j, \frac{1}{N_1}\sum_{k=1}^{N_1} y^k\big)\right)}{\sum_{j=1}^{N_1}\omega_\alpha\big(\widehat{x}^j, \frac{1}{N_1}\sum_{k=1}^{N_1}\widehat{y}^k\big)\sum_{j=1}^{N_1}\omega_\alpha\big(x^j, \frac{1}{N_1}\sum_{k=1}^{N_1} y^k\big)} \right\|_2 \\ &\leq \frac{2\alpha e^{2\alpha\Delta_R\mathcal{E}}\left\|\|\nabla_{\mathbf{z}}\mathcal{E}\|_2\right\|_{L^\infty(B_R)}}{N_1}\sqrt{N_1\,\|\widehat{x}^i\|_2^2 + \|\widehat{\mathbf{x}}\|_2^2}\big(\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|_2\big). \end{aligned}$$

Combining these bounds yields the result. Analogously we can bound $\left\|F^{i,Y}(\mathbf{z}) - F^{i,Y}(\widehat{\mathbf{z}})\right\|_2$. □

Proof of Theorem 3. The statement follows by invoking the standard result [10, Chapter 5, Theorem 3.1] (see Theorem A.1 in the Appendix) on the existence and pathwise uniqueness of a strong solution. That Condition (i) of Theorem A.1 about the local Lipschitz continuity and linear growth of $\mathbf{F}(\mathbf{Z}_t)$ and $\mathbf{M}(\mathbf{Z}_t)$ holds, follows immediately from Lemma 4. To ensure Condition (ii) of Theorem A.1 we make use of [10, Chapter 5, Theorem 3.2] (see Theorem A.2 in the Appendix) and verify that there exists a constant $b_{N_1,N_2} > 0$ such that $-2\boldsymbol{\lambda}\mathbf{Z}_t \cdot \mathbf{F}(\mathbf{Z}_t) + \operatorname{tr}(\boldsymbol{\sigma}\mathbf{M}(\mathbf{Z}_t)\mathbf{M}(\mathbf{Z}_t)^T\boldsymbol{\sigma}^T) \leq b_{N_1,N_2}(1 + \|\mathbf{Z}_t\|_2^2)$. Indeed, since

$$\begin{aligned} -\boldsymbol{\lambda}\mathbf{Z}_t \cdot \mathbf{F}(\mathbf{Z}_t) &\leq \lambda_1\sum_{i=1}^{N_1}\left\|X_t^i\right\|_2\left\|F^{i,X}(\mathbf{Z}_t)\right\|_2 + \lambda_2\sum_{i=1}^{N_2}\left\|Y_t^i\right\|_2\left\|F^{i,Y}(\mathbf{Z}_t)\right\|_2 \\ &\leq \left(\lambda_1\big(1 + \sqrt{N_1}\big) + \lambda_2\big(1 + \sqrt{N_2}\big)\right)\|\mathbf{Z}_t\|_2^2 \end{aligned}$$

and

$$\begin{aligned}\operatorname{tr}(\boldsymbol{\sigma}\mathbf{M}(\mathbf{Z}_t)\mathbf{M}(\mathbf{Z}_t)^T\boldsymbol{\sigma}^T) &= \sigma_1^2\sum_{i=1}^{N_1}\left\|F^{i,X}(\mathbf{Z}_t)\right\|_2^2 + \sigma_2^2\sum_{i=1}^{N_2}\left\|F^{i,Y}(\mathbf{Z}_t)\right\|_2^2\\ &\leq 2\left(\sigma_1^2(1+N_1)+\sigma_2^2(1+N_2)\right)\|\mathbf{Z}_t\|_2^2,\end{aligned}$$

the former holds with b_{N_1,N_2} defined as the sum of the two former upper bounds. □

2.2 Well-Posedness of the Mean-Field Dynamics

In what follows let us furthermore ensure the well-posedness of the mean-field dynamics (3) and (4), which is the main object of our studies in Section 3. We prove existence and uniqueness of a solution for objective functions $\mathcal{E}$ that satisfies the following conditions.

Definition 5 (Assumptions)**.** *In this section we consider functions $\mathcal{E}\in\mathcal{C}^1(\mathbb{R}^{d_1+d_2})$, which*

W1 *are bounded in the sense that there exist $\underline{\mathcal{E}}\in\mathcal{C}^1(\mathbb{R}^{d_2})$ and $\overline{\mathcal{E}}\in\mathcal{C}^1(\mathbb{R}^{d_1})$ such that*

$$\underline{\mathcal{E}}(y)\leq\mathcal{E}(x,y)\leq\overline{\mathcal{E}}(x)\quad\text{for all }(x,y)\in\mathbb{R}^{d_1+d_2}.$$

W2 *are locally Lipschitz continuous in the sense that there exists a constant $C_1>0$ such that for all $(x,y),(x',y')\in\mathbb{R}^{d_1+d_2}$ it holds*

$$\left|\mathcal{E}(x,y)-\mathcal{E}(x',y')\right|\leq C_1\left(1+\|x\|_2+\|x'\|_2+\|y\|_2+\|y'\|_2\right)\left(\|x-x'\|_2+\|y-y'\|_2\right).$$

W3 *have at most quadratic growth in the sense that there exists a constant $C_2>0$ obeying*

$$\begin{aligned}\mathcal{E}(x,y)-\underline{\mathcal{E}}(y+sy')&\leq C_2\left(1+\|x\|_2^2+\|y\|_2^2+\|y'\|_2^2\right)\quad\text{for all }(x,y),(x',y')\in\mathbb{R}^{d_1+d_2},s\in[0,1],\\ \overline{\mathcal{E}}(x+sx')-\mathcal{E}(x,y)&\leq C_2\left(1+\|x\|_2^2+\|x'\|_2^2+\|y\|_2^2\right)\quad\text{for all }(x,y),(x',y')\in\mathbb{R}^{d_1+d_2},s\in[0,1].\end{aligned}$$

For such objectives we have the following well-posedness result for the macroscopic SDE (3) and its associated Fokker-Planck equation (4).

Theorem 6. *Let $\mathcal{E}$ satisfy Assumptions W1–W3. Let $T>0$, $\rho_0\in\mathcal{P}_4(\mathbb{R}^{d_1+d_2})$. Then there exists a unique nonlinear process $(\overline{X},\overline{Y})\in\mathcal{C}([0,T],\mathbb{R}^{d_1+d_2})$ satisfying (3). The associated law $\rho=\operatorname{Law}(\overline{X},\overline{Y})$ has regularity $\rho\in\mathcal{C}([0,T],\mathcal{P}_4(\mathbb{R}^{d_1+d_2}))$ and is a weak solution to the Fokker-Planck equation (4).*

Before giving the proof of Theorem 6 let us first provide some auxiliary results.

Lemma 7. *Let $\varrho,\widehat{\varrho}\in\mathcal{P}_2(\mathbb{R}^{d_1+d_2})$ with $\iint\|x\|_2^2+\|y\|_2^2\,d\varrho(x,y)\leq K$ and $\iint\|\widehat{x}\|_2^2+\|\widehat{y}\|_2^2\,d\widehat{\varrho}(\widehat{x},\widehat{y})\leq K$. Then, under Assumptions W1 and W3 on $\mathcal{E}$, it holds for any $s\in[0,1]$ that*

$$\frac{\exp\left(-\alpha\underline{\mathcal{E}}\left(\int y d\varrho_Y(y)+s\left(\int\widehat{y}d\widehat{\varrho}_Y(\widehat{y})-\int y d\varrho_Y(y)\right)\right)\right)}{\int\omega_\alpha\left(x,\int y d\varrho_Y(y)\right)d\varrho_X(x)}\leq\exp\left(\alpha C_2(1+2K)\right):=C_K^\alpha \tag{7a}$$

and

$$\frac{\exp\left(\beta\overline{\mathcal{E}}\left(\int x d\varrho_X(x)+s\left(\int\widehat{x}d\widehat{\varrho}_X(\widehat{x})-\int x d\varrho_X(x)\right)\right)\right)}{\int\omega_{-\beta}\left(\int x d\varrho_X(x),y\right)d\varrho_X(x)}\leq\exp\left(\beta C_2(1+2K)\right):=C_K^\beta. \tag{7b}$$

Proof. By exploiting Assumption W3 and utilizing Jensen's inequality, we obtain

$$\begin{aligned}&\frac{\exp\left(-\alpha\underline{\mathcal{E}}\left(\int y d\varrho_Y(y)+s\left(\int\widehat{y}d\widehat{\varrho}_Y(\widehat{y})-\int y d\varrho_Y(y)\right)\right)\right)}{\int\omega_\alpha\left(x,\int y d\varrho_Y(y)\right)d\varrho_X(x)}\\ &\quad\leq\frac{1}{\exp\left(-\alpha C_2\left(\int 1+\|x\|_2^2+\left\|\int y d\varrho_Y(y)\right\|_2^2+\left\|\int\widehat{y}d\widehat{\varrho}_Y(\widehat{y})\right\|_2^2 d\varrho_X(x)\right)\right)}\leq\exp\left(\alpha C_2(1+2K)\right),\end{aligned}$$

where in the last inequality we integrated the moment bounds on ϱ and $\widehat{\varrho}$. A similar estimate gives (7b) by exploiting Assumption W3. □

Lemma 8. *Let $\mathcal{E}$ satisfy Assumption W1 and the assumptions of Theorem 6. Let $\varrho, \widehat{\varrho} \in \mathcal{P}_4(\mathbb{R}^{d_1+d_2})$ with $\iint \|x\|_2^4+\|y\|_2^4 \, d\varrho(x,y) \leq K$ and $\iint \|\widehat{x}\|_2^4+\|\widehat{y}\|_2^4 \, d\widehat{\varrho}(\widehat{x},\widehat{y}) \leq K$. Then it holds the stability estimate*

$$\left\|x_\alpha^Y(\varrho_X) - x_\alpha^Y(\widehat{\varrho}_X)\right\|_2 + \left\|y_\beta^X(\varrho_Y) - y_\beta^X(\widehat{\varrho}_Y)\right\|_2 \leq c_0 W_2(\varrho, \widehat{\varrho}) \tag{8}$$

with c_0 depending only on α, β, C_1, C_2 and K.

Proof. To keep the notation concise we write $\mathsf{E}\varrho_Y := \int y \, d\varrho_Y(y)$ and $\mathsf{E}\widehat{\varrho}_Y := \int \widehat{y} \, d\widehat{\varrho}_Y(\widehat{y})$ in what follows. According to the definition of the consensus point in (2a) we have

$$x_\alpha^Y(\varrho_X) - x_\alpha^Y(\widehat{\varrho}_X) = \iint \underbrace{\frac{x\omega_\alpha(x, \mathsf{E}\varrho_Y)}{\int \omega_\alpha(x, \mathsf{E}\varrho_Y) \, d\varrho_X(x)} - \frac{\widehat{x}\omega_\alpha(\widehat{x}, \mathsf{E}\widehat{\varrho}_Y)}{\int \omega_\alpha(\widehat{x}, \mathsf{E}\widehat{\varrho}_Y) \, d\widehat{\varrho}_X(\widehat{x})}}_{=: h(x) - h(\widehat{x})} d\pi(x, y, \widehat{x}, \widehat{y}),$$

where $\pi \in \Pi(\varrho, \widehat{\varrho})$ is any coupling of ϱ and $\widehat{\varrho}$. By adding and subtracting mixed terms, we obtain the decomposition

$$\begin{aligned} h(x) - h(\widehat{x}) &= \frac{(x-\widehat{x})\,\omega_\alpha(x, \mathsf{E}\varrho_Y)}{\int \omega_\alpha(x, \mathsf{E}\varrho_Y) \, d\varrho_X(x)} + \frac{\widehat{x}\left(\omega_\alpha(x, \mathsf{E}\varrho_Y) - \omega_\alpha(\widehat{x}, \mathsf{E}\widehat{\varrho}_Y)\right)}{\int \omega_\alpha(x, \mathsf{E}\varrho_Y) \, d\varrho_X(x)} \\ &\quad + \frac{\iint \omega_\alpha(\widehat{x}, \mathsf{E}\widehat{\varrho}_Y) - \omega_\alpha(x, \mathsf{E}\varrho_Y) \, d\pi(x, y, \widehat{x}, \widehat{y})}{\left(\int \omega_\alpha(x, \mathsf{E}\varrho_Y) \, d\varrho_X(x)\right)\left(\int \omega_\alpha(\widehat{x}, \mathsf{E}\widehat{\varrho}_Y) \, d\widehat{\varrho}_X(\widehat{x})\right)} \widehat{x}\omega_\alpha(\widehat{x}, \mathsf{E}\widehat{\varrho}_Y) =: I_1 + I_2 + I_3, \end{aligned}$$

where I_1, I_2 and I_3 correspond to the three summands. For I_1, by using Assumption W3 and Lemma 7 with $s = 0$, we obtain

$$\|I_1\|_2 \leq \frac{\omega_\alpha(x, \mathsf{E}\varrho_Y)}{\int \omega_\alpha(x, \mathsf{E}\varrho_Y) \, d\varrho_X(x)} \|x - \widehat{x}\|_2 \leq \frac{e^{-\alpha \underline{\mathcal{E}}(\mathsf{E}\varrho_Y)}}{\int \omega_\alpha(x, \mathsf{E}\varrho_Y) \, d\varrho_X(x)} \|x - \widehat{x}\|_2 \leq C_K^\alpha \|x - \widehat{x}\|_2 .$$

