Journal of Optimization Theory and Applications
https://doi.org/10.1007/s10957-023-02233-0

Polarization and Coherence in Mean Field Games Driven by Private and Social Utility

Paolo Dai Pra[1] · **Elena Sartori**[2] · **Marco Tolotti**[3]

Received: 28 June 2021 / Accepted: 25 April 2023

Abstract

We study a mean field game in continuous time over a finite horizon, T, where the state of each agent is binary and where players base their strategic decisions on two, possibly competing, factors: the willingness to align with the majority (conformism) and the aspiration of sticking with the own type (stubbornness). We also consider a quadratic cost related to the rate at which a change in the state happens: changing opinion may be a costly operation. Depending on the parameters of the model, the game may have more than one Nash equilibrium, even though the corresponding N-player game does not. Moreover, it exhibits a very rich phase diagram, where polarized/unpolarized, coherent/incoherent equilibria may coexist, except for T small, where the equilibrium is always unique. We fully describe such phase diagram in closed form and provide a detailed numerical analysis of the N-player counterpart of the mean field game. In this finite dimensional setting, the equilibrium selected by the population of players is always coherent (favoring the subpopulation whose type is aligned with the initial condition), but it does not necessarily minimize the cost functional. Rather, it seems that, among the coherent ones, the equilibrium prevailing is the one that most benefits the *underdog* subpopulation forced to change opinion.

Communicated by Xiaojun Chen.

✉ Elena Sartori
esartori@math.unipd.it

Paolo Dai Pra
paolo.daipra@univr.it

Marco Tolotti
tolotti@unive.it

[1] Department of Computer Science, University of Verona, 15, Strada Le Grazie, I-37134 Verona, Italy

[2] Department of Mathematics "Tullio Levi-Civita", University of Padova, 63, Via Trieste, I-35121 Padova, Italy

[3] Department of Management – Venice School of Management, Ca' Foscari University of Venice, Cannaregio 873, I-31121 Venice, Italy

Published online: 19 May 2023

Keywords Finite population dynamics · Mean field games · Multiple Nash equilibria · Phase transition · Social interaction

Mathematics Subject Classification 91A16 · 91B14

1 Introduction

In this paper, we analyze a simple continuous-time dynamic multi-agent model and study the limit as the number of agents goes to infinity. We consider a group of N interacting agents (*players*), who are allowed to control their *binary* state choosing the probability rate of "flipping" them. This rate is a feedback control that may depend on the state of all players and (measurably) on time. Each player aims at minimizing an individual cost, which is comprised by a *running cost* and a *final reward*. We consider a standard quadratic running cost. At some final time $T > 0$, each player gets a reward given as the sum of two different terms:

- the first mimics a *social driver* and favors imitation: the player gets a higher reward if she conforms with the majority, and if the majority becomes polarized (close to consensus); the majority is over the whole population, making the interaction between players of *mean field* type;
- the second models the private (individual) desire to align the state with the sign of a static and predetermined random variable denoting her personal type.

These two terms are possibly competing and represent a classical social dilemma: the former mimics *conformism*, i.e., the adherence to social norms and is often referred to as social utility; the latter models *stubbornness*, namely, the aspiration of the agent to stay as close as possible to the prescription of personal traits, hence mimicking a private (or individual) utility. As notion of optimality, we adopt that of *Nash equilibrium*, and our aim is to understand the system's behavior in the limit as $N \to +\infty$. This falls into the realm of *mean field games*, introduced by J.-M. Lasry and P.-L. Lions and, independently, by M. Huang, R.P. Malhamé and P.E. Caines (cf. [19, 20]), as limit models for symmetric many player dynamic games as the number of players tends to infinity; see, for instance, the lecture notes [6] and the two-volume work [7]. As concerns the finite state mean field games, we refer the reader to [14, 15].

The variable representing the type is introduced in the model as a random field and is treated as an observable static component of player' state. Therefore, this term introduces random disorder and, to the best of our knowledge, it is one of the first attempts to do it in mean field games.

In literature, this dilemma has been analyzed from different perspectives: [5] is usually considered as a pioneering study on the trade-off between private and social drivers in (static) binary choice models. In [4], a generalization to a dynamic continuous-time setting is proposed, which is not a mean field game, as agents play static games at random times. [11] proposes a model of consensus formation similar to our, where only the social component is present and individual preferences are not considered. [2] is one of the rare examples studying the interplay between stubbornness and imitation drivers in the realm of mean field games. However, their mathematical setting

is rather different from ours: in that paper state variables are real with Gaussian initial distribution; moreover, their linear quadratic optimization problem is solved by affine controls which preserve Gaussianity; as a consequence, their optimal control is always unique. Models close to the one proposed here, but without individual preferences, have also been introduced as examples of non-uniqueness of equilibria in mean field games (see, e.g., [1, 3, 10, 12, 17, 18, 21]). Closed in spirit are also mean field games of interacting rotators [8, 22, 23], which exhibit a synchronization/incoherence phase transition.

Similarly to what is contained in the last cited references, the model that we are proposing here has the following remarkable feature: for each N, there is a unique Nash equilibrium for the N-player game; however, the corresponding mean field game may have multiple equilibria. This is reminiscent of a common paradigm in Statistical Physics: finite volume Gibbs states are uniquely defined, but thermodynamic limit may be non-unique, indicating a *phase transition*. The analogy with models in Statistical Physics can be carried on further: the model that we propose corresponds to the mean field Ising model (or Curie–Weiss one), when there are no private signals/types, and to the random field Curie–Weiss model, when disorder is introduced. The time horizon T plays a role similar to the inverse temperature in the models cited above: the higher T, the smaller the contribution of the running cost. The N-player game as well as its mean field limit are presented in Sect. 2, following the general theory developed in [10, 14, 15]. In remarkable analogy with the Curie–Weiss model, we show that the mean field game has a unique equilibrium for small T, whereas several equilibria emerge as T increases. For mean field games with multiple equilibria at least four criteria for the selection of a "preferred" equilibrium have been proposed [12, 18, 21]:

- limit of the unique equilibrium of the N-player game,
- minimization of the player cost,
- regularization by vanishing *common noise*,
- stability for the best response map.

These criteria are not equivalent, and we are not aware of general results concerning their relations. We stress that selecting one equilibrium does not imply that the remaining equilibria are meaningless. Indeed, the feedback strategy corresponding to *any* equilibrium of the mean field game is an "approximate" Nash equilibrium for the N-player game, as shown in [9].

A detailed study of the different equilibria of the mean field limit is proposed in Sect. 3. In particular, we will see that a number of different types of equilibria can be identified: polarized/unpolarized (related to the size of the majority), coherent/incoherent (alignment of the final population state with the initial state). In Sect. 4, we discuss the selection of the equilibrium obtained by taking the limit of the unique equilibrium of the N-player game. In [10], in the absence of individual preferences and with a much simpler phase diagram, this question was rigorously answered, while a rigorous analysis is presently out of reach here. We, therefore, run numerical simulations to capture this selection. We see that there is, indeed, a unique equilibrium emerging from the N-player approximation: it is always coherent, but it could be polarized or not, depending on some parameters of the model. Notably, the prevailing equilibrium is not necessarily the one that minimizes the aggregate cost suffered by

the population of interacting agents. Some remarks about the rationale behind the selection of the equilibrium in the case of a finite population are collected in Sect. 4.2. Section 5 contains some concluding remarks. Appendix contains all technical proofs of the results stated in Sect. 3.

2 A Continuous-Time Binary Strategic Game

In this section, we apply the general theory in [15] to study the equilibria of the N-player game and of the mean field game in the specific model we propose.