For I_2 and I_3, on the other hand, let us first notice that for some $s, s' \in [0,1]$ it holds

$$\begin{aligned} &\left|\omega_\alpha(x, \mathsf{E}\varrho_Y) - \omega_\alpha(\widehat{x}, \mathsf{E}\widehat{\varrho}_Y)\right| \leq \left|\omega_\alpha(x, \mathsf{E}\varrho_Y) - \omega_\alpha(\widehat{x}, \mathsf{E}\varrho_Y))\right| + \left|\omega_\alpha(\widehat{x}, \mathsf{E}\varrho_Y) - \omega_\alpha(\widehat{x}, \mathsf{E}\widehat{\varrho}_Y)\right| \\ &\quad = \left|\partial_x \omega_\alpha(x + s(\widehat{x} - x), \mathsf{E}\varrho_Y)\right| \|x - \widehat{x}\|_2 + \left|\partial_y \omega_\alpha(\hat{x}, \mathsf{E}\varrho_Y + s'(\mathsf{E}\widehat{\varrho}_Y - \mathsf{E}\varrho_Y))\right| \|\mathsf{E}\widehat{\varrho}_Y - \mathsf{E}\varrho_Y\|_2 \\ &\quad \leq \alpha C_1 e^{-\alpha \underline{\mathcal{E}}(\mathsf{E}\varrho_Y)} 2\left(1 + \|x\|_2 + \|\widehat{x}\|_2 + \left\|\mathsf{E}\varrho_Y\right\|_2\right) \|x - \widehat{x}\|_2 \\ &\qquad + \alpha C_1 e^{-\alpha \underline{\mathcal{E}}(\mathsf{E}\varrho_Y + s'(\mathsf{E}\widehat{\varrho}_Y - \mathsf{E}\varrho_Y))} 2\left(1 + \|\mathsf{E}\widehat{\varrho}_Y\|_2 + \|\widehat{x}\|_2 + \left\|\mathsf{E}\varrho_Y\right\|_2\right) \|\mathsf{E}\widehat{\varrho}_Y - \mathsf{E}\varrho_Y\|_2 \end{aligned}$$

due to Assumptions W2 and W3. With this we immediately obtain the upper bounds

$$\begin{aligned} \|I_2\|_2 &\leq 2\alpha C_1 C_K^\alpha \|\widehat{x}\|_2 \left(1 + \|x\|_2 + \|\mathsf{E}\widehat{\varrho}_Y\|_2 + 2\|\widehat{x}\|_2 + 2\left\|\mathsf{E}\varrho_Y\right\|_2\right) \left(\|x - \widehat{x}\|_2 + \|\mathsf{E}\widehat{\varrho}_Y - \mathsf{E}\varrho_Y\|_2\right), \\ \|I_3\|_2 &\leq 2\alpha C_1 (C_K^\alpha)^2 \|\widehat{x}\|_2 \\ &\qquad \cdot \iint \left(1 + \|x\|_2 + \|\mathsf{E}\widehat{\varrho}_Y\|_2 + 2\|\widehat{x}\|_2 + 2\left\|\mathsf{E}\varrho_Y\right\|_2\right) \left(\|x - \widehat{x}\|_2 + \|\mathsf{E}\widehat{\varrho}_Y - \mathsf{E}\varrho_Y\|_2\right) d\pi(x, y, \widehat{x}, \widehat{y}). \end{aligned}$$

Collecting the latter three estimates for $\|I_1\|_2, \|I_2\|_2$ and $\|I_3\|_2$ eventually gives after an application of Jensen's and Cauchy-Schwarz inequality

$$\begin{aligned} \left\|x_\alpha^Y(\varrho_X) - x_\alpha^Y(\widehat{\varrho}_X)\right\|_2 &\leq C(\alpha, C_1, C_K^\alpha, K) \sqrt{\iint \|x - \widehat{x}\|_2^2 \, d\pi(x, y, \widehat{x}, \widehat{y}) + \|\mathsf{E}\widehat{\varrho}_Y - \mathsf{E}\varrho_Y\|_2^2} \\ &\leq C(\alpha, C_1, C_K^\alpha, K) \sqrt{\iint \|x - \widehat{x}\|_2^2 + \|y - \widehat{y}\|_2^2 \, d\pi(x, y, \widehat{x}, \widehat{y})}, \end{aligned}$$

where the last step is due to Jensen's inequality. $\left\|y_\beta^X(\varrho_Y) - y_\beta^X(\widehat{\varrho}_Y)\right\|_2$ can be bounded analogously. Eventually, optimizing over all couplings $\pi \in \Pi(\varrho, \widehat{\varrho})$ gives the claim. □

Proof of Theorem 6. The proof is based on the Leray-Schauder fixed point theorem and follows in the spirit of [5, Theorems 3.1, 3.2].
Step 1: For given functions $u^X \in \mathcal{C}([0,T],\mathbb{R}^{d_1})$, $u^Y \in \mathcal{C}([0,T],\mathbb{R}^{d_2})$ and measure $\rho_0 \in \mathcal{P}_4(\mathbb{R}^{d_1+d_2})$, according to standard SDE theory [1, Chapter 6], we can uniquely solve the SDE system

$$d\widetilde{X}_t = -\lambda_1\big(\widetilde{X}_t - u_t^X\big)\,dt + \sigma_1 D\big(\widetilde{X}_t - u_t^X\big)\,dB_t^X, \tag{9a}$$
$$d\widetilde{Y}_t = -\lambda_2\big(\widetilde{Y}_t - u_t^Y\big)\,dt + \sigma_2 D\big(\widetilde{Y}_t - u_t^Y\big)\,dB_t^Y \tag{9b}$$

with $(\widetilde{X}_0,\widetilde{Y}_0) \sim \rho_0$ as a consequence of the coefficients being locally Lipschitz continuous and having at most linear growth. This induces $\widetilde{\rho}_t = \mathrm{Law}((\widetilde{X}_t,\widetilde{Y}_t))$. Moreover, the regularity of the initial distribution $\rho_0 \in \mathcal{P}_4(\mathbb{R}^{d_1+d_2})$ allows for a fourth-order moment estimate of the form $\mathbb{E}\big[\|\widetilde{X}_t\|_2^4 + \|\widetilde{Y}_t\|_2^4\big] \le \big(1 + 2\mathbb{E}\big[\|\widetilde{X}_0\|_2^4 + \|\widetilde{Y}_0\|_2^4\big]\big)e^{ct}$, see, e.g. [1, Chapter 7]. So, in particular, $\widetilde{\rho} \in \mathcal{C}([0,T],\mathcal{P}_4(\mathbb{R}^{d_1+d_2}))$, i.e., $\sup_{t\in[0,T]}\iint \|x\|_2^4 + \|y\|_2^4\,d\widetilde{\rho}_t(x,y) \le K$ for some $K>0$.
Step 2: Let us now define the test function space

$$\begin{aligned}\mathcal{C}_*^2(\mathbb{R}^{d_1+d_2}) := \Big\{ &\phi \in \mathcal{C}^2(\mathbb{R}^{d_1+d_2}) :\ \|\nabla\phi(x,y)\|_2 \le C_\phi\big(1+\|x\|_2+\|y\|_2\big) \text{ for some } C_\phi>0 \\ &\text{and } \max\Big\{\max_{k=1,\dots,d_1}\big\|\partial^2_{x_kx_k}\phi\big\|_\infty, \max_{k=1,\dots,d_2}\big\|\partial^2_{y_ky_k}\phi\big\|_\infty\Big\} < \infty\Big\}.\end{aligned} \tag{10}$$

For any $\phi \in \mathcal{C}_*^2(\mathbb{R}^d)$, by Itô's formula, we can derive

$$\begin{aligned} d\phi(\widetilde{X}_t,\widetilde{Y}_t) =\ &\nabla_x\phi(\widetilde{X}_t,\widetilde{Y}_t)\cdot\Big(-\lambda_1\big(\widetilde{X}_t-u_t^X\big)\,dt + \sigma_1 D\big(\widetilde{X}_t-u_t^X\big)\,dB_t^X\Big) \\ &+ \nabla_y\phi(\widetilde{X}_t,\widetilde{Y}_t)\cdot\Big(-\lambda_2\big(\widetilde{Y}_t-u_t^Y\big)\,dt + \sigma_2 D\big(\widetilde{Y}_t-u_t^Y\big)\,dB_t^Y\Big) \\ &+ \frac{\sigma_1^2}{2}\sum_{k=1}^{d_1}\partial^2_{x_kx_k}\phi(\widetilde{X}_t,\widetilde{Y}_t)\big(\widetilde{X}_t-u_t^X\big)_k^2\,dt + \frac{\sigma_2^2}{2}\sum_{k=1}^{d_2}\partial^2_{y_ky_k}\phi(\widetilde{X}_t,\widetilde{Y}_t)\big(\widetilde{Y}_t-u_t^Y\big)_k^2\,dt. \end{aligned}$$

After taking the expectation, applying Fubini's theorem and observing that the stochastic integrals $\mathbb{E}\int_0^t \nabla_x\phi(\widetilde{X}_t,\widetilde{Y}_t)\cdot D\big(\widetilde{X}_t-u_t^X\big)\,dB_t^X$ and $\mathbb{E}\int_0^t \nabla_y\phi(\widetilde{X}_t,\widetilde{Y}_t)\cdot D\big(\widetilde{Y}_t-u_t^Y\big)\,dB_t^Y$ vanish as a consequence of [37, Theorem 3.2.1(iii)] due to the established regularity $\widetilde{\rho}\in\mathcal{C}([0,T],\mathcal{P}_4(\mathbb{R}^{d_1+d_2}))$ and $\phi\in\mathcal{C}_*^2(\mathbb{R}^{d_1+d_2})$, we obtain

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\phi(\widetilde{X}_t,\widetilde{Y}_t) =\ &-\lambda_1\mathbb{E}\nabla_x\phi(\widetilde{X}_t,\widetilde{Y}_t)\cdot\big(\widetilde{X}_t-u_t^X\big) - \lambda_2\mathbb{E}\nabla_y\phi(\widetilde{X}_t,\widetilde{Y}_t)\cdot\big(\widetilde{Y}_t-u_t^Y\big) \\ &+ \frac{\sigma_1^2}{2}\mathbb{E}\sum_{k=1}^{d_1}\partial^2_{x_kx_k}\phi(\widetilde{X}_t,\widetilde{Y}_t)\big(\widetilde{X}_t-u_t^X\big)_k^2 + \frac{\sigma_2^2}{2}\mathbb{E}\sum_{k=1}^{d_2}\partial^2_{y_ky_k}\phi(\widetilde{X}_t,\widetilde{Y}_t)\big(\widetilde{Y}_t-u_t^Y\big)_k^2\end{aligned}$$

as a consequence of the fundamental theorem of calculus. This shows that the measure $\widetilde{\rho}\in\mathcal{C}([0,T],\mathcal{P}_4(\mathbb{R}^{d_1+d_2}))$ satisfies the Fokker-Planck equation

$$\begin{aligned}\frac{d}{dt}\int\phi(x,y)\,d\widetilde{\rho}_t(x,y) =\ &-\int\lambda_1\nabla_x\phi(x,y)\cdot\big(x-u_t^X\big)+\lambda_2\nabla_y\phi(x,y)\cdot\big(y-u_t^Y\big)\,d\widetilde{\rho}_t(x,y) \\ &+\int\frac{\sigma_1^2}{2}\sum_{k=1}^{d_1}\partial^2_{x_kx_k}\phi(x,y)\big(x-u_t^X\big)_k^2+\frac{\sigma_2^2}{2}\sum_{k=1}^{d_2}\partial^2_{y_ky_k}\phi(x,y)\big(y-u_t^Y\big)_k^2\,d\widetilde{\rho}_t(x,y).\end{aligned} \tag{11}$$

Step 3: Setting $\mathcal{T}u := \big((x_\alpha^Y(\widetilde{\rho}_X))^T,(y_\beta^X(\widetilde{\rho}_Y))^T\big)^T \in \mathcal{C}([0,T],\mathbb{R}^{d_1+d_2})$ provides the self-mapping property of the map

$$\mathcal{T}:\mathcal{C}([0,T],\mathbb{R}^{d_1+d_2})\to\mathcal{C}([0,T],\mathbb{R}^{d_1+d_2}),\quad u\mapsto\mathcal{T}u=\big((x_\alpha^Y(\widetilde{\rho}_X))^T,(y_\beta^X(\widetilde{\rho}_Y))^T\big)^T.$$

By means of Lemma 8 we have

$$\big\|x_\alpha^Y(\widetilde{\rho}_{X,t})-x_\alpha^Y(\widetilde{\rho}_{X,s})\big\|_2+\big\|y_\beta^X(\widetilde{\rho}_{Y,t})-y_\beta^X(\widetilde{\rho}_{Y,s})\big\|_2 \le c_0W_2(\widetilde{\rho}_t,\widetilde{\rho}_s)\lesssim c_0\,|t-s|^{\frac{1}{2}},$$

which shows the Hölder-1/2 continuity of $\mathcal{T}$ due to the compact embedding $\mathcal{C}^{0,1/2}([0,T],\mathbb{R}^{d_1+d_2}) \hookrightarrow \mathcal{C}([0,T],\mathbb{R}^{d_1+d_2})$. The last inequality, we note that due to Itô's isometry it holds

$$W_2^2(\widetilde{\rho}_t,\widetilde{\rho}_s) \leq \mathbb{E}\left[\left\|\widetilde{X}_t-\widetilde{X}_s\right\|_2^2+\left\|\widetilde{Y}_t-\widetilde{Y}_s\right\|_2^2\right] \leq 4\big((\lambda_1^2+\lambda_2^2)T+(\sigma_1^2+\sigma_2^2)\big)\Big(K+\left\|u^X\right\|_{L^\infty}^2+\left\|u^Y\right\|_{L^\infty}^2\Big)|t-s|.$$

Step 4: Now, for $u \in \mathcal{C}([0,T],\mathbb{R}^{d_1+d_2})$ satisfying $u = \vartheta\mathcal{T}u$ with $\vartheta \in [0,1]$, there exists $\rho \in \mathcal{C}([0,T],\mathcal{P}_4(\mathbb{R}^{d_1+d_2}))$ satisfying (11) such that $u = \vartheta\big((x_\alpha^Y(\rho_X))^T,(y_\beta^X(\rho_Y))^T\big)^T$. As a consequence of Lemma 7 with $s=0$ we can show that

$$\|u_t\|_2^2 \leq \vartheta^2\left((C_K^\alpha)^2\int\|x\|_2^2\,d\rho_{X,t}(x)+(C_K^\beta)^2\int\|y\|_2^2\,d\rho_{Y,t}(y)\right) \leq \vartheta^2\left((C_K^\alpha)^2+(C_K^\beta)^2\right)\sqrt{K}$$

where we can use $K = \big(1+2\mathbb{E}\big[\|\widetilde{X}_0\|_2^4+\|\widetilde{Y}_0\|_2^4\big]\big)e^{cT}$ as of Step 1. This allows for a uniform estimate of $\|u\|_{L^\infty} < q$ for $q > 0$. An application of the Leray-Schauder fixed point theorem concludes the proof of existence by providing a solution to (3).

Step 5: For uniqueness, suppose we have two fixed points u^1 and u^2 (as specified in the previous step) together with corresponding processes $((\widetilde{X}^1)^T,(\widetilde{Y}^1)^T)^T$ and $((\widetilde{X}^2)^T,(\widetilde{Y}^2)^T)^T$ satisfying (9). Then, taking their difference while keeping the initial conditions and respective Brownian motion paths, we obtain after an application of Itô's isometry and employment of Lemma 8 the bound

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{X}_t^1-\widetilde{X}_t^2\right\|_2^2+\left\|\widetilde{Y}_t^2-\widetilde{Y}_t^2\right\|_2^2\right]\\
&\quad\leq c\mathbb{E}\int_0^t\left\|\widetilde{X}_\tau^1-\widetilde{X}_\tau^2\right\|_2^2+\left\|\widetilde{Y}_\tau^1-\widetilde{Y}_\tau^2\right\|_2^2+\left\|x_\alpha^Y(\widetilde{\rho}_{X,\tau}^1)-x_\alpha^Y(\widetilde{\rho}_{X,\tau}^2)\right\|_2^2+\left\|y_\beta^X(\widetilde{\rho}_{Y,\tau}^1)-y_\beta^X(\widetilde{\rho}_{Y,\tau}^2)\right\|_2^2\,d\tau\\
&\quad\lesssim c\mathbb{E}\int_0^t\left\|\widetilde{X}_\tau^1-\widetilde{X}_\tau^2\right\|_2^2+\left\|\widetilde{Y}_\tau^1-\widetilde{Y}_\tau^2\right\|_2^2\,d\tau
\end{aligned}$$

with $c = 4\left((\lambda_1^2+\lambda_2^2)T+(\sigma_1^2+\sigma_2^2)\right)$. Grönwall's inequality eventually shows uniqueness since $\mathbb{E}\big[\|\widetilde{X}_t^1-\widetilde{X}_t^2\|_2^2+\|\widetilde{Y}_t^2-\widetilde{Y}_t^2\|_2^2\big] = 0$ for all $t \in [0,T]$. □

3 Convergence to Saddle Points

Inspired by the theories of mean-field limits for consensus-based optimization methods (see [11, 14, 20] for instance), the convergence of particles systems (1) to the mean-field dynamics (3) follows in a similar way and thus the associated argument is omitted. In this section we present the main theoretical result of our paper concerned with the convergence of the macroscopic dynamics (3) towards saddle points of objective functions $\mathcal{E}$ that satisfy the following conditions.

Definition 9 (Assumptions)**.** *Throughout this section we are interested in objective functions $\mathcal{E} \in \mathcal{C}^2(\mathbb{R}^{d_1+d_2})$, for which*

A1 there exist two functions $\underline{\mathcal{E}} \in \mathcal{C}^1(\mathbb{R}^{d_2})$ and $\overline{\mathcal{E}} \in \mathcal{C}^1(\mathbb{R}^{d_1})$ such that

$$\underline{\mathcal{E}}(y) \leq \mathcal{E}(x,y) \leq \overline{\mathcal{E}}(x)$$

for all $(x,y) \in \mathbb{R}^{d_1+d_2}$. The functions $\underline{\mathcal{E}}$ and $\overline{\mathcal{E}}$ shall, for some constant $\overline{C}_{\nabla\mathcal{E}} > 0$, satisfy $\|\nabla\underline{\mathcal{E}}(y)\|_2 \leq \overline{C}_{\nabla\mathcal{E}}$ for all $y \in \mathbb{R}^{d_2}$ and $\|\nabla\overline{\mathcal{E}}(x)\|_2 \leq \overline{C}_{\nabla\mathcal{E}}$ for all $x \in \mathbb{R}^{d_1}$.

A2 there exist constants $C_{\nabla\mathcal{E}}, C_{\nabla^2\mathcal{E}} > 0$ such that

$$\begin{aligned}
&\max\left\{\sup_{(x,y)\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}}\left\|\nabla_x\mathcal{E}(x,y)\right\|_2,\sup_{(x,y)\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}}\left\|\nabla_y\mathcal{E}(x,y)\right\|_2\right\} \leq C_{\nabla\mathcal{E}},\ \text{and}\\
&\max\left\{\max_{k=1,\dots,d_1}\left\|\partial_{x_kx_k}^2\mathcal{E}\right\|_\infty,\max_{k=1,\dots,d_2}\left\|\partial_{y_ky_k}^2\mathcal{E}\right\|_\infty,\left\|\rho\left(\nabla_x^2\mathcal{E}\right)\right\|_\infty,\left\|\rho\left(\nabla_y^2\mathcal{E}\right)\right\|_\infty\right\} \leq C_{\nabla^2\mathcal{E}},
\end{aligned}$$

where $\|\cdot\|_\infty$ denotes the L^∞ norm on $\mathcal{C}\left(\mathbb{R}^{d_1+d_2}\right)$ and ρ denotes the spectral radius.

A3 *there exist constants* $\epsilon_0, \eta, \nu > 0$ *such that for each* $(x, y) \in \mathbb{R}^{d_1+d_2}$ *satisfying* $\mathcal{E}^* - \mathcal{E}(x^*, y) \leq \epsilon_0$ *and* $\mathcal{E}(x, y^*) - \mathcal{E}^* \leq \epsilon_0$ *for some saddle point* (x^*, y^*) *of* $\mathcal{E}$*, we have*

$$\|x - x^*\|_2 \leq \frac{1}{\eta}\big(|\mathcal{E}(x, y^*) - \mathcal{E}^*|\big)^\nu \text{ and } \|y - y^*\|_2 \leq \frac{1}{\eta}\big(|\mathcal{E}(x^*, y) - \mathcal{E}^*)|\big)^\nu.$$

Assumption A1 requires that the twice continuously differentiable objective function $\mathcal{E}$ is bounded from below by a function $\underline{\mathcal{E}}$, which depends only on the y coordinate, and from above by a function $\overline{\mathcal{E}}$, which depends only on the x coordinate. Moreover, the first order derivatives of $\underline{\mathcal{E}}$ and $\overline{\mathcal{E}}$ are assumed to be uniformly bounded.

Assumption A2 is a mere technical regularity assumption about $\mathcal{E}$ in terms of the first and second derivatives. In particular, it requires that the gradients as well as second order derivatives of $\mathcal{E}$ are uniformly bounded, which is however necessary only for theoretical analysis of the long time behavior of the algorithm. Analogous regularity assumptions may be found in the literature, see, e.g., [5, 6, 21]. However, as a purely zero-order derivative-free method, our CBO-SP algorithm only uses point-wise values of the objective function $\mathcal{E}$ in practical applications.

Assumption A3 on the other hand should be regarded as a tractability condition on the landscape of the objective function $\mathcal{E}$. It imposes coercivity of $\mathcal{E}$ around saddle points, which relates the distance from (x^*, y^*) with the value of the objective function. We refer to the discussion after [14, Remark 9] for related conditions in the machine learning literature.

In order to formulate in Theorem 11 below the result about the convergence of the dynamics (3) towards saddle points of the objective functions $\mathcal{E}$ satisfying the aforementioned assumption, let us define the variances

$$\mathrm{Var}^X(t) = \mathbb{E}\left\|\overline{X}_t - \mathbb{E}\overline{X}_t\right\|_2^2 \quad \text{and} \quad \mathrm{Var}^Y(t) = \mathbb{E}\left\|\overline{Y}_t - \mathbb{E}\overline{Y}_t\right\|_2^2, \tag{12}$$

which act as Lyapunov functionals of the dynamics. In addition, we require certain well-preparedness assumptions about the initial data and parameters. For this reason we introduce the notations

$$\widetilde{\mathcal{M}}^X(t) := \mathbb{E}\exp\left(-\alpha\mathcal{E}(\overline{X}_t, \mathbb{E}\overline{Y}_t)\right) \quad \text{and} \quad \widetilde{\mathcal{M}}^Y(t) := \mathbb{E}\exp\left(\beta\mathcal{E}(\mathbb{E}\overline{X}_t, \overline{Y}_t)\right), \tag{13a}$$

$$\mathcal{M}^X(t) := \widetilde{\mathcal{M}}^X(t)\, e^{\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)} \quad \text{and} \quad \mathcal{M}^Y(t) := \widetilde{\mathcal{M}}^Y(t)\, e^{-\beta\overline{\mathcal{E}}(\mathbb{E}\overline{X}_t)}, \tag{13b}$$

$$\mathcal{M}^X_*(t) := \mathbb{E}\exp\left(-\alpha\mathcal{E}(\overline{X}_t, y^*)\right) \quad \text{and} \quad \mathcal{M}^Y_*(t) := \mathbb{E}\exp\left(\beta\mathcal{E}(x^*, \overline{Y}_t)\right), \tag{13c}$$

where the definitions of $\mathcal{M}^X_*$ and $\mathcal{M}^Y_*$ in (13c) may be different for potentially different saddle points (x^*, y^*) (since our assumptions allow for the existence of multiple such points), meaning, in particular, that there may be multiple functionals $\mathcal{M}^X_*$ and $\mathcal{M}^Y_*$. This ambiguity is clarified in Remark 12.

Definition 10 (Well-preparedness of the initial data and parameters)**.** *The initial datum* $(\overline{X}_0, \overline{Y}_0)$ *and the parameters* $\alpha, \beta, \lambda_1, \lambda_2, \sigma_1$ *and* σ_2 *of the CBO-SP method are well-prepared if*

P1 $\mu_1 := 2\left(\lambda_1 - 4\sigma_1^2/\mathcal{M}^X(0)\right) > 0$ *and* $\mu_2 := 2\left(\lambda_2 - 4\sigma_2^2/\mathcal{M}^Y(0)\right) > 0$,

P2 *all saddle points* (x^*, y^*) *lie in* $\mathrm{supp}(\rho_0)$*, where* $\rho_0 := \mathrm{Law}(\overline{X}_0, \overline{Y}_0)$ *has marginals* ρ_0^X *and* ρ_0^Y*. Moreover, for any* $\delta > 0$*, there exists some constant* $C_\delta > 0$ *depending only on* δ *such that it holds*

$$\rho_0^X\left(\left\{x : \exp\left(-\mathcal{E}(x, \mathbb{E}(\overline{Y}_0))\right) > \exp\left(-\min_{x\in\mathbb{R}^{d_1}} \mathcal{E}(x, \mathbb{E}(\overline{Y}_0))\right) - \delta\right\}\right) \geq C_\delta,$$

$$\rho_0^Y\left(\left\{y : \exp\left(\mathcal{E}(\mathbb{E}(\overline{X}_0), y)\right) > \exp\left(\max_{y\in\mathbb{R}^{d_2}} \mathcal{E}(\mathbb{E}(\overline{X}_0), y)\right) - \delta\right\}\right) \geq C_\delta$$

as well as

$$\rho_0^X\left(\left\{x : \exp\left(-\mathcal{E}(x, y^*)\right) > \exp\left(-\mathcal{E}^*\right) - \delta\right\}\right) \geq C_\delta,$$

$$\rho_0^Y\left(\left\{y : \exp\left(\mathcal{E}(x^*, y)\right) > \exp\left(\mathcal{E}^*\right) - \delta\right\}\right) \geq C_\delta.$$

P3 *it holds that*

$$4\alpha C_{\nabla^2\mathcal{E}}\left(\lambda_1+\frac{\sigma_1^2}{2}\right)\frac{\mathrm{Var}^X(0)}{\mu_1\mathcal{M}^X(0)}+2\sqrt{2}\alpha\lambda_2 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^Y(0)}}{\mu_2\sqrt{\mathcal{M}^Y(0)}}\leq\frac{1}{8}\mathcal{M}^X(0),$$

$$4\beta C_{\nabla^2\mathcal{E}}\left(\lambda_2+\frac{\sigma_2^2}{2}\right)\frac{\mathrm{Var}^Y(0)}{\mu_2\mathcal{M}^Y(0)}+2\sqrt{2}\beta\lambda_1 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^X(0)}}{\mu_1\sqrt{\mathcal{M}^X(0)}}\leq\frac{1}{8}\mathcal{M}^Y(0),$$

P4 *it holds for any fixed* (x^*,y^*) *that*

$$\begin{aligned}4\alpha\sigma_1^2 C_{\nabla^2\mathcal{E}}\left(\lambda_1+\frac{\sigma_1^2}{2}\right)\frac{\mathrm{Var}^X(0)}{\mu_1\mathcal{M}^X(0)}+8\alpha\lambda_1 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^X(0)}}{\mu_1\sqrt{\mathcal{M}^X(0)}}+\sqrt{2}\alpha\lambda_2 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^Y(0)}}{\mu_2\sqrt{\mathcal{M}^Y(0)}}\\ \leq\frac{1}{8}\min\left\{\mathcal{M}_*^X(0)e^{\alpha\underline{\mathcal{E}}(y^*)},\widetilde{\mathcal{M}}^X(0)e^{-\alpha\mathcal{E}_M}\right\},\\ 4\beta C_{\nabla^2\mathcal{E}}\left(\lambda_2+\frac{\sigma_2^2}{2}\right)\frac{\mathrm{Var}^Y(0)}{\mu_2\mathcal{M}^Y(0)}+8\beta\lambda_2 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^Y(0)}}{\mu_2\sqrt{\mathcal{M}^Y(0)}}+\sqrt{2}\beta\lambda_1 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^X(0)}}{\mu_1\sqrt{\mathcal{M}^X(0)}}\\ \leq\frac{1}{8}\min\left\{\mathcal{M}_*^Y(0)e^{-\beta\bar{\mathcal{E}}(x^*)},\widetilde{\mathcal{M}}^Y(0)e^{-\beta\mathcal{E}_M}\right\},\end{aligned}$$

where $\mathcal{E}_M$ *depends on* $(\widetilde{x},\widetilde{y})$ *as well as* $\underline{\mathcal{E}}$ *and* $\bar{\mathcal{E}}$ *(see Theorem 11 below).*

Condition P1 can be satisfied appropriately big choices of λ_1 and λ_2. Condition P2 is valid if the initial distribution ρ_0 has some mass at the saddle points (x^*,y^*). While this may have a certain locality flavor, in the case of functions $\mathcal{E}$ having multiple saddle points, the condition is generically satisfied at least for one of them, allowing essentially to obtain a global result. It is actually sufficient if there is at least one saddle point satisfying Condition P2, in which case the method is agnostic to saddle points not in $\mathrm{supp}(\rho_0)$. Conditions P3–P4 on the other hand may be ensured if the initial variances $\mathrm{Var}^X(0)$ and $\mathrm{Var}^Y(0)$ are sufficiently small. Well-preparedness conditions similar to P2–P4 can be found in the literature about a convergence analysis of CBO for minimization, see, e.g., [5, 6, 12, 21], while we note that the coupling of $(\overline{X},\overline{Y})$ due to the intrinsic difference between games and optimizations prompts us to use some different techniques in the proof.

We are now ready to state the result about the convergence of the dynamics (3) towards saddle points of the objective functions $\mathcal{E}$. The proof details are deferred to Section 4.