2.1 The *N*-Player Game

We consider N players whose binary state vector is denoted by $\mathbf{x} := (x_1, \dots, x_N) \in \{-1, 1\}^N$. To each player is also assigned a variable $y_i \in \{-\epsilon, \epsilon\}$, where $\epsilon > 0$ is a given constant, and we set $\mathbf{y} := (y_1, \dots, y_N)$; the components of $\mathbf{y}$ will be referred to as *local fields*. The vector state $\mathbf{x} = \mathbf{x}(t)$ evolves in continuous time, while $\mathbf{y}$ is static. Each player is allowed to control her state with a feedback control $u_i(t, \mathbf{x}, \mathbf{y})$ which may depend on time, and on the values of $\mathbf{x}$ and $\mathbf{y}$. We assume each u_i, as function of t, to be nonnegative, measurable and locally integrable. Thus, for a given control $\mathbf{u} = (u_1, \dots, u_N)$, the state of the system, $\mathbf{x}(t)$, evolves as a Markov process, whose law is uniquely determined as the solution of the martingale problem for the time-dependent generator

$$\mathcal{L}_t f(\mathbf{x}) := \sum_{i=1}^{N} u_i(t, \mathbf{x}, \mathbf{y}) \Big[f(\mathbf{x}^i) - f(\mathbf{x}) \Big] =: \sum_{i=1}^{N} u_i(t, \mathbf{x}, \mathbf{y}) \nabla^i f(\mathbf{x}),$$

where $\mathbf{x}^i$ is the vector state obtained from $\mathbf{x}$ by replacing the component x_i with $-x_i$. In order to fully define the dynamics, we prescribe the joint distribution of the initial states $\mathbf{x}(0)$ and of the local fields $\mathbf{y}$. For simplicity, we assume all these variables are independent: all $x_i(0)$ have mean $m_0 \in [-1, 1]$, whereas all y_i have mean 0. Each player aims at minimizing her own cost, depending on the controls of all players, which is given by

$$J_i(\mathbf{u}) := \mathbb{E}\left[\frac{1}{2} \int_0^T u_i^2(t, \mathbf{x}(t), \mathbf{y}) \mathrm{d}t - x_i(T)(m_N(T) + y_i) \right], \tag{1}$$

where $m_N(t) := \frac{1}{N} \sum_{i=1}^{N} x_i(t)$ is the mean state of the population. Here, $T > 0$ is the time horizon of the game. Besides the standard quadratic running cost in the control, two other terms contribute to the cost:

– the term $-x_i(T) m_N(T)$ favors *polarization*: each agent profits from being aligned with the majority at the final time T;

– the term $-x_i(T)\, y_i$ incentivizes each agent to align with her own local field. As $\mathbf{y}$ is uniformly distributed on $\{-\epsilon, \epsilon\}^N$, this term inhibits alignment of behaviors, hence polarization.

From a technical viewpoint, we note that, by rescaling time, one could normalize to 1 the time horizon and multiplying by T the reward. Thus, the time horizon T may be seen as tuning the relevance of the final reward as compared to the "natural inertia" expressed by the running cost.

Given a control vector $\mathbf{u}$ and a measurable and locally integrable function

$$\beta : [0, T] \times \{-1, 1\}^N \times \{-\epsilon, \epsilon\}^N \to [0, +\infty),$$

we define the control vector $\left[\mathbf{u}^i, \beta\right]$ by

$$\left[\mathbf{u}^i, \beta\right]_j = \begin{cases} u_j & \text{for } j \neq i \\ \beta & \text{for } j = i \end{cases}.$$

Definition 2.1 A control vector $\mathbf{u}$ is a *Nash equilibrium* if for each β as above,

$$J_i(\mathbf{u}) \leq J_i\left([\mathbf{u}^i, \beta]\right), \quad \forall i = 1, \ldots, N.$$

Nash equilibria may be obtained via the Hamilton–Jacobi–Bellman equation (see for details [13])

$$\begin{cases} \frac{\partial v_i}{\partial t}(t, \mathbf{x}, \mathbf{y}) + \sum_{j=1}^N a_*(\nabla^j v_j(t, \mathbf{x}, \mathbf{y}))\nabla^j v_i(t, \mathbf{x}, \mathbf{y}) + \frac{1}{2}a_*^2(\nabla^i v_i(t, \mathbf{x}, \mathbf{y})) = 0 \\ v_i(T, \mathbf{x}, \mathbf{y}) = -x_i(m_N + y_i), \end{cases} \tag{2}$$

where $\nabla^j v_j(t, \mathbf{x}, \mathbf{y}) = v_j(t, \mathbf{x}^j, \mathbf{y}) - v_j(t, \mathbf{x}, \mathbf{y})$, $m_N = \frac{1}{N}\sum_{i=1}^N x_i$, and

$$a_*(p) := \arg\min_{a \geq 0}\left[ap - \frac{1}{2}a^2\right] = p^-,$$

with p^- denoting the negative part of $p \in \mathbb{R}$. Note that (2) is a system of $N \times 2^N \times 2^N$ ordinary differential equations with locally Lipschitz vector field and global solutions. Therefore, it admits a unique solution $\mathbf{v} := (v_1, \ldots, v_N)$; moreover, there exists a unique Nash equilibrium $\mathbf{u}$ given by

$$u_i(t, \mathbf{x}, \mathbf{y}) := a_*(\nabla^i v_i(t, \mathbf{x}, \mathbf{y})), \quad i = 1, \ldots, N.$$

2.2 The Mean Field Game

The mean field game is the formal limit of the above N-player game, as seen from a representative player. Denote, respectively, by $x \in \{-1, 1\}$ and $y \in \{-\epsilon, \epsilon\}$ the state

and the local field of a representative player. Given a (deterministic) $m \in [-1, 1]$, the player aims at minimizing the cost

$$J(u) := \mathbb{E}\left[\frac{1}{2}\int_0^T u^2(t, x(t), y)\mathrm{d}t - x(T)(m + y)\right] \tag{3}$$

under the Markovian dynamics with infinitesimal generator

$$L_t^u f(x, y) := u(t, x, y)[f(t, -x, y) - f(t, x, y)], \tag{4}$$

where $\mathbb{E}[\,\cdot\,]$ denotes the expectation with respect to the noise of the dynamics and to the distribution of the local field y. As above, admissible controls are measurable and locally integrable functions. Consistently with the N-player game, the initial state $x(0)$ and the local field y are independent, with means m_0 and 0, respectively. As we will see below, the convexity of the running cost implies uniqueness of the optimal control u_*^m, which actually depends on the choice of m. Denoting by $x_*^m(t, y)$ the evolution of the state for the optimal control, the solution of the mean field game is completed by finding the solutions of the Consistency Equation

$$m = \mathbb{E}\left(x_*^m(T, y)\right). \tag{5}$$

Therefore, we handle the mean field game first by solving, for m fixed, the optimal control problem (3)–(4) via Dynamic Programming. This leads to the Hamilton–Jacobi–Bellman equation

$$\begin{cases} \frac{\partial V(t,x,y)}{\partial t} + \min_{u\geq 0}\left[u\nabla V(t, x, y) + \frac{1}{2}u^2\right] = 0 \\ V(T, x, y) = -x(m + y) \end{cases} \tag{6}$$

with $\nabla V(t, x, y) := V(t, -x, y) - V(t, x, y)$. The optimal feedback control is given by

$$u_*^m(t, x, y) = (\nabla V(t, x, y))^-.$$

Then, we solve (5) using (4) to obtain the evolution of $m(t, y) := \mathbb{E}_y\left(x_*^m(t, y)\right)$, where $\mathbb{E}_y$ denotes the expectation conditioned to the local field $y \in \{-\epsilon, \epsilon\}$. It follows that

$$\mathbb{E}\left(x_*^m(t, y)\right) = \mathbb{E}(\mathbb{E}_y(x_*^m(t, y))) = \frac{1}{2}\left[m(t, \epsilon) + m(t, -\epsilon)\right].$$

By (4), we obtain the Kolmogorov forward equation

$$\frac{\mathrm{d}}{\mathrm{d}t}m(t, y) = \mathbb{E}_y\left[-2u_*(t, x_*^m(t, y), y)x_*^m(t, y)\right]. \tag{7}$$

Now, we proceed with the explicit solutions of the steps just described. Next, we discuss the solutions of the Consistency Equation (5) in terms of the three parameters of the model: T, ϵ and m_0.