Theorem 11. *Let* $\mathcal{E}$ *satisfy Assumptions A1 and A2 and let* $(\overline{X}_t,\overline{Y}_t)_{t\geq0}$ *be a solution to the SDE (3). Then the following statements hold.*

(1) *Under the assumption of well-preparedness of the initial datum* $(\overline{X}_0,\overline{Y}_0)$ *and the parameters* $\alpha,\beta,\lambda_1,\lambda_2,\sigma_1$ *and* σ_2 *in the sense of P1–P3,* Var^X *and* Var^Y *as defined in (12) converge exponentially fast to* 0 *as* $t\to\infty$*. More precisely, it holds*

$$\mathrm{Var}^X(t)+\mathrm{Var}^Y(t)\leq\mathrm{Var}^X(0)e^{-\mu_1 t}+\mathrm{Var}^Y(0)e^{-\mu_2 t}. \tag{14}$$

Moreover, there exists some $(\widetilde{x},\widetilde{y})$ *depending in particular on* α *and* β *such that*

$$\left(\mathbb{E}\overline{X}_t,\mathbb{E}\overline{Y}_t\right)\to(\widetilde{x},\widetilde{y})\quad\text{and}\quad\left(x_\alpha^Y(\rho_{X,t}),y_\beta^X(\rho_{Y,t})\right)\to(\widetilde{x},\widetilde{y}) \tag{15}$$

as $t\to\infty$*.*

(2) *For any given accuracy* $\varepsilon>0$*, there exist some* $\alpha_0,\beta_0>0$ *such that for all* $\alpha\geq\alpha_0$ *and* $\beta\geq\beta_0$ *the point* $(\widetilde{x},\widetilde{y})$ *from (1) (which may depend on* α *and* β*) satisfies*

$$|\mathcal{E}(\widetilde{x},\widetilde{y})-\mathcal{E}^*|\leq\varepsilon\quad\text{as well as}\quad\mathcal{E}^*-\mathcal{E}(x^*,\widetilde{y})\leq\varepsilon\quad\text{and}\quad\mathcal{E}(\widetilde{x},y^*)-\mathcal{E}^*\leq\varepsilon \tag{16}$$

provided that the well-preparedness Assumptions P1–P4 hold for such α *and* β *together with the initial datum* $(\overline{X}_0,\overline{Y}_0)$*.*

(3) *If $\mathcal{E}$ satisfies Assumption A3 with respect to $(\widetilde{x},\widetilde{y})$ from (2) with $\varepsilon \leq \epsilon_0$, i.e., there exists some saddle point (x^*,y^*) depending on $(\widetilde{x},\widetilde{y})$ such that*

$$\|\widetilde{x}-x^*\|_2 \leq \frac{1}{\eta}\left(|\mathcal{E}(\widetilde{x},y^*)-\mathcal{E}^*|\right)^\nu \quad \text{and} \quad \|\widetilde{y}-y^*\|_2 \leq \frac{1}{\eta}\left(|\mathcal{E}(x^*,\widetilde{y})-\mathcal{E}^*|\right)^\nu,$$

then we have

$$\|(\widetilde{x},\widetilde{y})-(x^*,y^*)\|_2 \leq \frac{2}{\eta}\varepsilon^\nu \tag{17}$$

provided that the well-preparedness Assumptions P1–P4 hold for sufficiently large α and β together with the initial datum $(\overline{X}_0,\overline{Y}_0)$.

The first part (1) of Theorem 11 states that under suitable well-preparedness conditions on the initialization and the parameters, the mean-field dynamics (3) reaches consensus at *some* location $(\widetilde{x},\widetilde{y})$, which may depend in particular on α and β, as time evolves. In the second part (2) of the statement, for sufficiently large α and β as well as under certain well-preparedness conditions, properties of *the* corresponding point $(\widetilde{x},\widetilde{y})$ are specified which are typical for saddle points, see (16). These properties eventually allow to conclude in the third part (3) of the result, that *the* $(\widetilde{x},\widetilde{y})$ from before is arbitrarily close to *any* saddle point (x^*,y^*) which satisfies the inverse continuity property A3.

Remark 12. *It is worth mentioning at this point that in order to prove (16), any saddle point (x^*,y^*) satisfying Assumption P4 can be used in the definitions of $\mathcal{M}_*^X$ and $\mathcal{M}_*^Y$. For the proof of (17), on the other hand, it is necessary to use in the definitions of $\mathcal{M}_*^X$ and $\mathcal{M}_*^Y$ a specific saddle point (x^*,y^*) that satisfies the inverse continuity property A3 with respect to $(\widetilde{x},\widetilde{y})$ as well as Assumption P4, so, in this case, the saddle point (x^*,y^*) does depend on $(\widetilde{x},\widetilde{y})$.*

4 Proof Details for Section 3

In this section we provide the proof details for the convergence result of the mean-field dynamics (3) to a saddle point of the objective function $\mathcal{E}$. Sections 4.1–4.3 present individual statements which are necessary in the proof of our main theorem, Theorem 11, which is then given in Section 4.4.

4.1 Time-Evolution of the Variances Var^X and Var^Y

In order to ensure consensus formation of the mean-field dynamics (3) we show that the variances $\mathrm{Var}^X(t)=\mathbb{E}\|\overline{X}_t-\mathbb{E}\overline{X}_t\|_2^2$ and $\mathrm{Var}^Y(t)=\mathbb{E}\|\overline{Y}_t-\mathbb{E}\overline{Y}_t\|_2^2$ of the particle distribution decay to 0 as $t\to\infty$. For this we need to analyze their time-evolutions, which is the content of the following statement.

Lemma 13. *Let Var^X and Var^Y be as defined in (12) and let us recall the definitions of $\mathcal{M}^X$ and $\mathcal{M}^Y$ from (13b). Then, under Assumption A1, it holds*

$$\frac{d}{dt}\mathrm{Var}^X(t) \leq -2\left(\lambda_1-\frac{2\sigma_1^2}{\mathcal{M}^X(t)}\right)\mathrm{Var}^X(t) \quad \text{and} \quad \frac{d}{dt}\mathrm{Var}^Y(t) \leq -2\left(\lambda_2-\frac{2\sigma_2^2}{\mathcal{M}^Y(t)}\right)\mathrm{Var}^Y(t). \tag{18}$$

Proof. By means of Itô's calculus we have

$$d\left\|\overline{X}_t-\mathbb{E}\overline{X}_t\right\|_2^2 = 2\left(\overline{X}_t-\mathbb{E}\overline{X}_t\right)\cdot d\overline{X}_t + \sigma_1^2\left\|\overline{X}_t-x_\alpha^Y(\rho_{X,t})\right\|_2^2 dt.$$

Since the appearing stochastic integral vanishes when taking the expectation as a consequence of the regularity established in Theorem 6 and Assumption A1, we obtain

$$\begin{aligned}\frac{d}{dt}\mathrm{Var}^X(t) &= -2\lambda_1\mathbb{E}\left[\left(\overline{X}_t-\mathbb{E}\overline{X}_t\right)\cdot\left(\overline{X}_t-x_\alpha^Y(\rho_{X,t})\right)\right]+\sigma_1^2\mathbb{E}\left\|\overline{X}_t-x_\alpha^Y(\rho_{X,t})\right\|_2^2\\ &= -2\lambda_1\mathrm{Var}^X(t)+\sigma_1^2\mathbb{E}\left\|\overline{X}_t-x_\alpha^Y(\rho_{X,t})\right\|_2^2,\end{aligned} \tag{19}$$

where we used that $\mathbb{E}\left[\left(\overline{X}_t - \mathbb{E}\overline{X}_t\right)\cdot\left(\mathbb{E}\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right)\right] = 0$. Analogously, we derive

$$\frac{d}{dt}\mathrm{Var}^Y(t) = -2\lambda_2\mathrm{Var}^Y(t) + \sigma_2^2\mathbb{E}\left\|\overline{Y}_t - y_\beta^X(\rho_{Y,t})\right\|_2^2. \tag{20}$$

In order to control the terms $\mathbb{E}\left\|\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right\|_2^2$ and $\mathbb{E}\left\|\overline{Y}_t - y_\beta^X(\rho_{Y,t})\right\|_2^2$ appearing in (19) and (20), let us first observe that, for any $\widehat{x}\in\mathbb{R}^{d_1}$, $\widehat{y}\in\mathbb{R}^{d_2}$, Jensen's inequality gives

$$\left\|\widehat{x} - x_\alpha^Y(\rho_{X,t})\right\|_2^2 \le \frac{1}{\mathbb{E}\omega_\alpha(\overline{X}_t,\mathbb{E}\overline{Y}_t)}\int\|\widehat{x}-x\|_2^2\,\omega_\alpha(x,\mathbb{E}\overline{Y}_t)\,d\rho_{X,t}(x), \tag{21}$$

$$\left\|\widehat{y} - y_\beta^X(\rho_{Y,t})\right\|_2^2 \le \frac{1}{\mathbb{E}\omega_{-\beta}(\mathbb{E}\overline{X}_t,\overline{Y}_t)}\int\|\widehat{y}-y\|_2^2\,\omega_{-\beta}(\mathbb{E}\overline{X}_t,y)\,d\rho_{Y,t}(y). \tag{22}$$

Exploiting the boundedness of $\mathcal{E}$ as of Assumption A1, the two latter bounds in particular imply

$$\begin{aligned}\mathbb{E}\left\|\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right\|_2^2 &\le \frac{2}{\mathbb{E}\omega_\alpha(\overline{X}_t,\mathbb{E}\overline{Y}_t)}\int\left(\mathbb{E}\left\|\overline{X}_t-\mathbb{E}\overline{X}_t\right\|_2^2 + \left\|\mathbb{E}\overline{X}_t - x\right\|_2^2\right)\omega_\alpha(x,\mathbb{E}\overline{Y}_t)\,d\rho_{X,t}(x)\\ &\le 2\mathrm{Var}^X(t) + \frac{2}{\mathbb{E}\omega_\alpha(\overline{X}_t,\mathbb{E}\overline{Y}_t)}e^{-\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}\mathrm{Var}^X(t) \le 4\frac{\mathrm{Var}^X(t)}{\mathcal{M}^X(t)}\end{aligned} \tag{23}$$

and analogously

$$\mathbb{E}\left\|\overline{Y}_t - y_\beta^X(\rho_{Y,t})\right\|_2^2 \le 4\frac{\mathrm{Var}^Y(t)}{\mathcal{M}^Y(t)}, \tag{24}$$

which allow to conclude the proof when being inserted into (19) and (20), respectively. □

4.2 Time-Evolution of the Functionals $\mathcal{M}^X$ and $\mathcal{M}^Y$ from (13b)

In the time-evolutions (18) of the variances Var^X and Var^Y there appear the functionals $\mathcal{M}^X$ and $\mathcal{M}^Y$ as defined in (13b), which need to be controlled in order to ensure that the decay rates can be bounded from below by a positive constant, which eventually leads to at least exponential decay of the variances and therefore consensus of the dynamics (3). We therefore investigate the evolutions of $\mathcal{M}^X$ and $\mathcal{M}^Y$ next. To do so, let us recall from (13b) that $\mathcal{M}^X(t) = \widetilde{\mathcal{M}}^X(t)\,e^{\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}$ and $\mathcal{M}^Y(t) = \widetilde{\mathcal{M}}^Y(t)\,e^{-\beta\bar{\mathcal{E}}(\mathbb{E}\overline{X}_t)}$. We first bound in Lemma 14 the evolutions of $\widetilde{\mathcal{M}}^X$ and $\widetilde{\mathcal{M}}^Y$ as defined in (13a), before we use product rule to obtain a lower bound for the evolutions of $\mathcal{M}^X$ and $\mathcal{M}^Y$ in Lemma 15.

Let us furthermore remark, that $\widetilde{\mathcal{M}}^X$ and $\widetilde{\mathcal{M}}^Y$ will later allow to characterize the convergence point of the dynamics (3).

Lemma 14. *Let* Var^X *and* Var^Y *be as defined in* (12), *and* $\widetilde{\mathcal{M}}^X$ *and* $\widetilde{\mathcal{M}}^Y$ *as in* (13a). *Then, under Assumptions A1 and A2, it holds*

$$\frac{d}{dt}\widetilde{\mathcal{M}}^X(t) \ge -4\alpha e^{-\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}C_{\nabla^2\mathcal{E}}\left(\lambda_1 + \frac{\sigma_1^2}{2}\right)\frac{\mathrm{Var}^X(t)}{\mathcal{M}^X(t)} - \alpha\lambda_2 e^{-\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^Y(t)}}{\sqrt{\mathcal{M}^Y(t)}} \tag{25}$$

as well as

$$\frac{d}{dt}\widetilde{\mathcal{M}}^Y(t) \ge -4\beta e^{\beta\bar{\mathcal{E}}(\mathbb{E}\overline{X}_t)}C_{\nabla^2\mathcal{E}}\left(\lambda_2 + \frac{\sigma_2^2}{2}\right)\frac{\mathrm{Var}^Y(t)}{\mathcal{M}^Y(t)} - \beta\lambda_1 e^{\beta\bar{\mathcal{E}}(\mathbb{E}\overline{X}_t)}C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^X(t)}}{\sqrt{\mathcal{M}^X(t)}}. \tag{26}$$

Proof. With Itô's formula and chain rule we first note that

$$\begin{aligned}d\widetilde{\mathcal{M}}^X(t) &= -\alpha\mathbb{E}\left[\exp\left(-\alpha\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\right)\nabla_x\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\cdot d\overline{X}_t\right]\\ &+ \frac{\sigma_1^2}{2}\sum_{k=1}^{d_1}\mathbb{E}\Big[\exp\left(-\alpha\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\right)\left(\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right)_k^2\cdot\Big(\alpha^2\left(\partial_{x_k}\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\right)^2 - \alpha\partial_{x_kx_k}^2\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\Big)\Big]\,dt\\ &- \alpha\mathbb{E}\left[\exp\left(-\alpha\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\right)\nabla_y\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\cdot d\mathbb{E}\overline{Y}_t\right] =: (T_1+T_2+T_3)\,dt,\end{aligned}$$

where for the definition in the last step we exploited that the appearing stochastic integrals have expectation 0 as a consequence of the regularity established in Theorem 6 and Assumptions A1 and A2. Noticing that $\mathbb{E}\left[\exp\left(-\alpha\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\right)\left(\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right)\right] = 0$ and $\nabla_x\mathcal{E}(x_\alpha^Y(\rho_{X,t}),\mathbb{E}\overline{Y}_t)$ is deterministic, we obtain for T_1 the lower bound

$$\begin{aligned}
T_1 &= \alpha\lambda_1\mathbb{E}\left[\exp\left(-\alpha\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\right)\nabla_x\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\cdot\left(\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right)\right]\\
&= \alpha\lambda_1\mathbb{E}\left[\exp\left(-\alpha\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\right)\left(\nabla_x\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t) - \nabla_x\mathcal{E}(x_\alpha^Y(\rho_{X,t}),\mathbb{E}\overline{Y}_t)\right)\cdot\left(\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right)\right]\\
&\geq -\alpha\lambda_1 e^{-\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}C_{\nabla^2\mathcal{E}}\mathbb{E}\left\|\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right\|_2^2,
\end{aligned}$$

where we made use of the assumptions again. For T_2 it holds

$$\begin{aligned}
T_2 &\geq -\alpha\frac{\sigma_1^2}{2}\sum_{k=1}^{d_1}\mathbb{E}\left[\exp\left(-\alpha\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\right)\left(\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right)_k^2\partial^2_{x_kx_k}\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\right]dt\\
&\geq -\alpha\frac{\sigma_1^2}{2}e^{-\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}C_{\nabla^2\mathcal{E}}\mathbb{E}\left\|\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right\|_2^2.
\end{aligned}$$

And, eventually, for T_3 we have the following bound from below

$$\begin{aligned}
T_3 &= \alpha\lambda_2\mathbb{E}\left[\exp\left(-\alpha\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\right)\nabla_y\mathcal{E}(\overline{X}_t,\mathbb{E}\overline{Y}_t)\cdot\left(\mathbb{E}\overline{Y}_t - y_\beta^X(\rho_{Y,t})\right)\right]\\
&\geq -\alpha\lambda_2 e^{-\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}C_{\nabla\mathcal{E}}\left\|\mathbb{E}\overline{Y}_t - y_\beta^X(\rho_{Y,t})\right\|_2,
\end{aligned}$$

where we used the bounds on the gradient of $\mathcal{E}$ required through Assumption A1 in the last step. Collecting the estimates for T_1, T_2 and T_3, and inserting them into the first equation gives

$$\frac{d}{dt}\widetilde{\mathcal{M}}^X(t) \geq -\alpha e^{-\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}C_{\nabla^2\mathcal{E}}\left(\lambda_1 + \frac{\sigma_1^2}{2}\right)\mathbb{E}\left\|\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right\|_2^2 - \alpha\lambda_2 e^{-\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}C_{\nabla\mathcal{E}}\left\|\mathbb{E}\overline{Y}_t - y_\beta^X(\rho_{Y,t})\right\|_2.$$

The two appearing norms can be bounded by recalling (23) and noticing that (22) gives

$$\left\|\mathbb{E}\overline{Y}_t - y_\beta^X(\rho_{Y,t})\right\|_2^2 \leq \frac{1}{\mathbb{E}\omega_{-\beta}(\mathbb{E}\overline{X}_t,\overline{Y}_t)}e^{\beta\overline{\mathcal{E}}(\mathbb{E}\overline{X}_t)}\int\left\|\mathbb{E}\overline{Y}_t - y\right\|_2^2 d\rho_{Y,t}(y) \leq \frac{\mathrm{Var}^Y(t)}{\mathcal{M}^Y(t)}. \tag{27}$$

Inserting these two latter estimates allows to continue the former as desired as

$$\frac{d}{dt}\widetilde{\mathcal{M}}^X(t) \geq -4\alpha e^{-\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}C_{\nabla^2\mathcal{E}}\left(\lambda_1 + \frac{\sigma_1^2}{2}\right)\frac{\mathrm{Var}^X(t)}{\mathcal{M}^X(t)} - \alpha\lambda_2 e^{-\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^Y(t)}}{\sqrt{\mathcal{M}^Y(t)}}. \tag{28}$$

The estimate for $\frac{d}{dt}\widetilde{\mathcal{M}}^Y(t)$ can be obtained analogously. □

As mentioned already before we derive in the next lemma the time-evolutions of the functionals $\mathcal{M}^X$ and $\mathcal{M}^Y$ as defined in (13b). This is an immediate consequence of product rule and Lemma 14.

Lemma 15. *Let* Var^X *and* Var^Y *be as defined in* (12)*, and* $\mathcal{M}^X$ *and* $\mathcal{M}^Y$ *as in* (13b)*. Then, under Assumptions A1 and A2, it holds*

$$\frac{d}{dt}\mathcal{M}^X(t) \geq -4\alpha C_{\nabla^2\mathcal{E}}\left(\lambda_1 + \frac{\sigma_1^2}{2}\right)\frac{\mathrm{Var}^X(t)}{\mathcal{M}^X(t)} - 2\alpha\lambda_2 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^Y(t)}}{\sqrt{\mathcal{M}^Y(t)}} \tag{29}$$

as well as

$$\frac{d}{dt}\mathcal{M}^Y(t) \geq -4\beta C_{\nabla^2\mathcal{E}}\left(\lambda_2 + \frac{\sigma_2^2}{2}\right)\frac{\mathrm{Var}^Y(t)}{\mathcal{M}^Y(t)} - 2\beta\lambda_1 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^X(t)}}{\sqrt{\mathcal{M}^X(t)}}. \tag{30}$$

Proof. By product rule we have $\frac{d}{dt}\mathcal{M}^X(t) = e^{\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}\frac{d}{dt}\widetilde{\mathcal{M}}^X(t) + \widetilde{\mathcal{M}}^X(t)\,\frac{d}{dt}e^{\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}$. While the first summand is controlled by recalling Lemma 14, for the second term we straightforwardly compute

$$\frac{d}{dt}e^{\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)} = \alpha e^{\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}\nabla\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)\cdot\frac{d}{dt}\mathbb{E}\overline{Y}_t \geq -\alpha\lambda_2 e^{\alpha\underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)}C_{\nabla\mathcal{E}}\left\|\mathbb{E}\overline{Y}_t - y_\beta^X(\rho_{Y,t})\right\|_2,$$

where we used the bounds on the gradient of $\underline{\mathcal{E}}$ required through Assumption A1 together with the regularity from Theorem 6. Recalling (27) and putting everything together yields

$$\frac{d}{dt}\mathcal{M}^X(t) \geq -4\alpha C_{\nabla^2\mathcal{E}}\left(\lambda_1 + \frac{\sigma_1^2}{2}\right)\frac{\mathrm{Var}^X(t)}{\mathcal{M}^X(t)} - \left(1+\mathcal{M}^X(t)\right)\alpha\lambda_2 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^Y(t)}}{\sqrt{\mathcal{M}^Y(t)}},$$

which gives the claim after noting that $\mathcal{M}^X(t)\leq 1$. The proceeding for $\frac{d}{dt}\mathcal{M}^Y(t)$ is identical. □

4.3 Time-Evolution of the Functionals $\mathcal{M}_*^X$ and $\mathcal{M}_*^Y$ from (13c)

Similarly to the preceding sections we study the time-evolution of two functionals $\mathcal{M}_*^X$ and $\mathcal{M}_*^Y$ as defined in (13c), which aids to prove properties of the limit point of the mean-field dynamics (3).

Lemma 16. *Let* Var^X *and* Var^Y *be as defined in* (12), $\mathcal{M}^X$ *and* $\mathcal{M}^Y$ *as in* (13b), *and* $\mathcal{M}_*^X$ *and* $\mathcal{M}_*^Y$ *as in* (13c). *Then, under Assumptions A1 and A2, it holds*

$$\frac{d}{dt}\mathcal{M}_*^X(t) \geq -4\alpha\lambda_1 e^{-\alpha\underline{\mathcal{E}}(y^*)}C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^X(t)}}{\sqrt{\mathcal{M}^X(t)}} - 2\alpha\sigma_1^2 e^{-\alpha\underline{\mathcal{E}}(y^*)}C_{\nabla^2\mathcal{E}}\frac{\mathrm{Var}^X(t)}{\mathcal{M}^X(t)} \tag{31}$$

as well as

$$\frac{d}{dt}\mathcal{M}_*^Y(t) \geq -4\beta\lambda_2 e^{\beta\bar{\mathcal{E}}(x^*)}C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^Y(t)}}{\sqrt{\mathcal{M}^Y(t)}} - 2\beta\sigma_2^2 e^{\beta\bar{\mathcal{E}}(x^*)}C_{\nabla^2\mathcal{E}}\frac{\mathrm{Var}^Y(t)}{\mathcal{M}^Y(t)}. \tag{32}$$

Proof. With Itô's formula and chain rule we first note that

$$\begin{aligned} d\mathcal{M}_*^X(t) = &-\alpha\mathbb{E}\left[\exp\left(-\alpha\mathcal{E}(\overline{X}_t,y^*)\right)\nabla_x\mathcal{E}(\overline{X}_t,y^*)\cdot d\overline{X}_t\right] + \frac{\sigma_1^2}{2}\sum_{k=1}^{d_1}\mathbb{E}\left[\exp\left(-\alpha\mathcal{E}(\overline{X}_t,y^*)\right)\cdot\right.\\ &\left.\cdot\left(\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right)_k^2\left(\alpha^2\left(\partial_{x_k}\mathcal{E}(\overline{X}_t,y^*)\right)^2 - \alpha\partial_{x_kx_k}^2\mathcal{E}(\overline{X}_t,y^*)\right)\right]dt =: (T_1+T_2)\,dt, \end{aligned} \tag{33}$$

where for the definition in the last step we again exploited that the appearing stochastic integral has expectation 0 as a consequence of the assumptions. For T_1 we have the lower bound

$$T_1 \geq -\alpha\lambda_1 e^{-\alpha\underline{\mathcal{E}}(y^*)}\mathbb{E}\left[\left\|\nabla_x\mathcal{E}(\overline{X}_t,y^*)\right\|_2\left\|\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right\|_2\right] \geq -\alpha\lambda_1 e^{-\alpha\underline{\mathcal{E}}(y^*)}C_{\nabla\mathcal{E}}\sqrt{\mathbb{E}\left\|\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right\|_2^2}.$$

For T_2 it holds

$$\begin{aligned} T_2 &\geq -\alpha\frac{\sigma_1^2}{2}\sum_{k=1}^{d_1}\mathbb{E}\left[\exp\left(-\alpha\mathcal{E}(\overline{X}_t,y^*)\right)\left(\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right)_k^2\partial_{x_kx_k}^2\mathcal{E}(\overline{X}_t,y^*)\right]dt\\ &\geq -\alpha\frac{\sigma_1^2}{2}e^{-\alpha\underline{\mathcal{E}}(y^*)}C_{\nabla^2\mathcal{E}}\mathbb{E}\left\|\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right\|_2^2. \end{aligned}$$

Collecting the two former estimates for the terms T_1 and T_2, and inserting them into (33) gives

$$\frac{d}{dt}\mathcal{M}_*^X(t) \geq -\alpha\lambda_1 e^{-\alpha\underline{\mathcal{E}}(y^*)}C_{\nabla\mathcal{E}}\sqrt{\mathbb{E}\left\|\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right\|_2^2} - \alpha\frac{\sigma_1^2}{2}e^{-\alpha\underline{\mathcal{E}}(y^*)}C_{\nabla^2\mathcal{E}}\mathbb{E}\left\|\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right\|_2^2, \tag{34}$$

where the last expression can be bounded by employing (23). The estimate for $\frac{d}{dt}\mathcal{M}_*^Y(t)$ can be obtained analogously. □

4.4 Proof of Theorem 11

Proof of Theorem 11. **Step 1a:** Let us define the time horizon

$$T := \inf\left\{t \geq 0 : \mathcal{M}^X(t) < \frac{1}{2}\mathcal{M}^X(0) \text{ or } \mathcal{M}^Y(t) < \frac{1}{2}\mathcal{M}^Y(0)\right\} \quad \text{with } \inf\emptyset = \infty, \tag{35}$$

where $\mathcal{M}^X$ and $\mathcal{M}^Y$ are as defined in (13b). Obviously, by continuity, $T > 0$. We claim that $T = \infty$, which is shown by contradiction in what follows. Therefore, let us assume $T < \infty$. Then, as a consequence of the definition of the time horizon T, the prefactors of $\mathrm{Var}^X(t)$ and $\mathrm{Var}^Y(t)$ in Lemma 13 are upper bounded by $-\mu_1$ and $-\mu_2$, respectively, for all $t \in [0, T]$. Consequently, Lemma 13 permits the upper bounds

$$\frac{d}{dt}\mathrm{Var}^X(t) \leq -\mu_1 \mathrm{Var}^X(t) \quad \text{and} \quad \frac{d}{dt}\mathrm{Var}^Y(t) \leq -\mu_2 \mathrm{Var}^Y(t) \tag{36}$$

for the time-evolution of the functionals Var^X and Var^Y. The negativity of the rate is ensured by the well-preparedness condition P1. An application of Grönwall's inequality gives

$$\mathrm{Var}^X(t) \leq \mathrm{Var}^X(0)e^{-\mu_1 t} \quad \text{and} \quad \mathrm{Var}^Y(t) \leq \mathrm{Var}^Y(0)e^{-\mu_2 t}. \tag{37}$$

Let us now derive the contradiction. It follows from Lemma 15 for $\mathcal{M}^X$ and $\mathcal{M}^Y$ from (13b) that

$$\begin{aligned}
\frac{d}{dt}\mathcal{M}^X(t) &\geq -8\alpha C_{\nabla^2\mathcal{E}}\left(\lambda_1 + \frac{\sigma_1^2}{2}\right)\frac{\mathrm{Var}^X(0)e^{-\mu_1 t}}{\mathcal{M}^X(0)} - 2\sqrt{2}\alpha\lambda_2 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^Y(0)}e^{-\mu_2 t/2}}{\sqrt{\mathcal{M}^Y(0)}}, \\
\frac{d}{dt}\mathcal{M}^Y(t) &\geq -8\beta C_{\nabla^2\mathcal{E}}\left(\lambda_2 + \frac{\sigma_2^2}{2}\right)\frac{\mathrm{Var}^Y(0)e^{-\mu_2 t}}{\mathcal{M}^Y(0)} - 2\sqrt{2}\beta\lambda_1 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^X(0)}e^{-\mu_1 t/2}}{\sqrt{\mathcal{M}^X(0)}}
\end{aligned} \tag{38}$$

where we used the formerly derived (37) as well as that $\mathcal{M}^X(t) \geq \mathcal{M}^X(0)/2$ and $\mathcal{M}^Y(t) \geq \mathcal{M}^Y(0)/2$ for all $t \in [0, T]$ by definition of T. Integrating (38) and employing the well-preparedness condition P3 shows for all $t \in [0, T]$

$$\begin{aligned}
\mathcal{M}^X(t) &\geq \mathcal{M}^X(0) - 8\alpha C_{\nabla^2\mathcal{E}}\left(\lambda_1 + \frac{\sigma_1^2}{2}\right)\frac{\mathrm{Var}^X(0)}{\mu_1\mathcal{M}^X(0)} - 4\sqrt{2}\alpha\lambda_2 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^Y(0)}}{\mu_2\sqrt{\mathcal{M}^Y(0)}} \geq \frac{3}{4}\mathcal{M}^X(0), \\
\mathcal{M}^Y(t) &\geq \mathcal{M}^Y(0) - 8\beta C_{\nabla^2\mathcal{E}}\left(\lambda_2 + \frac{\sigma_2^2}{2}\right)\frac{\mathrm{Var}^Y(0)}{\mu_2\mathcal{M}^Y(0)} - 4\sqrt{2}\beta\lambda_1 C_{\nabla\mathcal{E}}\frac{\sqrt{\mathrm{Var}^X(0)}}{\mu_1\sqrt{\mathcal{M}^X(0)}} \geq \frac{3}{4}\mathcal{M}^Y(0).
\end{aligned}$$

This entails that there exists $\delta > 0$ such that $\mathcal{M}^X(t) \geq \mathcal{M}^X(0)/2$ and $\mathcal{M}^Y(t) \geq \mathcal{M}^Y(0)/2$ hold for all $t \in [T, T + \delta]$ as well, contradicting the definition of T and therefore showing $T = \infty$. Consequently (37) as well as

$$\mathcal{M}^X(t) \geq \frac{1}{2}\mathcal{M}^X(0) \quad \text{and} \quad \mathcal{M}^Y(t) \geq \frac{1}{2}\mathcal{M}^Y(0) \tag{39}$$

hold for all $t \geq 0$, which proves (14).