2.3 Solving the Hamilton–Jacobi–Bellman Equation

Our aim here is to determine the value function $V(t, x, y)$ which solves (6). Note that this value function also depends on m. It is convenient to set $z(t, y) := V(t, -1, y) - V(t, 1, y)$. Note also that $\nabla V(t, x, y) = xz(t, y)$. Using (6), we can subtract the two equations for $V(t, -1, y)$ and $V(t, 1, y)$, obtaining a closed equation for $z(t, y)$:

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} z(t, y) = \frac{1}{2} z(t, y)|z(t, y)| \\ z(T, y) = 2(m + y). \end{cases} \tag{8}$$

It is a key fact that the equations for $z(t, \epsilon)$ and $z(t, -\epsilon)$ are decoupled, so they can be solved separately by separation of variables. Indeed, observing that, by uniqueness, the sign of $z(t, y)$ is constant in $t \in [0, T]$, we can rewrite (8) as

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{z(t, y)} \right) = \frac{1}{2} \mathrm{sign}(m + y)$$

that, integrated over $[t, T]$, yields

$$z(t, y) = \frac{2(m + y)}{|m + y|(T - t) + 1}. \tag{9}$$

At this point, we can also compute the value function $V(t, x, y)$. Plugging the optimal control in (6), we get

$$\frac{\partial}{\partial t} V(t, 1, y) = \frac{1}{2} \left(z^-(t, y) \right)^2 = \begin{cases} 0 & \text{if } m + y \geq 0 \\ \left(\frac{2(m+y)}{|m+y|(T-t)+1} \right)^2 & \text{if } m + y < 0. \end{cases}$$

Integrating this last identity from t to T, we get $V(t, 1, y)$. Having $V(t, 1, y)$ and $z(t, y) = V(t, -1, y) - V(t, 1, y)$, we can also obtain $V(t, -1, y)$. The final result is

$$V(t, x, y) = \begin{cases} -|m + y| & \text{if } \mathrm{sign}(m + y) \in \{0, x\} \\ -|m + y| + \frac{2|m+y|}{|m+y|(T-t)+1} & \text{if } \mathrm{sign}(m + y) = -x. \end{cases} \tag{10}$$

2.4 Solving the Kolmogorov Forward Equation

We begin by observing that

$$\begin{aligned} u_*(t, x, y) &= (\nabla V(t, x, y))^- = (xz(t, y))^- \\ &= \frac{1 + x}{2} z^-(t, y) + \frac{1 - x}{2} z^+(t, x) = \frac{1}{2}|z(t, y)| - \frac{x}{2} z(t, y). \end{aligned}$$

Plugging this into (7), we obtain

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} m(t, y) = -m(t, y)|z(t, y)| + z(t, y) \\ m(0, y) = m_0, \end{cases}$$

where we used the fact that $x^2 = 1$. Recalling that the sign of $z(t, y)$ is constant and equals $\rho := \mathrm{sign}(m + y)$, we can rewrite this last equation as

$$-\rho \frac{\mathrm{d}}{\mathrm{d}t} \log(1 - \rho m(t, y)) = z(t, y) = \frac{2(m + y)}{\rho(m + y)(T - t) + 1},$$

which, integrated from 0 to t, yields

$$m(t, y) = \rho \left[1 - (1 - \rho m_0) \left(\frac{|m + y|(T - t) + 1}{|m + y|T + 1} \right)^2 \right]. \tag{11}$$

Hence,

$$\mathbb{E}\left(x_*^m(t, y)\right) = \frac{1}{2} \left(m(t, \epsilon) + m(t, -\epsilon)\right). \tag{12}$$

3 Equilibria and Phase Diagram

This section is entirely devoted to the analysis of the solutions to the Consistency Equation (5). As said, m corresponds to a solution of the MFG if and only if it solves such equation. Relying on (11) and (12), the Consistency Equation can be rewritten as

$$\begin{aligned} &\frac{1}{2}\mathrm{sign}(m + \epsilon) \left[1 - \frac{(1 - \mathrm{sign}(m + \epsilon)m_0)}{(|m + \epsilon|T + 1)^2} \right] \\ &\quad + \frac{1}{2}\mathrm{sign}(m - \epsilon) \left[1 - \frac{(1 - \mathrm{sign}(m - \epsilon)m_0)}{(|m - \epsilon|T + 1)^2} \right] = m. \end{aligned} \tag{13}$$

Solutions of (13) will be called *equilibria*. We are now going to identify all such equilibria. Moreover, depending on the values of the parameters of the model, we classify them emphasizing the presence or the absence of two different features:

- *polarization*, which expresses the fact that agent alignment outscores individual preference, namely, $|m| > \epsilon$;
- *coherence*, which indicates the fact that the majority of the agents aligns with the sign of the initial condition m_0.

We begin by pointing out a symmetry property: if we denote by $F(m, \epsilon, T, m_0)$ the l.h.s. of (13), then

$$F(-m, \epsilon, T, -m_0) = F(m, \epsilon, T, m_0). \tag{14}$$

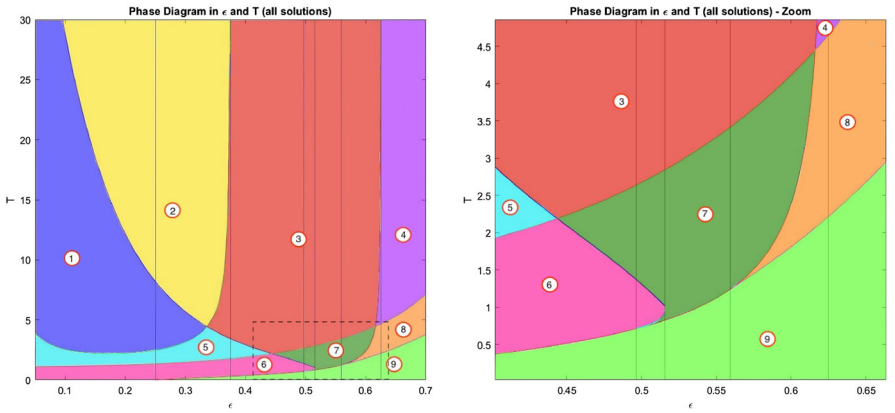

Fig. 1 Full phase diagram for $m_0 = 0.25$. Different regions identify different properties of the solutions to (13), both in terms of numerosity and classification

Therefore, without loss of generality, in the remainder of this article we study (13) in the case $m_0 \geq 0$. We can thus specify precisely four classes of equilibria:

- equilibria $m \in (\epsilon, 1]$ will be called *polarized coherent*;
- equilibria $m \in [-1, -\epsilon]$ will be called *polarized incoherent*;
- equilibria $m \in (0, \epsilon]$ will be called *unpolarized coherent*;
- equilibria $m \in [-\epsilon, 0]$ will be called *unpolarized incoherent*.

Before stating the formal results describing in detail all the solutions to (13), we provide a visual example, where all the possible situations are depicted. To this aim, in Fig. 1 we plot the full phase diagram in the parameters ϵ, T, having fixed $m_0 = 0.25$. The right picture is a zoom of the region within the dashed lines. We can identify nine regions, each of them corresponding to a specific typology of solutions to (13). For example, *region 1*, characterized by an intermediate value of T and ϵ small, shows the presence of three equilibria, one polarized/coherent, one polarized/incoherent, and one unpolarized/incoherent.

In Table 1, we summarize the results in terms of number of equilibria and their typology for all the regions on the phase diagram as depicted in Fig. 1. We note that in regions 6 and 9, for T small, there is a unique equilibrium which is always coherent, and it is polarized in 6, for ϵ small, whereas it is unpolarized in 9, for ϵ large. On the opposite, in zones 2, 3 and 4, for T large, there are five equilibria, and three of them are always coherent. Finally, in zones 1, 5, 7 and 8 (for T intermediate), there are three equilibria. In this situation, we see two different zones: in regions 1 and 5 (for ϵ small), only one equilibrium is coherent; in regions 7 and 8 (for ϵ large), they are all coherent.

Note that the way the number of equilibria depends of the parameters ϵ and T is far from obvious. For instance, fixing $\epsilon \simeq 0.5$, the number of equilibria is not monotonic in T. Similarly, fixing $T \simeq 3$, there is no monotonicity in ϵ.

In the remainder of this section, we state the results describing the phase diagram, specifying at what times the phase transitions occur. This will also serve to specify the algebraic form of all the curves separating the regions depicted in Fig. 1. To ease

Table 1 Total number of equilibria for each region of the phase diagram as in Fig. 1, and number of equilibria disentangled for their typology (coherent/incoherent; polarized/unpolarized)

Region	Number of equilibria	Polarized coherent	Polarized incoherent	Unpolarized coherent	Unpolarized incoherent
①	3	1	1	0	1
②	5	1	1	2	1
③	5	1	2	2	0
④	5	2	2	1	0
⑤	3	1	2	0	0
⑥	1	1	0	0	0
⑦	3	1	0	2	0
⑧	3	2	0	1	0
⑨	1	0	0	1	0

readability, we organize them in four propositions, one for each type of equilibrium of the MFG, i.e., one for each class identified by the possible polarization or coherence of the equilibria, which are listed in Table 1. As already mentioned, we restrict to the case $m_0 \geq 0$.