Step 1b: With Jensen's inequality and by making use of the bounds (23) and (24) combined with (37) and (39) we further observe that

$$\begin{aligned}
\left\|\frac{d}{dt}\mathbb{E}\overline{X}_t\right\|_2 &\leq \lambda_1\mathbb{E}\left\|\overline{X}_t - x_\alpha^Y(\rho_{X,t})\right\|_2 \leq 2\lambda_1\frac{\sqrt{\mathrm{Var}^X(0)}e^{-\mu_1 t/2}}{\sqrt{\mathcal{M}^X(0)}} \to 0 \quad \text{as } t \to \infty, \\
\left\|\frac{d}{dt}\mathbb{E}\overline{Y}_t\right\|_2 &\leq \lambda_2\mathbb{E}\left\|\overline{Y}_t - y_\beta^X(\rho_{Y,t})\right\|_2 \leq 2\lambda_2\frac{\sqrt{\mathrm{Var}^Y(0)}e^{-\mu_2 t/2}}{\sqrt{\mathcal{M}^Y(0)}} \to 0 \quad \text{as } t \to \infty.
\end{aligned}$$

We therefore have $\left(\mathbb{E}\overline{X}_t, \mathbb{E}\overline{Y}_t\right) \to (\widetilde{x}, \widetilde{y})$ for *some* $(\widetilde{x}, \widetilde{y}) \in \mathbb{R}^{d_1+d_2}$. In fact, following from (37), $\left(\overline{X}_t, \overline{Y}_t\right) \to (\widetilde{x}, \widetilde{y})$ and $\left(x^Y_\alpha(\rho_{X,t}), y^X_\beta(\rho_{Y,t})\right) \to (\widetilde{x}, \widetilde{y})$ in L^2 thanks to (23) and (24). This shows (15).
Step 2a: It remains to verify (16) for the point $(\widetilde{x}, \widetilde{y})$. With similar arguments as in Step 1a let us first derive analogous statements as in (39) for $\widetilde{\mathcal{M}}^X$ and $\widetilde{\mathcal{M}}^Y$ as defined in (13a) as well as $\mathcal{M}^X_*$ and $\mathcal{M}^Y_*$ as defined in (13c), respectively. To do so, we first notice that $(\mathbb{E}\overline{X}_t, \mathbb{E}\overline{Y}_t)$ is continuous and since it converges to $(\widetilde{x}, \widetilde{y})$ as $t \to \infty$, there exists $M > 0$, potentially depending on $(\widetilde{x}, \widetilde{y})$, such that $\left\|\mathbb{E}\overline{X}_t\right\|_2 + \left\|\mathbb{E}\overline{Y}_t\right\|_2 \leq M$ for all $t \geq 0$. Since moreover $\underline{\mathcal{E}}$ and $\overline{\mathcal{E}}$ are continuous, there exists $\mathcal{E}_M > 0$ such that $-\mathcal{E}_M \leq \underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t) \leq \overline{\mathcal{E}}(\mathbb{E}\overline{X}_t) \leq \mathcal{E}_M$ for all $t > 0$. Utilizing this together with (37) and (39) we derive from Lemma 14 for $\widetilde{\mathcal{M}}^X$ and $\widetilde{\mathcal{M}}^Y$ from (13a) that

$$\begin{aligned}
\frac{d}{dt}\widetilde{\mathcal{M}}^X(t) &\geq -8\alpha e^{\alpha\mathcal{E}_M} C_{\nabla^2\mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2}\right) \frac{\mathrm{Var}^X(0) e^{-\mu_1 t}}{\mathcal{M}^X(0)} - \sqrt{2}\alpha\lambda_2 e^{\alpha\mathcal{E}_M} C_{\nabla\mathcal{E}} \frac{\sqrt{\mathrm{Var}^Y(0)} e^{-\mu_2 t/2}}{\sqrt{\mathcal{M}^Y(0)}}, \\
\frac{d}{dt}\widetilde{\mathcal{M}}^Y(t) &\geq -8\beta e^{\beta\mathcal{E}_M} C_{\nabla^2\mathcal{E}} \left(\lambda_2 + \frac{\sigma_2^2}{2}\right) \frac{\mathrm{Var}^Y(0) e^{-\mu_2 t}}{\mathcal{M}^Y(0)} - \sqrt{2}\beta\lambda_1 e^{\beta\mathcal{E}_M} C_{\nabla\mathcal{E}} \frac{\sqrt{\mathrm{Var}^X(0)} e^{-\mu_1 t/2}}{\sqrt{\mathcal{M}^X(0)}}.
\end{aligned} \tag{40}$$

Analogously, by using (37) and (39) it follows directly from Lemma 16 for $\mathcal{M}^X_*$ and $\mathcal{M}^Y_*$ from (13c) that

$$\begin{aligned}
\frac{d}{dt}\mathcal{M}^X_*(t) &\geq -8\alpha\lambda_1 e^{-\alpha\underline{\mathcal{E}}(y^*)} C_{\nabla\mathcal{E}} \frac{\sqrt{\mathrm{Var}^X(0)} e^{-\mu_1 t/2}}{\sqrt{\mathcal{M}^X(0)}} - 4\alpha\sigma_1^2 e^{-\alpha\underline{\mathcal{E}}(y^*)} C_{\nabla^2\mathcal{E}} \frac{\mathrm{Var}^X(0) e^{-\mu_1 t}}{\mathcal{M}^X(0)}, \\
\frac{d}{dt}\mathcal{M}^Y_*(t) &\geq -8\beta\lambda_2 e^{\beta\overline{\mathcal{E}}(x^*)} C_{\nabla\mathcal{E}} \frac{\sqrt{\mathrm{Var}^Y(0)} e^{-\mu_2 t/2}}{\sqrt{\mathcal{M}^Y(0)}} - 4\beta\sigma_2^2 e^{\beta\overline{\mathcal{E}}(x^*)} C_{\nabla^2\mathcal{E}} \frac{\mathrm{Var}^Y(0) e^{-\mu_2 t}}{\mathcal{M}^Y(0)}.
\end{aligned} \tag{41}$$

Integrating (40) and employing the well-preparedness condition P3 shows for all $t \geq 0$ that

$$\begin{aligned}
\widetilde{\mathcal{M}}^X(t) &\geq \widetilde{\mathcal{M}}^X(0) - 8\alpha e^{\alpha\mathcal{E}_M} C_{\nabla^2\mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2}\right) \frac{\mathrm{Var}^X(0)}{\mu_1 \mathcal{M}^X(0)} - 2\sqrt{2}\alpha\lambda_2 e^{\alpha\mathcal{E}_M} C_{\nabla\mathcal{E}} \frac{\sqrt{\mathrm{Var}^Y(0)}}{\mu_2\sqrt{\mathcal{M}^Y(0)}} \geq \frac{3}{4}\widetilde{\mathcal{M}}^X(0), \\
\widetilde{\mathcal{M}}^Y(t) &\geq \widetilde{\mathcal{M}}^Y(0) - 8\beta e^{\beta\mathcal{E}_M} C_{\nabla^2\mathcal{E}} \left(\lambda_2 + \frac{\sigma_2^2}{2}\right) \frac{\mathrm{Var}^Y(0)}{\mu_2 \mathcal{M}^Y(0)} - 2\sqrt{2}\beta\lambda_1 e^{\beta\mathcal{E}_M} C_{\nabla\mathcal{E}} \frac{\sqrt{\mathrm{Var}^X(0)}}{\mu_1\sqrt{\mathcal{M}^X(0)}} \geq \frac{3}{4}\widetilde{\mathcal{M}}^Y(0).
\end{aligned}$$

Analogously, using (41) together with P3 shows for all $t \geq 0$ that

$$\begin{aligned}
\mathcal{M}^X_*(t) &\geq \mathcal{M}^X_*(0) - 16\alpha\lambda_1 e^{-\alpha\underline{\mathcal{E}}(y^*)} C_{\nabla\mathcal{E}} \frac{\sqrt{\mathrm{Var}^X(0)}}{\mu_1\sqrt{\mathcal{M}^X(0)}} - 4\alpha\sigma_1^2 e^{-\alpha\underline{\mathcal{E}}(y^*)} C_{\nabla^2\mathcal{E}} \frac{\mathrm{Var}^X(0)}{\mu_1\mathcal{M}^X(0)} \geq \frac{3}{4}\mathcal{M}^X_*(0), \\
\mathcal{M}^Y_*(t) &\geq \mathcal{M}^Y_*(0) - 16\beta\lambda_2 e^{\beta\overline{\mathcal{E}}(x^*)} C_{\nabla\mathcal{E}} \frac{\sqrt{\mathrm{Var}^Y(0)}}{\mu_2\sqrt{\mathcal{M}^Y(0)}} - 4\beta\sigma_2^2 e^{\beta\overline{\mathcal{E}}(x^*)} C_{\nabla^2\mathcal{E}} \frac{\mathrm{Var}^Y(0)}{\mu_2\mathcal{M}^Y(0)} \geq \frac{3}{4}\mathcal{M}^Y_*(0).
\end{aligned}$$

Thus, in particular it holds for all $t \geq 0$

$$\widetilde{\mathcal{M}}^X(t) \geq \frac{1}{2}\widetilde{\mathcal{M}}^X(0) \quad \text{and} \quad \widetilde{\mathcal{M}}^Y(t) \geq \frac{1}{2}\widetilde{\mathcal{M}}^Y(0) \tag{42}$$

as well as

$$\mathcal{M}^X_*(t) \geq \frac{1}{2}\mathcal{M}^X_*(0) \quad \text{and} \quad \mathcal{M}^Y_*(t) \geq \frac{1}{2}\mathcal{M}^Y_*(0). \tag{43}$$

Step 2b: By Chebyshev's inequality, for each $\delta > 0$ it holds that

$$\begin{aligned}
\rho_t(\{\|(x - \widetilde{x}, y - \widetilde{y})\|_2 \geq \delta\}) &\leq \frac{2}{\delta^2}\left(\mathrm{Var}^X(t) + \mathrm{Var}^Y(t) + \left\|\mathbb{E}\overline{X}_t - \widetilde{x}\right\|_2^2 + \left\|\mathbb{E}\overline{Y}_t - \widetilde{y}\right\|_2^2\right) \\
&\to 0, \text{ as } t \to \infty.
\end{aligned}$$

Thus, the pair $(\overline{X}_t, \overline{Y}_t)$ converges to $(\widetilde{x}, \widetilde{y})$ in probability as t tends to infinity. Recall the convergence $\left(\mathbb{E}\overline{X}_t, \mathbb{E}\overline{Y}_t\right) \to (\widetilde{x}, \widetilde{y})$, the continuity of $\mathcal{E}$, and the fact that for all $t \geq 0$,

$$\exp(-\alpha\mathcal{E}(\overline{X}_t, \mathbb{E}\overline{Y}_t)) \leq \exp(\alpha\mathcal{E}_M), \quad \text{a.s.}$$

By the dominated convergence theorem, one can pass to the limit in t to obtain $\lim_{t\to\infty} \widetilde{\mathcal{M}}^X(t) = \exp\left(-\alpha\mathcal{E}(\widetilde{x}, \widetilde{y})\right)$. Analogously, one may get $\widetilde{\mathcal{M}}^Y(t) \to \exp\left(\beta\mathcal{E}(\widetilde{x}, \widetilde{y})\right)$ as $t \to \infty$. Using this when taking the limit $t \to \infty$ in the bounds (42) after applying the logarithm and multiplying both sides with $-1/\alpha$ and $1/\beta$, respectively, we obtain

$$\begin{aligned}
\mathcal{E}(\widetilde{x}, \widetilde{y}) &= \lim_{t\to\infty} \left(-\frac{1}{\alpha} \log \widetilde{\mathcal{M}}^X(t)\right) \leq \frac{1}{\alpha}\log 2 - \frac{1}{\alpha}\log \widetilde{\mathcal{M}}^X(0), \\
\mathcal{E}(\widetilde{x}, \widetilde{y}) &= \lim_{t\to\infty} \left(\frac{1}{\beta} \log \widetilde{\mathcal{M}}^Y(t)\right) \geq -\frac{1}{\beta}\log 2 + \frac{1}{\beta}\log \widetilde{\mathcal{M}}^Y(0).
\end{aligned} \tag{44}$$

Due to the first set of well-preparedness conditions from P2, the Laplace principle in form of Lemmas A.3 and A.4 when choosing μ^α as the law of the initial data $\overline{X}_0$ and μ^β as the law of $\overline{Y}_0$, now allows to choose $\alpha \geq (2\log 2)/\varepsilon$ and $\beta \geq (2\log 2)/\varepsilon$ large enough such that for given $\varepsilon > 0$ it moreover holds

$$\begin{aligned}
-\frac{1}{\alpha}\log \widetilde{\mathcal{M}}^X(0) - \min_{x\in\mathbb{R}^{d_1}} \mathcal{E}(x, \mathbb{E}\overline{Y}_0) &= -\frac{1}{\alpha}\log \mathbb{E}\exp\left(-\alpha\mathcal{E}(\overline{X}_0, \mathbb{E}\overline{Y}_0)\right) - \min_{x\in\mathbb{R}^{d_1}} \mathcal{E}(x, \mathbb{E}\overline{Y}_0) \leq \varepsilon/2, \\
-\frac{1}{\beta}\log \widetilde{\mathcal{M}}^Y(0) + \max_{y\in\mathbb{R}^{d_2}} \mathcal{E}(\mathbb{E}\overline{X}_0, y) &= -\frac{1}{\beta}\log \mathbb{E}\exp\left(\beta\mathcal{E}(\mathbb{E}\overline{X}_0, \overline{Y}_0)\right) + \max_{y\in\mathbb{R}^{d_2}} \mathcal{E}(\mathbb{E}\overline{X}_0, y) \leq \varepsilon/2.
\end{aligned} \tag{45}$$

Notice here that we well-prepare α and β simultaneously with the initial data $(\overline{X}_0, \overline{Y}_0)$ (therewith $(\overline{X}_0, \overline{Y}_0)$ depends on α, β). However due to the well-preparedness conditions P2, α and β can still be taken sufficiently large as ensured in Lemmas A.3 and A.4.