In Proposition 3.1, we study polarized coherent equilibria m, i.e., those for which $m > \epsilon$. We identify four regions for the parameter ϵ:

- *Small* ϵ: $\epsilon \leq m_0$. For every value of T, there is a unique equilibrium.
- *Low intermediate* ϵ: $m_0 < \epsilon \leq \epsilon_*^{(1)}$. There exists a critical time below which there is no equilibrium, and above which there is a unique equilibrium.
- *High intermediate* ϵ: $\epsilon_*^{(1)} < \epsilon < \frac{1+m_0}{2}$. There are two critical times, and the number of equilibria varies from zero to two to one as T increases and crosses these critical values.
- *Large* ϵ: $\epsilon \geq \frac{1+m_0}{2}$. There is a unique critical time with the number of equilibria going from zero to two as T crosses it.

Proposition 3.1 (Polarized coherent equilibria: $m > \epsilon$)

(i) Suppose $\epsilon \leq m_0$. *Then,* $\forall T > 0$, *there is a unique equilibrium* $m = M(T, \epsilon, m_0)$ *in* $(\epsilon, 1]$, *[regions 6, 5, 1, 2 in Fig. 1]. Moreover,* $\lim_{T \to +\infty} M(T, \epsilon, m_0) = 1$.

(ii) Suppose $\epsilon \geq \frac{1+m_0}{2}$. *Then, there exists a unique* $T_c^{(1)} = T_c^{(1)}(\epsilon, m_0) > 0$ *such that the graph of the curve in the plane* (z, m) *of equation* $z = F(m, \epsilon, T_c^{(1)}, m_0)$ *is tangent to the line of equation* $z = 0$. *Moreover,*

- *for* $T < T_c^{(1)}(\epsilon, m_0)$, *there is no equilibrium in* $(\epsilon, 1]$, *[region 9];*
- *for* $T = T_c^{(1)}(\epsilon, m_0)$, *there is a unique equilibrium in* $(\epsilon, 1]$, *[separatrix of regions 9 and 8];*
- *for* $T > T_c^{(1)}(\epsilon, m_0)$, *there are two equilibria in* $(\epsilon, 1]$, *[regions 8, 4].*

(iii) Suppose $m_0 < \epsilon < \frac{1+m_0}{2}$. *Define*

$$T^*(\epsilon, m_0) := -\frac{1}{2\epsilon} + \frac{1}{2\epsilon}\sqrt{\frac{1-m_0}{1+m_0-2\epsilon}} > 0. \tag{15}$$

Then, there exists a unique $\epsilon_*^{(1)} \in \left(m_0, \frac{1+m_0}{2}\right)$ *such that* $\frac{\partial}{\partial m} F(\epsilon, \epsilon, T^*(\epsilon, m_0), m_0) = 0$. *Moreover, if* $m_0 < \epsilon \leq \epsilon_*^{(1)}$,

- *for* $T \leq T^*(\epsilon, m_0)$, *there is no equilibrium in* $(\epsilon, 1]$, *[region 9];*
- *for* $T > T^*(\epsilon, m_0)$, *there is a unique equilibrium in* $(\epsilon, 1]$, *[regions 6, 7, 5, 3, 1, 2].*

If $\epsilon_*^{(1)} < \epsilon < \frac{1+m_0}{2}$, *there exists a unique* $T_c^{(1)}(\epsilon, m_0) > 0$ *defined as for the case* $\epsilon \geq \frac{1+m_0}{2}$. *Furthermore,* $T_c^{(1)}(\epsilon, m_0) < T^*(\epsilon, m_0)$, *for each* m_0 *the map* $\epsilon \mapsto T_c^{(1)}(\epsilon, m_0)$ *is continuous,* $T_c^{(1)}(\epsilon, m_0) \to T^*(\epsilon_*^{(1)}, m_0)$ *as* $\epsilon \downarrow \epsilon_*^{(1)}$, $T_c^{(1)}$ *defined here for* $\epsilon \in (\epsilon_*^{(1)}, \frac{1+m_0}{2})$ *connects continuously at* $\epsilon = \frac{1+m_0}{2}$ *with* $T_c^{(1)}$ *defined for* $\epsilon \geq \frac{1+m_0}{2}$, *and*

- *for* $T < T_c^{(1)}(\epsilon, m_0)$*, there is no equilibrium in* $(\epsilon, 1]$*, [region 9];*
- *for* $T = T_c^{(1)}(\epsilon, m_0)$*, there is a unique equilibrium in* $(\epsilon, 1]$*, [separatrix of 9 and 8];*
- *for* $T_c^{(1)}(\epsilon, m_0) < T < T^*(\epsilon, m_0)$*, there are two equilibria in* $(\epsilon, 1]$*, [regions 8, 4];*
- *for* $T \geq T^*(\epsilon, m_0)$*, there is a unique equilibrium in* $(\epsilon, 1]$*, [regions 7, 3].*

In Proposition 3.2, we study polarized incoherent equilibria m, i.e., those for which $m < -\epsilon$: the population polarizes in disagreement with the initial majority. We identify two regions for the parameter ϵ:

- *Small* ϵ: $\epsilon < \frac{1-m_0}{2}$. There are two critical times, and the number of equilibria varies from zero to two to one as T increases and crosses these critical values.
- *Large* ϵ: $\epsilon \geq \frac{1-m_0}{2}$. There is a unique critical time with the number of equilibria going from zero to two as T crosses it.

Proposition 3.2 (Polarized incoherent equilibria: $m < -\epsilon$)

(i) Suppose $\epsilon \geq \frac{1-m_0}{2}$*. Then,*

- *for* $T < T_c^{(1)}(\epsilon, -m_0)$*, there is no equilibrium in* $[-1, -\epsilon)$*, [regions 9, 6, 7, 8];*
- *for* $T = T_c^{(1)}(\epsilon, -m_0)$*, there is a unique equilibrium in* $[-1, -\epsilon)$*, [separatrix of 6, 7, 8 from 5, 3, 4];*
- *for* $T > T_c^{(1)}(\epsilon, -m_0))$*, there are two equilibria in* $[-1, -\epsilon)$*, [regions 5, 3, 4].*

(ii) Suppose $0 < \epsilon < \frac{1-m_0}{2}$*. Then,* $T_c^{(1)}(\epsilon, -m_0) < T_*^{(1)}(\epsilon, -m_0)$*, and,*

- *for* $T < T_c^{(1)}(\epsilon, -m_0)$*, there is no equilibrium in* $[-1, -\epsilon)$*, [region 6];*
- *for* $T = T_c^{(1)}(\epsilon, -m_0)$*, there is a unique equilibrium in* $[-1, -\epsilon)$*, [separatrix of 6 and 5];*
- *for* $T_c^{(1)}(\epsilon, -m_0) < T < T_*^{(1)}(\epsilon, -m_0)$*, there are two equilibria in* $[-1, -\epsilon)$*, [region 5];*
- *for* $T \geq T^*(\epsilon, -m_0)$*, there is a unique equilibrium in* $[-1, -\epsilon)$*, [regions 1, 2].*

In Proposition 3.3, we study unpolarized coherent equilibria m, i.e., those for which $0 \leq m \leq \epsilon$. We identify five regions for the parameter ϵ:

- *Small* ϵ: $\epsilon \leq m_0$. There is a unique critical time with the number of equilibria going from zero to two as T crosses it.
- *Low intermediate* ϵ: $m_0 < \epsilon \leq \epsilon_*^{(2)}$. There are two critical times, and the number of equilibria varies from one to zero to two as T increases and crosses these critical values.
- *Intermediate* ϵ: $\epsilon_*^{(2)} < \epsilon < \epsilon_*^{(3)}$. There are three critical times, and the number of equilibria varies from one to two to two to zero to two as T increases and crosses these critical values.
- *High intermediate* ϵ: $\epsilon_*^{(3)} \leq \epsilon < \frac{1+m_0}{2}$. There are two equilibria for all values of T.

– *Large* ϵ: $\epsilon \geq \frac{1+m_0}{2}$. There is a unique equilibrium for all values of T.

Proposition 3.3 (Unpolarized coherent equilibria: $0 \leq m \leq \epsilon$)

(i) Suppose $\epsilon \leq m_0$. *Then, there exists a unique* $T_c^{(2)} = T_c^{(2)}(\epsilon, m_0) > 0$ *such that the graph of the curve in the plane* (z, m) *of equation* $z = F(m, \epsilon, T_c^{(2)}, m_0)$ *is tangent to the line of equation* $z = 0$. *Moreover,*

- *for* $T < T_c^{(2)}(\epsilon, m_0)$, *there is no equilibrium in* $(0, \epsilon)$, *[regions 6, 5, 1];*
- *for* $T = T_c^{(2)}(\epsilon, m_0)$, *there is a unique equilibrium in* $(0, \epsilon)$, *[separatrix of 1 and 2];*
- *for* $T > T_c^{(2)}(\epsilon, m_0)$, *there are two equilibria in* $(0, \epsilon)$, *[region 2].*

(ii) Suppose $\epsilon \geq \frac{1+m_0}{2}$. *Then,* $\forall T > 0$, *there is a unique equilibrium* $m = M(T, \epsilon, m_0)$ *in* $(0, \epsilon)$, *[regions 9, 8, 4]. Moreover,* $\lim_{T \to +\infty} M(T, \epsilon, m_0) = 0$.

(iii) Suppose $m_0 < \epsilon < \frac{1+m_0}{2}$ *and let* $T^*(\epsilon, m_0)$ *given by* (15). *For* $T \leq T^*(\epsilon, m_0)$, *there is a unique solution in* $(0, \epsilon]$, *[regions 9, 8, 4]; the solution is* ϵ *if and only if* $T = T^*(\epsilon, m_0)$. *Consider now the function*

$$V(s, \epsilon, m_0) := \frac{\sqrt{s}}{8}\left\{16 + \sqrt{s}\left[16 + 3m_0\left(\frac{2m_0}{\epsilon s}\right)^{1/3}\right]\right\}$$
$$+\frac{\epsilon}{2}(1+\sqrt{s})^2\left\{-2 - 4\sqrt{s} + s\left[-2 + 3\left(\frac{2m_0}{\epsilon s}\right)^{1/3}\right]\right\}.$$

Then,

- *the equation for the unknown* ϵ, $V\left(\frac{m_0}{4\epsilon}, \epsilon, m_0\right) = 0$ *has a unique solution* $\epsilon_*^{(2)}(m_0) \in \left(m_0, \frac{1+m_0}{2}\right)$ *(unless* $m_0 = 0$ *for which* $\epsilon_*^{(2)}(m_0) = 0$*);*
- *there is a unique* $\epsilon_*^{(3)}(m_0) \in \left(\epsilon_*^{(2)}(m_0), \frac{1+m_0}{2}\right)$ *such that the curve in the plane* (s, z) *of equation* $z = V(s, \epsilon_*^{(3)}(m_0), m_0)$ *is tangent to the line of equation* $z = 0$.

Moreover,

(a) if $m_0 < \epsilon \leq \epsilon_*^{(2)}(m_0)$, *the critical time* $T_c^{(2)} = T_c^{(2)}(\epsilon, m_0) > 0$ *defined in point (i) is well defined, and* $T^*(\epsilon, m_0) < T_c^{(2)}(\epsilon, m_0)$*;*
- *for* $T^*(\epsilon, m_0) < T \leq T_c^{(2)}(\epsilon, m_0)$, *there is no equilibrium in* $(0, \epsilon]$, *[regions 6, 5, 1];*
- *for* $T = T_c^{(2)}(\epsilon, m_0)$, *there is a unique equilibrium in* $(0, \epsilon)$, *[separatrix of 1 from 2, 3];*
- *for* $T > T_c^{(2)}(\epsilon, m_0)$, *there are two equilibria in* $(0, \epsilon)$, *[regions 3, 2];*

(b) if $\epsilon_*^{(2)}(m_0) < \epsilon < \epsilon_*^{(3)}(m_0)$, $T_c^{(2)}$ *is not well defined. This implies that there are two times* $T_c^{(2)}(\epsilon, m_0) < T_c^{(3)}(\epsilon, m_0)$ *such that the graph of the curve in the plane* (z, m) *of equation* $z = F(m, \epsilon, T_c^{(2)}, m_0)$ *is tangent to the line of equation* $z = 0$ *and*

- *for* $T^*(\epsilon, m_0) < T \leq T_c^{(2)}(\epsilon, m_0)$*, there are two equilibria in* $(0, \epsilon]$*, [region 7 in Fig. 1];*
- *for* $T = T_c^{(2)}(\epsilon, m_0)$*, there is a unique equilibrium in* $(0, \epsilon)$*, [separatrix of 7 and 6];*
- *for* $T_c^{(2)}(\epsilon, m_0) < T < T_c^{(3)}(\epsilon, m_0)$*, there is no equilibrium in* $(0, \epsilon)$*, [region 6];*
- *for* $T = T_c^{(3)}(\epsilon, m_0)$*, there is a unique equilibrium in* $(0, \epsilon)$*, [separatrix of 6 and 7];*
- *for* $T > T_c^{(3)}(\epsilon, m_0)$*, there are two equilibria in* $(0, \epsilon])$*, [region 3];*

(c) if $\epsilon_*^{(3)}(m_0) \leq \epsilon < \frac{1+m_0}{2}$*, there are two equilibria in* $(0, \epsilon])$*,* $\forall T > T^*(\epsilon, m_0)$*, [regions 7, 3].*

In Proposition 3.4, we study unpolarized incoherent equilibria m, i.e., those for which $-\epsilon \leq m < 0$. We identify two regions for the parameter ϵ:

- *Small* ϵ: $\epsilon < \frac{1-m_0}{2}$. There is a unique critical time with the number of equilibria going from zero to one as T crosses it.
- *Large* ϵ: $\epsilon \geq \frac{1-m_0}{2}$. There is no equilibrium for all values of T.

Proposition 3.4 (Unpolarized incoherent equilibria: $-\epsilon \leq m < 0$)

(i) Suppose $\epsilon \geq \frac{1-m_0}{2}$*. Then, there is no equilibrium in* $[-\epsilon, 0)$*,* $\forall T > 0$*, [reg. 9, 6, 7, 8, 5, 3, 4].*

(ii) Suppose $\epsilon < \frac{1-m_0}{2}$ *and let* $T^*(\epsilon, m_0)$ *be given by* (15)*. Then,*

- *for* $T \leq T^*(\epsilon, -m_0)$*, there is no equilibrium in* $(-\epsilon, 0)$ *and* $m = -\epsilon$ *is an equilibrium if and only if* $T = T^*(\epsilon, -m_0)$*, [regions 9, 6, 5, 3];*
- *for* $T > T^*(\epsilon, -m_0)$*, there is a unique equilibrium in* $[-\epsilon, 0)$*, [regions 1, 2].*

4 The *N*-Player Game: HJB and Numerical Results

In this section, we provide a different glance to the problem and consider again the representative agent in a setting where she is best responding to a population of N opponents (note that, in doing this, the population is thus formed by $N + 1$ players). We now derive a new *large dimensional* HJB equation used to run simulations of a finite population in order to identify the emerging (unique) equilibrium in the finite dimensional model. This approach is inspired by [17, 18]. Different numerical methods for finite state N-player games are developed in [16].

All parameters and variables are as described in the previous sections. Recall that each agent j is characterized by a predetermined *local field* $y_j \in \{-\epsilon, \epsilon\}$, where $\epsilon \in [0, 1]$ and by a time-varying state variable $x_j(t) \in \{-1, 1\}$. We take agent i playing the role of the representative agent. Concerning the remaining population of N players, we introduce two summary statistics as the number of "ones" in the two subpopulations with different local fields (i.e., different ϵ). To this aim, we define

$$n_N^+ = \sum_{j \neq i} \mathbb{I}_{\{x_j = 1\}} \mathbb{I}_{\{y_j = \epsilon\}}; \quad n_N^- = \sum_{j \neq i} \mathbb{I}_{\{x_j = 1\}} \mathbb{I}_{\{y_j = -\epsilon\}}; \quad n_N^\epsilon = \sum_{j \neq i} \mathbb{I}_{\{y_j = \epsilon\}}.$$