Such choices of parameters in Equation (44) immediately give

$$\mathcal{E}(\widetilde{x}, \widetilde{y}) \leq \min_{x\in\mathbb{R}^{d_1}} \mathcal{E}(x, \mathbb{E}\overline{Y}_0) + \varepsilon \quad \text{and} \quad \mathcal{E}(\widetilde{x}, \widetilde{y}) \geq \max_{y\in\mathbb{R}^{d_2}} \mathcal{E}(\mathbb{E}\overline{X}_0, y) - \varepsilon \tag{46}$$

and consequently

$$\mathcal{E}(\widetilde{x}, \widetilde{y}) \leq \min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} \mathcal{E}(x, y) + \varepsilon \quad \text{and} \quad \mathcal{E}(\widetilde{x}, \widetilde{y}) \geq \max_{y\in\mathbb{R}^{d_2}} \min_{x\in\mathbb{R}^{d_1}} \mathcal{E}(x, y) - \varepsilon, \tag{47}$$

which proves the first part of (16). Secondly, following an analogous argumentation for $\mathcal{M}_*^X$ and $\mathcal{M}_*^Y$ as defined in (13c), we obtain the remainder of (16). More precisely, we first note that, as $t \to \infty$,

$$\mathcal{M}_*^X(t) \to \exp\left(-\alpha\mathcal{E}(\widetilde{x}, y^*)\right) \quad \text{and} \quad \mathcal{M}_*^Y(t) \to \exp\left(\beta\mathcal{E}(x^*, \widetilde{y})\right). \tag{48}$$

Taking now the limit $t \to \infty$ in (43) after suitable algebraic manipulations, we obtain

$$\begin{aligned}
\mathcal{E}(\widetilde{x}, y^*) &= \lim_{t\to\infty} \left(-\frac{1}{\alpha} \log \mathcal{M}_*^X(t)\right) \leq \frac{1}{\alpha}\log 2 - \frac{1}{\alpha}\log \mathcal{M}_*^X(0), \\
\mathcal{E}(x^*, \widetilde{y}) &= \lim_{t\to\infty} \left(\frac{1}{\beta} \log \mathcal{M}_*^Y(t)\right) \geq -\frac{1}{\beta}\log 2 + \frac{1}{\beta}\log \mathcal{M}_*^Y(0).
\end{aligned} \tag{49}$$

A potentially larger choice of α and β allows (again by the Laplace principle in form of Lemmas A.3 and A.4, which applies due to the second set of well-preparedness conditions from P2) to guarantee

$$\begin{aligned}
-\frac{1}{\alpha}\log \mathcal{M}_*^X(0) - \min_{x\in\mathbb{R}^{d_1}} \mathcal{E}(x, y^*) &= -\frac{1}{\alpha}\log \mathbb{E}\exp\left(-\alpha\mathcal{E}(\overline{X}_0, y^*)\right) - \min_{x\in\mathbb{R}^{d_1}} \mathcal{E}(x, y^*) \leq \varepsilon/2, \\
-\frac{1}{\beta}\log \mathcal{M}_*^Y(0) + \max_{y\in\mathbb{R}^{d_2}} \mathcal{E}(x^*, y) &= -\frac{1}{\beta}\log \mathbb{E}\exp\left(\beta\mathcal{E}(x^*, \overline{Y}_0)\right) + \max_{y\in\mathbb{R}^{d_2}} \mathcal{E}(x^*, y) \leq \varepsilon/2
\end{aligned}$$

for the specified ε. Such choices of parameters in Equation (49) immediately give

$$\mathcal{E}(\widetilde{x}, y^*) \leq \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y^*) + \varepsilon \quad \text{and} \quad \mathcal{E}(x^*, \widetilde{y}) \geq \max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(x^*, y) - \varepsilon\,, \tag{50}$$

which completes the proof of (16).

Step 3: Finally, under the inverse continuity property A3 and making use of what we just proved, we additionally obtain (17), which concludes the proof. □

5 Implementation of CBO-SP and Numerical Experiments

5.1 Numerical Algorithm and Implementation

In order to implement and run CBO-SP on a computer, we first fix a discrete time step size Δt as well as a number of iterations K or define any other suitable stopping criterion. Then, by discretizing the interacting particle system (1) via an Euler-Maruyama time discretization [19, 41] as

$$\widehat{X}^i_{k+1} = \widehat{X}^i_k - \lambda_1 \Delta t \left(\widehat{X}^i_k - x^Y_\alpha(\widehat{\rho}^{N_1}_{X,k})\right) + \sigma_1 D\left(\widehat{X}^i_k - x^Y_\alpha(\widehat{\rho}^{N_1}_{X,k})\right) B^{X,i}_k, \tag{51a}$$

$$\widehat{Y}^i_{k+1} = \widehat{Y}^i_k - \lambda_2 \Delta t \left(\widehat{Y}^i_k - y^X_\beta(\widehat{\rho}^{N_2}_{Y,k})\right) + \sigma_2 D\left(\widehat{Y}^i_k - y^X_\beta(\widehat{\rho}^{N_2}_{Y,k})\right) B^{Y,i}_k, \tag{51b}$$

where $\widehat{\rho}^{N_1}_{X,k}$ and $\widehat{\rho}^{N_2}_{Y,k}$ denote the empirical averages of the iterates $(\widehat{X}^i_k)_{i=1,\dots,N_1}$ and $(\widehat{Y}^i_k)_{i=1,\dots,N_2}$ and where

$$\widehat{x}^Y_\alpha(\widehat{\rho}^{N_1}_{X,k}) = \int x \, \frac{\omega_\alpha\big(x, \int y \, d\widehat{\rho}^{N_2}_{Y,k}(y)\big)}{\big\|\omega_\alpha\big(\,\cdot\,, \int y \, d\widehat{\rho}^{N_2}_{Y,k}(y)\big)\big\|_{L_1(\widehat{\rho}^{N_1}_{X,k})}} \, d\widehat{\rho}^{N_1}_{X,k}(x), \tag{52a}$$

$$\widehat{y}^X_\beta(\widehat{\rho}^{N_2}_{Y,k}) = \int y \, \frac{\omega_{-\beta}\big(\int x \, d\widehat{\rho}^{N_1}_{X,k+1}(x), y\big)}{\big\|\omega_{-\beta}\big(\int x \, d\widehat{\rho}^{N_1}_{X,k+1}(x), \,\cdot\,\big)\big\|_{L_1(\widehat{\rho}^{N_2}_{Y,k})}} \, d\widehat{\rho}^{N_2}_{Y,k}(y), \tag{52b}$$

we obtain the implementable iterative scheme, which is used in the formulation of Algorithm 1. Moreover, $\big((B^{X,i}_k)_{k=1,\dots,K}\big)_{i=1,\dots,N_1}$ and $\big((B^{Y,i}_k)_{k=1,\dots,K}\big)_{i=1,\dots,N_2}$ in (51) are independent Gaussian vectors in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively, with covariance matrix $\Delta t\mathrm{Id}$. Note that in Equation (52b) we could also use the old iterates $\widehat{\rho}^{N_1}_{X,k}$ instead of the new ones $\widehat{\rho}^{N_1}_{X,k+1}$ for the computation.

Algorithm 1 CBO-SP

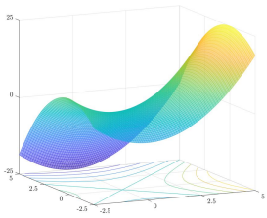
Input: Objective $\mathcal{E}$, discrete time step size Δt, number of iterates K, parameters $\lambda_1, \lambda_2, \sigma_1, \sigma_2, \alpha, \beta$, number of particles N_1 and N_2, initialization ρ_0

Output: Approximation $\big(\widehat{x}^Y_\alpha(\widehat{\rho}^{N_1}_{X,k}), \widehat{y}^X_\beta(\widehat{\rho}^{N_2}_{Y,k})\big)$ of the saddle point (x^*, y^*) of $\mathcal{E}$

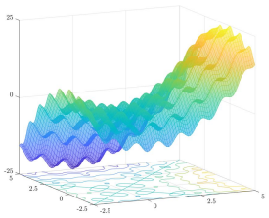
Generate the particles' initial positions $(X^i_0)_{i=1,\dots,N_1}$ and $(Y^i_0)_{i=1,\dots,N_2}$ according to the initial laws $\rho_{X,0}$ and $\rho_{Y,0}$, respectively. Set $k = 0$.
while $k \leq K$ or stopping criterion not fulfilled
    Compute the component $\widehat{x}^Y_\alpha(\widehat{\rho}^{N_1}_{X,k})$ of the consensus point according to (52a).
    Update the X-positions by computing $\big(\widehat{X}^i_{k+1}\big)_{i=1,\dots,N_1}$ according to (51a).
    Compute the component $\widehat{y}^X_\beta(\widehat{\rho}^{N_2}_{Y,k})$ of the consensus point according to (52b).
    Update the Y-positions by computing $\big(\widehat{Y}^i_{k+1}\big)_{i=1,\dots,N_2}$ according to (51b).
    Check the stopping criterion and **break** if fulfilled. If not, continue and set $k = k + 1$.
end while
Compute consensus point $\big(\widehat{x}^Y_\alpha(\widehat{\rho}^{N_1}_{X,k}), \widehat{y}^X_\beta(\widehat{\rho}^{N_2}_{Y,k})\big)$ as final approximation to saddle point (x^*, y^*).

5.2 Illustrative Numerical Experiments for CBO-SP

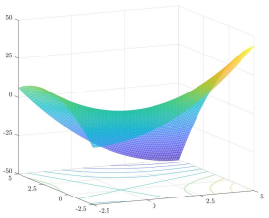
To visualize the behavior of the CBO-SP algorithm in practice, we depict in Figure 1 below snapshots of the positions of the particles for four different types of saddle point functions, which are plotted in the first row of the figure. The experiments include two *nonconvex-nonconcave* examples, which is in general the setting of particular interest in modern applications. We observe



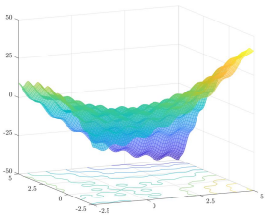
(a) The objective function $\mathcal{E}(x,y) = x^2 - y^2$



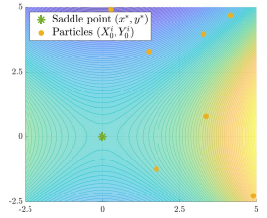
(b) The objective function $\mathcal{E}(x,y) = R(x) - R(y)$



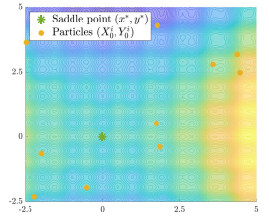
(c) The objective function $\mathcal{E}(x,y) = x^2 - 2xy - y^2$



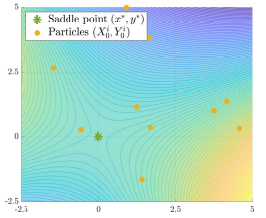
(d) The objective function $\mathcal{E}(x,y) = R(x) - 2xy - R(y)$



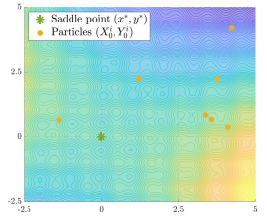
(a0) Initial configuration of the particles for **(a)**



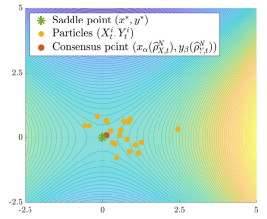
(b0) Initial configuration of the particles for **(b)**



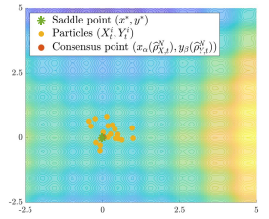
(c0) Initial configuration of the particles for **(c)**



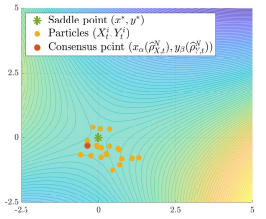
(d0) Initial configuration of the particles for **(d)**



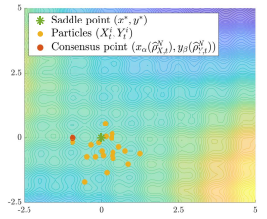
(a1) Positions of the particles at $t = 2$ for **(a)**



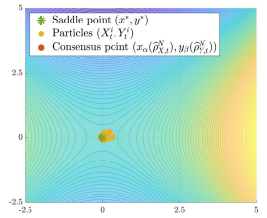
(b1) Positions of the particles at $t = 2$ for **(b)**



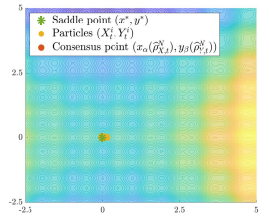
(c1) Positions of the particles at $t = 2$ for **(c)**



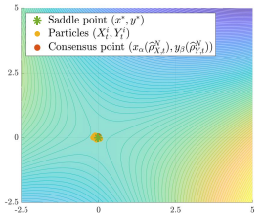
(d1) Positions of the particles at $t = 2$ for **(d)**



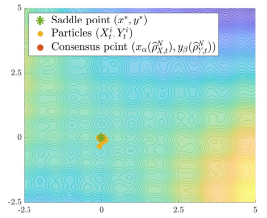
(a2) Final configuration of the particles for **(a)**



(b2) Final configuration of the particles for **(b)**



(c2) Final configuration of the particles for **(c)**



(d2) Final configuration of the particles for **(d)**

Figure 1: Illustration of the dynamics of CBO-SP when searching the global Nash equilibrium of four different saddle point functions plotted in **(a)**–**(d)** in $d = 1$, where $R(x) = \sum_{k=1}^{d} x_k^2 + \frac{5}{2}\big(1 - \cos(2\pi x_k)\big)$ is the Rastrigin function. Each column visualizes the positions of the $N = 20$ particles when running CBO-SP with parameters $\alpha = \beta = 10^{15}$, $\lambda_1 = \lambda_2 = 1$, $\sigma_1 = \sigma_2 = \sqrt{0.1}$ and time step size $\Delta t = 0.1$ at three different points in time ($t = 0$, $t = 2$ and $t = T = 4$). The particles are sampled initially from $\rho_0 \sim \mathcal{N}(2,4) \times \mathcal{N}(2,4)$.

that in all cases (also in case of different initializations) the saddle point is found fast and reliably.

5.3 Solving a Quadratic Game with CBO-SP

To demonstrate the practicability of CBO-SP, we solve a strongly-monotone quadratic game [27, Section 5] of the form

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}x^T A_i x + x^T B_i y - \frac{1}{2}x^T C_i y \tag{53}$$

with sample size $n = 100$ and for various dimensions d_1 and d_2. The matrices $B_i \in \mathbb{R}^{d_1\times d_2}$ have random Gaussian entries and the positive definite matrices $A_i \in \mathbb{R}^{d_1\times d_1}$ and $C_i \in \mathbb{R}^{d_2\times d_2}$ are of the form $A_i = \widetilde{A}_i^T\widetilde{A}_i$ and $C_i = \widetilde{C}_i^T\widetilde{C}_i$ with $\widetilde{A}_i \in \mathbb{R}^{d_1\times d_1}$ and $\widetilde{C}_i \in \mathbb{R}^{d_2\times d_2}$ having random Gaussian entries. We employ CBO-SP with parameters $\alpha = \beta = 10^{15}$, $\lambda_1 = \lambda_2 = 1$, $\sigma_1 = \sigma_2 = 2$ using $N \in \{40, 80, 120, 200\}$ particles, and with time horizon $T = 100$ and discrete time step size $\Delta t = 0.1$. The particles are sampled initially from $\rho_0 \sim \mathcal{N}(4, 2\mathsf{Id}) \times \mathcal{N}(4, 2\mathsf{Id})$ (i.e., they are initialized substantially far from the saddle point). We depict in Table 1 below the success rates, average ℓ^∞-error and average run time of CBO-SP algorithm computed on the basis of 100 runs. A run is considered successful if the obtained solution has an accuracy of 10^{-3} w.r.t. the ℓ^∞-norm. In brackets, we indicate the average (over the runs) runtime in milliseconds (ms) as well as the average (over the runs) ℓ^∞-error.