Note that n_N^ϵ is a static variable, whereas n_N^+ and n_N^- change in time and take values, respectively, in $\{0,\ 1,\ \ldots,\ n_N^\epsilon\}$ and $\{0,\ 1,\ \ldots,\ N - n_N^\epsilon\}$. By taking advantage of the symmetries of the model, we search equilibrium controls that, for the representative player i, are feedback depending on the state x_i, on the local field y_i and on the aggregate variables n_N^+, n_N^- and n_N^ϵ, and symmetrically for all other players. We denote by $\alpha(x_i,\ y_i,\ n^+,\ n^-,\ n^\epsilon,\ t)$ the feedback control strategy of player i, while each other player $j \neq i$ uses the feedback control

$$\begin{aligned}&\beta(x_j,\ y_j,\ n^+ - \mathbb{I}_{\{x_j=1\}}\mathbb{I}_{\{y_j=\epsilon\}} + \mathbb{I}_{\{x_i=1\}}\mathbb{I}_{\{y_i=\epsilon\}},\\ &n^- - \mathbb{I}_{\{x_j=1\}}\mathbb{I}_{\{y_j=-\epsilon\}} + \mathbb{I}_{\{x_i=1\}}\mathbb{I}_{\{y_i=-\epsilon\}},\ n^\epsilon - \mathbb{I}_{\{y_j=\epsilon\}} + \mathbb{I}_{\{y_i=\epsilon\}},\ t),\end{aligned}$$

where, for instance, we have used the fact that

$$n^+ - \mathbb{I}_{\{x_j=1\}}\mathbb{I}_{\{y_j=\epsilon\}} + \mathbb{I}_{\{x_i=1\}}\mathbb{I}_{\{y_i=\epsilon\}} = \sum_{k\neq j} \mathbb{I}_{\{x_k=1\}}\mathbb{I}_{\{y_k=\epsilon\}}.$$

Under these assumptions, the triple (x_i, n^+, n^-) is a sufficient statistics, in the sense that its time evolution is Markovian, with transition rates:

$$\begin{aligned}(x, n^+, n^-) &\mapsto (-x, n^+, n^-) \text{ with rate } u(t) = \alpha(x, y, n^+, n^-, n^\epsilon, t);\\ (x, n^+, n^-) &\mapsto (x, n^+ + 1, n^-) \text{ with rate } \gamma^+(x, n^+, n^-, n^\epsilon, t)\\ &\qquad = \Big(n^\epsilon - n^+\Big)\cdot \beta\Big(-1, +\epsilon, n^+ + \mathbb{I}_{\{x=1\}}, n^-, n^\epsilon, t\Big);\\ (x, n^+, n^-) &\mapsto (x, n^+ - 1, n^-) \text{ with rate } \delta^+(x, n^+, n^-, n^\epsilon, t)\\ &\qquad = n^+ \cdot \beta\Big(1, +\epsilon, n^+ - \mathbb{I}_{\{x=-1\}}, n^-, n^\epsilon, t\Big);\\ (x, n^+, n^-) &\mapsto (x, n^+, n^- + 1) \text{ with rate } \gamma^-(x, n^+, n^-, n^\epsilon, t)\\ &\qquad = \Big(N - n^\epsilon - n^-\Big)\cdot \beta\Big(-1, -\epsilon, n^+, n^- + \mathbb{I}_{\{x=1\}}, n^\epsilon, t\Big);\\ (x, n^+, n^-) &\mapsto (x, n^+, n^- - 1) \text{ with rate } \delta^-(x, n^+, n^-, n^\epsilon, t)\\ &\qquad = n^- \cdot \beta\Big(1, -\epsilon, n^+, n^- - \mathbb{I})\{x = -1\}, t\Big).\end{aligned}$$

This allows a considerable reduction in the cardinality of the state space, from 2^{N+1} to $O(N^2)$. The best response $u(t) = \alpha(x(t),\ y, n^+(t), n^-(t), n^\epsilon, t)$ for player i is the one minimizing the cost

$$\mathbb{E}\left[\int_0^T \frac{u^2(t)}{2}\,\mathrm{d}t - x(T)\,(m_{N+1}(T) + y)\right],$$

with

$$m_{N+1}(T) = 2\frac{(n^+(t) + n^-(t)) + \mathbb{I}_{\{x=1\}}}{N+1} - 1.$$

By Dynamic Programming, the Value Function for this stochastic optimal control problem solves

$$\frac{\partial V}{\partial t} + \min_u \left[\frac{u^2}{2} + u\nabla_x V + \gamma^+ \nabla_\gamma^+ V + \delta^+ \nabla_\delta^+ V + \gamma^- \nabla_\gamma^- V + \delta^- \nabla_\delta^- V \right] = 0$$
$$V(x, y, n^+, n^-, T) = -x \left(\frac{2(n^+ + n^- + \mathbb{I}_{\{x=1\}})}{N+1} - 1 \right).$$

The minimization over u leads to the optimal feedback

$$\alpha^*(x, y, n^+, n^-, t) = \left[V(-x, y, n^+, n^-, t) - V(x, y, n^+, n^-, t) \right]^- \tag{16}$$

and, finally, to the HJB equation

$$\begin{aligned} \dot{V} = \frac{1}{2} &\left(\left[V(-x, y, n^+, n^-, t) - V(x, y, n^+, n^-, t) \right]^- \right)^2 \\ &- \gamma^+(x, n^+, n^-, t) \cdot [V(x, +\epsilon, n^+ + 1, n^-, t) - V(x, +\epsilon, n^+, n^-, t)] \\ &- \delta^+(x, n^+, n^-, t) \cdot [V(x, +\epsilon, n^+ - 1, n^-, t) - V(x, +\epsilon, n^+, n^-, t)] \\ &- \gamma^-(x, n^+, n^-, t) \cdot [V(x, -\epsilon, n^+, n^- + 1, t) - V(x, -\epsilon, n^+, n^-, t)] \\ &- \delta^-(x, n^+, n^-, t) \cdot [V(x, -\epsilon, n^+, n^- - 1, t) - V(x, -\epsilon, n^+, n^-, t)]. \end{aligned} \tag{17}$$

The unique Nash equilibrium of the game is obtained by setting $\alpha = \beta = \alpha^*$, under which the HJB reduces to a system of $4\,(n^\epsilon + 1)(N - n^\epsilon + 1)$ ordinary differential equations in the state variables V and can be solved numerically. Specifically, we use the software Matlab to solve an ODE system backward in time, meaning that the final conditions play the role of initial conditions and the variation is inverted in time (the r.h.s of (17) is multiplied by -1).

4.1 Simulations of the *N*-Player System

Having obtained the Nash equilibrium feedback control α^* for $N + 1$ players, we rescale the problem to N players and simulate the evolution of the sufficient statistics

$$n^+(t) = \sum_{i=1}^{N} \mathbb{I}_{\{x_i=1\}} \mathbb{I}_{\{y_i=\epsilon\}}, \quad n^-(t) = \sum_{i=1}^{N} \mathbb{I}_{\{x_i=1\}} \mathbb{I}_{\{y_i=-\epsilon\}},$$

which has the Markovian evolution

$$(n^+, n^-) \mapsto (n^+ + 1, n^-) \text{ with rate } \left(n^\epsilon - n^+ \right) \cdot \alpha^*(-1, +\epsilon, n^+, n^-, n^\epsilon, t);$$
$$(n^+, n^-) \mapsto (n^+ - 1, n^-) \text{ with rate } n^+ \cdot \alpha^*(1, +\epsilon, n^+ - 1, n^-, n^\epsilon, t);$$
$$(n^+, n^-) \mapsto (n^+, n^- + 1) \text{ with rate } \left(N - n^\epsilon - n^- \right) \cdot \alpha^*(-1, -\epsilon, n^+, n^-, n^\epsilon, t);$$
$$(n^+, n^-) \mapsto (n^+, n^- - 1) \text{ with rate } n^- \cdot \alpha^*(1, -\epsilon, n^+, n^- - 1, n^\epsilon, t).$$

Initializing appropriately $(n^+(0), n^-(0))$, by independently assigning to each player $x_i(0) = 1$ with probability $\frac{1+m_0}{2}$ and $y_i = \epsilon$ with probability $\frac{1}{2}$, we run simulations of the above dynamics, eventually obtaining samples for $m_N(T) = 2\frac{n^+(T)+n^-(T)}{N} - 1$. Averaging over S independent simulations we estimate the expectation $\mathbb{E}(m_N(T))$.