Table 1: Success rates, average runtime (in ms) and average ℓ^∞-error of the CBO-SP algorithm when solving a quadratic game as specified in (53) for different dimensions d_1 and d_2, and with different numbers of particles N. All results are computed on the basis of 100 runs of the algorithm.

	$N = 40$	$N = 80$	$N = 120$	$N = 200$
$d_1 = 20,\ d_2 = 8$	31% (32ms, $1.5 \cdot 10^{-2}$)	100% (50ms, $2.4 \cdot 10^{-7}$)	100% (182ms, $6.1 \cdot 10^{-8}$)	100% (229ms, $2.9 \cdot 10^{-8}$)
$d_1 = 20,\ d_2 = 20$	7% (41ms, $3.0 \cdot 10^{-2}$)	100% (63ms, $4.5 \cdot 10^{-7}$)	100% (339ms, $3.7 \cdot 10^{-8}$)	100% (462ms, $2.4 \cdot 10^{-8}$)
$d_1 = 40,\ d_2 = 8$	0% (154ms, 1.1)	1% (190ms, $3.6 \cdot 10^{-2}$)	53% (226ms, $2.6 \cdot 10^{-3}$)	100% (418ms, $4.8 \cdot 10^{-5}$)
$d_1 = 40,\ d_2 = 20$	0% (162ms, 1.2)	0% (285ms, $4.9 \cdot 10^{-2}$)	52% (436ms, $3.8 \cdot 10^{-3}$)	100% (539ms, $8.2 \cdot 10^{-5}$)
$d_1 = 40,\ d_2 = 40$	0% (330ms, 1.9)	0% (336ms, $1.2 \cdot 10^{-1}$)	25% (421ms, $5.4 \cdot 10^{-3}$)	100% (606ms, $7.9 \cdot 10^{-5}$)

We observe that with already moderately many particles, the CBO-SP algorithm is capable of consistently finding the desired saddle point for relatively high-dimensional minimax problems.

Experiments in much higher dimensions and more applied settings coming for instance from economics or arising when training GANs are left to future and more experimental research, which focuses on benchmarking rather than providing rigorous convergence guarantees.

6 Conclusions

In this paper we propose consensus-based optimization for saddle point problems (CBO-SP) and analyze its global convergence behavior to global Nash equilibria. As apparent from the proof, our technique requires the equilibrium to satisfy the saddle point property, i.e., that $\min_x \max_y$ and $\max_y \min_x$ coincide. We leave to further research the extension of the results to sequential games, where the latter condition does not hold. This is in particular relevant in, for instance, the training of GANs, which are formulated as non-simultaneous games.

Acknowledgements

J.Q. is partially supported by the National Science and Engineering Research Council of Canada (NSERC) and by the PIMS-Europe Fellowship funding. K.R. acknowledges financial supports from the Technical University of Munich-Institute for Ethics in Artificial Intelligence (IEAI).

The authors moreover sincerely thank the two referees for their careful and insightful comments and suggestions which helped improve the article.

References

[1] L. Arnold. *Stochastic differential equations: Theory and applications*. Wiley-Interscience [John Wiley & Sons], New York-London-Sydney, 1974. Translated from the German.

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

[3] G. Borghi, M. Herty, and L. Pareschi. Constrained consensus-based optimization. *SIAM Journal on Optimization*, 33(1):211–236, 2023.

[4] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[5] J. A. Carrillo, Y.-P. Choi, C. Totzeck, and O. Tse. An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.*, 28(6):1037–1066, 2018.

[6] J. A. Carrillo, S. Jin, L. Li, and Y. Zhu. A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM Control Optim. Calc. Var.*, 27(suppl.):Paper No. S5, 22, 2021.

[7] J. A. Carrillo, C. Totzeck, and U. Vaes. Consensus-based optimization and ensemble kalman inversion for global optimization problems with constraints. In *Modeling and Simulation for Collective Dynamics*, pages 195–230. World Scientific, 2023.

[8] T.-H. Chang, M. Hong, H.-T. Wai, X. Zhang, and S. Lu. Distributed learning in the nonconvex world: From batch data to streaming and beyond. *IEEE Signal Processing Magazine*, 37(3):26–38, 2020.

[9] C. Daskalakis and I. Panageas. The limit points of (optimistic) gradient descent in min-max optimization. *Advances in neural information processing systems*, 31, 2018.

[10] R. Durrett. *Stochastic calculus*. Probability and Stochastics Series. CRC Press, Boca Raton, FL, 1996. A practical introduction.

[11] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. Consensus-based optimization on hypersurfaces: Well-posedness and mean-field limit. *Math. Models Methods Appl. Sci.*, 30(14):2725–2751, 2020.

[12] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. Consensus-based optimization on the sphere: Convergence to global minimizers and machine learning. *Journal of Machine Learning Research*, 22(237):1–55, 2021.

[13] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. Anisotropic diffusion in consensus-based optimization on the sphere. *SIAM Journal on Optimization*, 32(3):1984–2012, 2022.

[14] M. Fornasier, T. Klock, and K. Riedl. Consensus-based optimization methods converge globally. *arXiv:2103.15130*, 2021.

[15] M. Fornasier, T. Klock, and K. Riedl. Convergence of anisotropic consensus-based optimization in mean-field law. In J. L. Jiménez Laredo, J. I. Hidalgo, and K. O. Babaagba, editors, *Applications of Evolutionary Computation*, pages 738–754, Cham, 2022. Springer.

[16] R. H. Gohary, Y. Huang, Z.-Q. Luo, and J.-S. Pang. A generalized iterative water-filling algorithm for distributed power control in the presence of a jammer. *IEEE Transactions on Signal Processing*, 57(7):2660–2674, 2009.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[18] E. Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[19] D. J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Rev.*, 43(3):525–546, 2001.

[20] H. Huang and J. Qiu. On the mean-field limit for the consensus-based optimization. *Mathematical Methods in the Applied Sciences*, pages 1–18, 2022.

[21] H. Huang, J. Qiu, and K. Riedl. On the global convergence of particle swarm optimization methods. *Applied Mathematics & Optimization*, 88(2):30, 2023.

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

[23] R. A. Krohling, F. Hoffmann, and L. S. Coelho. Co-evolutionary particle swarm optimization for min-max problems using gaussian distribution. In *Proceedings of the 2004 congress on evolutionary computation (IEEE cat. no. 04TH8753)*, volume 1, pages 959–964. IEEE, 2004.

[24] E. C. Laskari, K. E. Parsopoulos, and M. N. Vrahatis. Particle swarm optimization for minimax problems. In *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*, volume 2, pages 1576–1581. IEEE, 2002.

[25] M. Liu, H. Rafique, Q. Lin, and T. Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *J. Mach. Learn. Res.*, 22:169–1, 2021.

[26] Y.-F. Liu, Y.-H. Dai, and Z.-Q. Luo. Max-min fairness linear transceiver design for a multi-user mimo interference channel. *IEEE Transactions on Signal Processing*, 61(9):2413–2423, 2013.

[27] N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, and S. Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34:19095–19108, 2021.

[28] N. Loizou, H. Berard, A. Jolicoeur-Martineau, P. Vincent, S. Lacoste-Julien, and I. Mitliagkas. Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pages 6370–6381. PMLR, 2020.

[29] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[30] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[31] E. V. Mazumdar, M. I. Jordan, and S. S. Sastry. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *arXiv:1901.00838*, 2019.

[32] P. D. Miller. *Applied asymptotic analysis*, volume 75 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2006.

[33] K. G. Murty and S. N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming: Series A and B*, 39(2):117–129, 1987.

[34] R. B. Myerson. *Game theory: analysis of conflict*. Harvard university press, 1997.

[35] J. F. Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[36] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[37] B. Øksendal. *Stochastic differential equations: An introduction with applications*. Springer-Verlag, Berlin, sixth edition, 2003.

[38] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pages 2681–2690. PMLR, 2017.

[39] J.-S. Pang and G. Scutari. Nonconvex games with side constraints. *SIAM Journal on Optimization*, 21(4):1491–1522, 2011.

[40] R. Pinnau, C. Totzeck, O. Tse, and S. Martin. A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.*, 27(1):183–204, 2017.

[41] E. Platen. An introduction to numerical methods for stochastic differential equations. In *Acta numerica, 1999*, volume 8 of *Acta Numer.*, pages 197–246. Cambridge Univ. Press, Cambridge, 1999.

[42] M. Razaviyayn, T. Huang, S. Lu, M. Nouiehed, M. Sanjabi, and M. Hong. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66, 2020.

[43] K. Riedl. Leveraging memory effects and gradient information in consensus-based optimization: On global convergence in mean-field law. *arXiv:2211.12184*, 2022.

[44] Y. Shi and R. A. Krohling. Co-evolutionary particle swarm optimization to solve min-max problems. In *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*, volume 2, pages 1682–1687. IEEE, 2002.

[45] M. Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

[46] J. Von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

[47] J. Von Neumann and O. Morgenstern. Theory of games and economic behavior. In *Theory of games and economic behavior*. Princeton university press, 2007.

A Appendix

A.1 Existence and Uniqueness of Solutions to SDEs

For the sake of a self-consistent presentation, let us recall two results from [10] about the existence and pathwise uniqueness of a strong solution of a SDE of the form

$$Z_t = Z_0 + \int_0^t b(Z_s)\,ds + \int_0^t \sigma(Z_s)\,dB_s. \tag{$\star$}$$

These results are used in the proof of Theorem 3. Note that here we adopted the notation of [10], i.e., in our setting we have $Z_t = \mathbf{Z}_t$ as well as $b(\mathbf{Z}_t) = -\boldsymbol{\lambda}\mathbf{F}(\mathbf{Z}_t)$ and $\sigma(\mathbf{Z}_t) = \boldsymbol{\sigma}\mathbf{M}(\mathbf{Z}_t)$.

Theorem A.1 ([10, Chapter 5, Theorem 3.1]). *Suppose that*

(i) for any $n < \infty$ we have $|b_i(z) - b_i(z')| \le K_n |z - z'|$ and $|\sigma_{ij}(z) - \sigma_{ij}(z')| \le K_n |z - z'|$ for $|z|, |z'| \le n$,

(ii) there is a constant $A < \infty$ and a function $\varphi(z) \ge 0$ so that if Z_t is a solution of $(\star)$, then $e^{-At}\varphi(Z_t)$ is a local supermartingale.

Then $(\star)$ has a strong solution and pathwise uniqueness holds.

Theorem A.2 ([10, Chapter 5, Theorem 3.2]). *Let $a = \sigma\sigma^T$ and suppose that $\sum_{i=1}^{d} 2z_i b_i(z) + a_{ii}(z) \le B(1 + |z|^2)$. Then (ii) in Theorem A.1 holds with $A = B$ and $\varphi(z) = 1 + |z|^2$.*

A.2 The Laplace Principle

Lemma A.3. *For any fixed $y \in \mathbb{R}^{d_2}$ and any $\delta > 0$, define the set*

$$S_{y,\delta} = \left\{ x \in \mathbb{R}^{d_1} : \exp\left(-\mathcal{E}(x,y)\right) > \exp\left(- \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x,y) \right) - \delta \right\}.$$

Let $\{\mu^\alpha\}_{\alpha \ge 1}$ be a family of measures in $\mathcal{P}(\mathbb{R}^{d_1})$ and assume that there exists some constant $C_\delta > 0$ depending only on δ such that $\mu^\alpha(S_{y,\delta}) \ge C_\delta$ for all $\alpha \ge 1$. Then it holds

$$\lim_{\alpha \to \infty} -\frac{1}{\alpha} \log \left(\int_{\mathbb{R}^{d_1}} \exp\left(-\alpha \mathcal{E}(x,y)\right) d\mu^\alpha(x) \right) = \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x,y).$$

Proof. We first notice that by definition of the set $S_{y,\delta}$ it holds

$$\begin{aligned}
\left(\int_{\mathbb{R}^{d_1}} \exp\left(-\alpha \mathcal{E}(x,y)\right) d\mu^\alpha(x) \right)^{\frac{1}{\alpha}} &\ge \left(\int_{S_{y,\delta}} \left(\exp\left(- \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x,y) \right) - \delta \right)^\alpha d\mu^\alpha(x) \right)^{\frac{1}{\alpha}} \\
&= \left(\exp\left(- \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x,y) \right) - \delta \right) \mu^\alpha(S_{y,\delta})^{\frac{1}{\alpha}} \\
&\ge \left(\exp\left(- \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x,y) \right) - \delta \right) C_\delta^{\frac{1}{\alpha}} \\
&\to \exp\left(- \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x,y) \right) - \delta \quad \text{as} \quad \alpha \to \infty.
\end{aligned}$$

Thus, for any $\delta > 0$, we have

$$\liminf_{\alpha \to \infty} \left(\int_{\mathbb{R}^{d_1}} \exp\left(-\alpha \mathcal{E}(x,y)\right) d\mu^\alpha(x) \right)^{\frac{1}{\alpha}} \ge \exp\left(- \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x,y) \right) - \delta.$$

On the other hand, clearly

$$\limsup_{\alpha \to \infty} \left(\int_{\mathbb{R}^{d_1}} \exp\left(-\alpha \mathcal{E}(x,y)\right) d\mu^\alpha(x) \right)^{\frac{1}{\alpha}} \le \exp\left(- \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x,y) \right).$$

Since δ was arbitrary, this implies

$$\lim_{\alpha \to \infty} \left(\int_{\mathbb{R}^{d_1}} \exp\left(-\alpha \mathcal{E}(x,y)\right) d\mu^\alpha(x) \right)^{\frac{1}{\alpha}} = \exp\left(- \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x,y) \right),$$

giving the result after taking the logarithm on both sides. □

Analogously we obtain the following.

Lemma A.4. *For any fixed $x \in \mathbb{R}^{d_1}$ and any $\delta > 0$, define the set*

$$S_{x,\delta} = \left\{ y \in \mathbb{R}^{d_2} : \exp\left(\mathcal{E}(x,y)\right) > \exp\left(\max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(x,y) \right) - \delta \right\}.$$

Let $\{\mu^\beta\}_{\beta \ge 1}$ be a family of measures in $\mathcal{P}(\mathbb{R}^{d_2})$ and assume that there exists some constant $C_\delta > 0$ depending only on δ such that $\mu^\beta(S_{x,\delta}) \ge C_\delta$ for all $\beta \ge 1$. Then it holds

$$\lim_{\beta \to \infty} \frac{1}{\beta} \log \left(\int_{\mathbb{R}^{d_2}} \exp\left(\beta \mathcal{E}(x,y)\right) d\mu^\beta(y) \right) = \max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(x,y).$$