More in detail, in our first series of experiments we fix $m_0 = 0.25$, and we take different values of ϵ and T. Concerning the number of simulations, we set $S = 100$, whereas the number of agents in the population is $N = 30$. This figure could appear too small to describe a *large population*. However, what we see in our simulations is that the expected values of m_N are approximating very well the asymptotic equilibrium of the mean field game, except for some *transition windows* that we will discuss in more detail. Later, in a second series of experiments, we will also consider $N = 60$. Note that, by increasing N, the numerical problem becomes quickly intractable because of the high dimension of the HJB associated with the N-dimensional system.[1]

In Fig. 2, we compare $\mathbb{E}(m_N(T))$ (red circles, estimated by averaging over the S simulations), the equilibrium emerging in the finite dimensional model, with the mean field equilibria described as the solutions to the Fixed Point Equation (13) (black lines). Specifically, we consider three different values of $\epsilon \in \{0.5,\ 0.52,\ 0.6\}$ and we let T vary from $T = 0$ to a large time, where the largest equilibrium value of $m(T)$ is approaching the limit value of 1. We expect that $\mathbb{E}(m_N(T))$ converges in N towards one of the mean field equilibria solving (13). This is verified in our simulations; notably, for certain values of the parameter ϵ, we see a clear transition from a polarized to a unpolarized equilibrium or vice versa. In fact, by looking at the four panels of Fig. 2, if ϵ is large enough (see panel with $\epsilon = 0.6$), the individual behavior prevails for all T and the population sticks with the smallest (unpolarized and coherent) equilibrium. When ϵ is small enough (see panel with $\epsilon = 0.5$), the selected equilibrium changes continuously in T, as in the large ϵ case, but this equilibrium is unpolarized for small T and polarized for larger T. More interesting is the case of intermediate values of ϵ (see panel with $\epsilon = 0.52$). Here, we see a continuous branch of unpolarized and coherent equilibria existing for all $T > 0$, while a branch of polarized coherent equilibria emerges for T sufficiently large. In this case, the N-player game agrees with the unique unpolarized equilibrium for small T, jumps to the branch of polarized equilibria as it emerges, but for larger T it jumps back to the less polarized equilibrium. We actually see "smooth transitions" rather than "jumps", but this could be due to the small value of N in simulations.

Note that this switch from polarized to unpolarized is not seen for all values of the initial condition m_0, as seen in Fig. 3.

In the next section, we provide a justification behind the emergence of one selected equilibrium, in case that multiple equilibria are present in the mean field limit.

4.2 A Rationale Behind the Equilibrium Selection

In all the numerical experiments, we have performed about $\mathbb{E}(m_N(T))$, the equilibrium emerging in the N-dimensional system, we can recognize two important properties:

[1] For $N = 30$, we have a system of 1024 equations. For $N = 60$, the number increases to 3844.

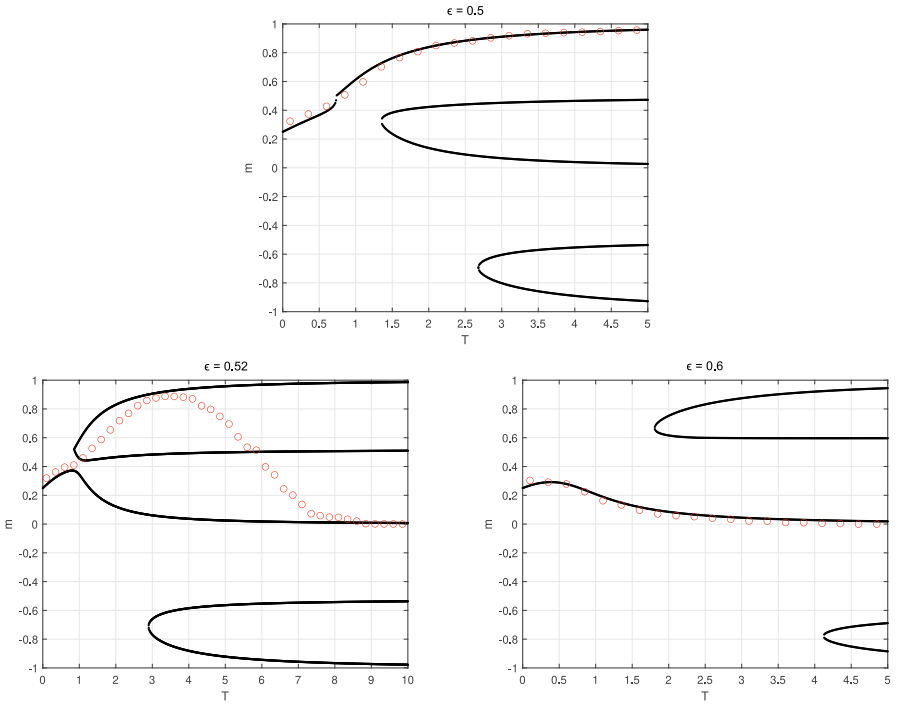

Fig. 2 Values of $m(T)$ (black dots) and $\mathbb{E}(m_N(T))$ (red large circles)

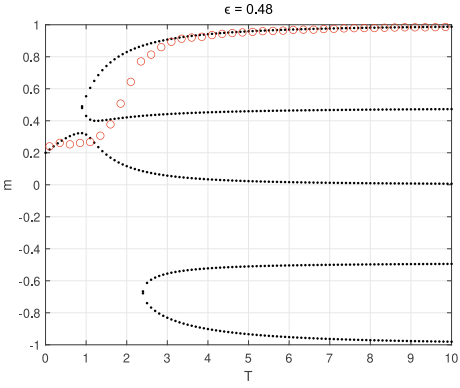

Fig. 3 Values of $m(T)$ (black dots) and $\mathbb{E}(m_N(T))$ (red large circles). Here, $m_0 = 0.2$ and $\epsilon = 0.48$

Property 1. The equilibrium $\mathbb{E}(m_N(T))$ is always coherent. Namely, if $m_0 > 0$, then $m_N(T) > 0$.

Property 2. For some values of ϵ, $\mathbb{E}(m_N(T))$ switches from a polarized to an unpolarized equilibrium (or vice versa) depending on the length of the time horizon T.

Concerning Property 1, when we look at the finite dimensional system, the equilibrium $\mathbb{E}(m_N(T))$ converges, for N large, to one of the values $m(T)$ solving (13). Note that, among those values, there is at least one coherent equilibrium (i.e., an equilibrium with the same sign as m_0). It is then plausible to presume that the finite population will

select one of the coherent equilibria, in that conveying to an incoherent equilibrium would ask for a (implausible) mobilization of the subpopulation ex-ante aligned with the sign of m_0. Property 2., instead, deals with the eventual polarization of the coherent equilibrium selected when playing the finite dimensional game. Here, the discussion is more subtle, since we do not have a clear and trivial explanation of the evident phase transitions we see in the simulations. The simplest explanation could be that the population chooses the equilibrium which minimizes the *total cost*, $J(u)$, defined in (3). Interestingly, this functional, being a function of the control, can be rewritten in terms of $m(T)$. We can take advantage of (10) to see that, depending on the value of $x \in \{-1, +1\}$ and $y \in \{-\epsilon, +\epsilon\}$, we can specify the cost needed to reach a certain equilibrium value $m(T)$:

$$v_{m(T)}(x, y) = \begin{cases} -|m(T) + y| & \text{if } \operatorname{sign}(m(T) + y) \in \{0, x\} \\ -|m(T) + y| + \frac{2|m(T)+y|}{T\,|m(T)+y|+1} & \text{if } \operatorname{sign}(m(T) + y) = -x, \end{cases} \tag{18}$$

where $v_{m(T)}(x, y) = V(0, x, y)$ evaluated at $m = m(T)$. This functional describes the total cost sustained by each subpopulation indexed by x and y to reach the equilibrium $m(T)$. We can now derive the costs sustained by the *underdog* subpopulation (the one whose local filed is opposite in sign to m_0) and by its opponent, namely the one whose local filed is aligned with m_0. With a slight abuse of notation, we denote such quantities with $J^{(-\epsilon)}(m(T))$ and $J^{(+\epsilon)}(m(T))$ to emphasize the dependence on the prevailing equilibrium. Accordingly, we will also write $J(m(T))$ as the total cost sustained by the entire system. We have

$$J^{(-\epsilon)}(m(T)) = \frac{1 - m_0}{2} v_{m(T)}(-1, -\epsilon) + \frac{1 + m_0}{2} v_{m(T)}(+1, -\epsilon)$$
$$J^{(+\epsilon)}(m(T)) = \frac{1 - m_0}{2} v_{m(T)}(-1, +\epsilon) + \frac{1 + m_0}{2} v_{m(T)}(+1, +\epsilon)$$

and

$$J(m(T)) = \frac{1}{2} J^{(-\epsilon)}(m(T)) + \frac{1}{2} J^{(-\epsilon)}(m(T)).$$

It is not difficult to see that, when considering only coherent equilibria (i.e., $m(T)$ such that $\operatorname{sign}(m(T)) = \operatorname{sign}(m_0)$), $J(m(T))$ decreases in $m(T)$, in the sense that the more polarized the equilibrium is, the lower the cost to reach it. Therefore, we could expect the polarized equilibrium to prevail. However, as said, for certain values of the parameters, this is not the case. In line with our simulations, we now shape a different conjecture. We see that the prevailing equilibrium is the one that, among the coherent ones, minimizes the functional $J^{(-\epsilon)}$, namely the cost related to the *underdog* subpopulation. We rephrase this conjecture in the following fact, which embraces both the previous properties.

Property 3. The equilibrium $\mathbb{E}(m_N(T))$ of the N-dimensional system converges to the coherent solution of (13) that minimizes $J^{(-\epsilon)}$.

In some sense, abstracting a two-player game played between the favorite player ($y = +\epsilon$) and the underdog one ($y = -\epsilon$), we can say that the former *imposes* that the equilibrium will be coherent (and this minimizes her effort), whereas the latter decides about polarization (again, minimizing effort given the previous selection).

To provide evidence about the goodness of Property 3, in Fig. 4, we plot the phase diagram of $J^{(-\epsilon)}$ for $m_0 = 0.25$ and for the same values of ϵ and T seen in Fig. 2. As said before, for each equilibrium value $m(T)$ of the mean field limit, we have the corresponding value of $J^{(-\epsilon)}$.

Note that, for $\epsilon = 0.52$, we see two transitions corresponding to the points where the two branches of $J^{(-\epsilon)}$ related to the polarized and unpolarized coherent equilibria intersect themselves. In the lower panel of Fig. 4, we zoom on the right-bottom panel to better recognize such intersection. We see that the two branches of $J^{(-\epsilon)}$ related to the polarized and unpolarized coherent equilibria intersect themselves at $T \approx 8.9$. Notably, this point lies in the time interval, where the emerging equilibrium of the N-dimensional system jumps from the polarized equilibrium to the unpolarized one (panel with $\epsilon = 0.6$ of Fig. 2). We do not report all the figures related to the other values of ϵ, but the same fact still appears, thus corroborating Property 3.

Finally, we show that the solution $m(T)$ related to the prevailing solution $m_N(T)$ does not necessarily minimize to total cost $J(m(T))$. In Fig. 5 (left panel), for $\epsilon = 0.52$, we plot the value of $m(T)$ that minimizes $J^{(-\epsilon)}$ (bold blue circles) and J (thin black line). It can be seen that the two cases coincide as soon as T exceed the value $T \approx 8.9$ discussed above. In the right panel of the same figure, we plot two different branches of J, one corresponding to $m(T)$ which minimizes $J^{(-\epsilon)}$, and the second one related to $m(T)$ which minimizes J (thin black line). We see that the two curves differ exactly for T larger than the intersection value depicted above; this is the value of T where we know that the equilibrium $m_N(T)$ in the finite dimensional model jumps from the polarized to the unpolarized one. This shows that the equilibrium emerging in the population dynamics does not always minimize the total cost. In some sense, the unpolarized equilibrium partially favors the underdog subpopulation, in that $J^{(-\epsilon)}$ is minimized among the coherent equilibria.

We now run a second series of experiment; the aim is still to reinforce the goodness of Property 3, showing that $\mathbb{E}(m_N(T))$ fairly well approximates the coherent mean field game equilibrium minimizing $J^{(-\epsilon)}$. For this round of simulations, we increase the number of agents taking $N = 60$. Concerning the other parameters, we choose different values of m_0, ϵ and T, in order to consider cases where there are one, three or five solutions to (13). Specifically, in the first panel of simulations, we fix $T = 1$ and let m_0 and ϵ vary (cf. the first six experiments reported in Table 2); in the second panel, we fix $m_0 = 0.2$ and let T and ϵ vary (cf. the second six experiments reported in Table 2). In Table 2, we report all results. The first three columns summarize the parameters of each experiment. In the fourth, we report the value of $m(T)$ which matches what prescribed by Property 3. Columns five and six pertain to the finite dimensional model; we indicate by $\overline{m}_N(T)$ the average value of $m_N(T)$ as resulting by running $S = 100$ simulations, with an indication of its standard deviation, $SD(m_N(T))$. Finally, in columns seven and eight, we measure the goodness of the approximation by reporting $|\overline{m}_N(T) - m(T)|$ and $|\overline{m}_N(T) - m(T)|/SD(m_N(T))$, respectively. By looking at these latter columns, it is evident that the equilibrium prevailing in the finite dimensional

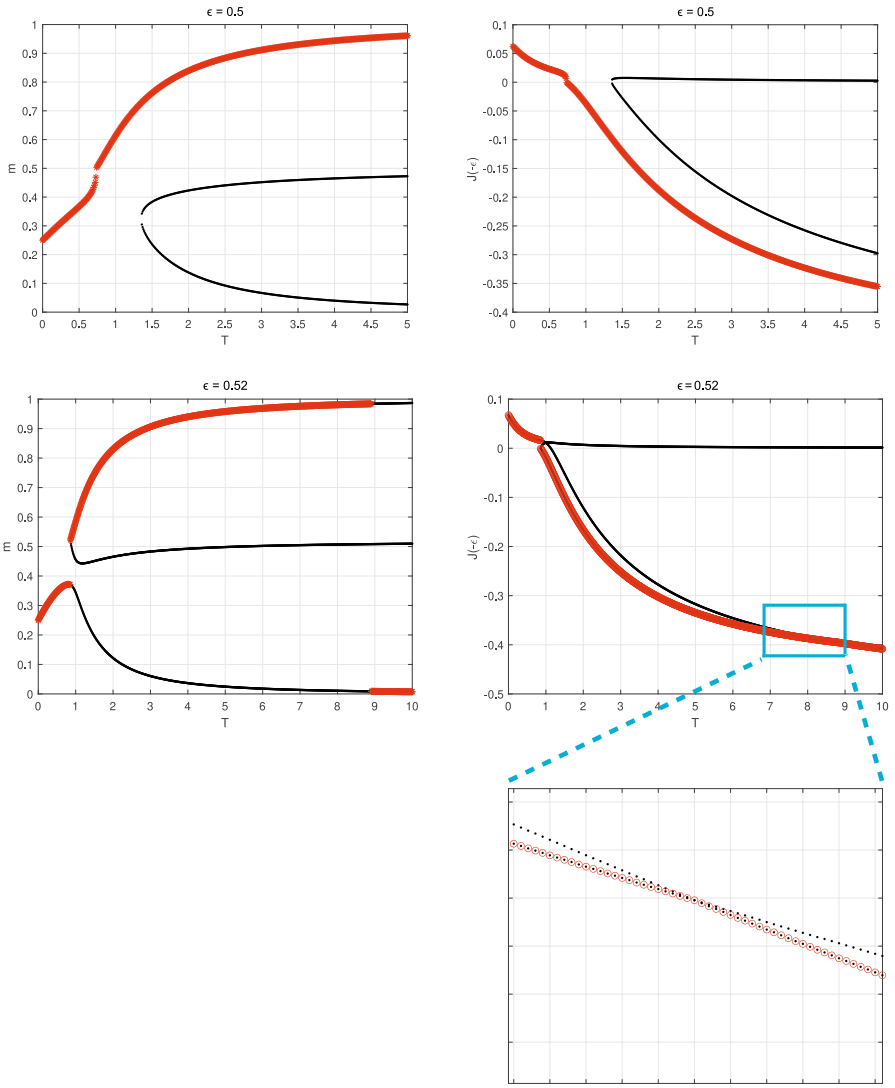

Fig. 4 Phase diagram for $m(T)$ (left panels) and $J^{(-\epsilon)}$ (right panels) for two different values of ϵ. Here, $\epsilon = 0.5$ (top panels) and $\epsilon = 0.52$ (bottom panels). In the left panels, the red points denote the solution $m(T)$ corresponding to the minimum value of $J^{(-\epsilon)}$ (depicted in red in the corresponding right panel)

model aligns with what prescribed by Property 3. This is testified by the fact that the difference between $\overline{m}_N(T)$ and $m(T)$ is close to 0 (the highest value is 0.038 in experiment n.7) and that such difference is always below one standard deviation (the highest ratio is, again, in experiment n.7).

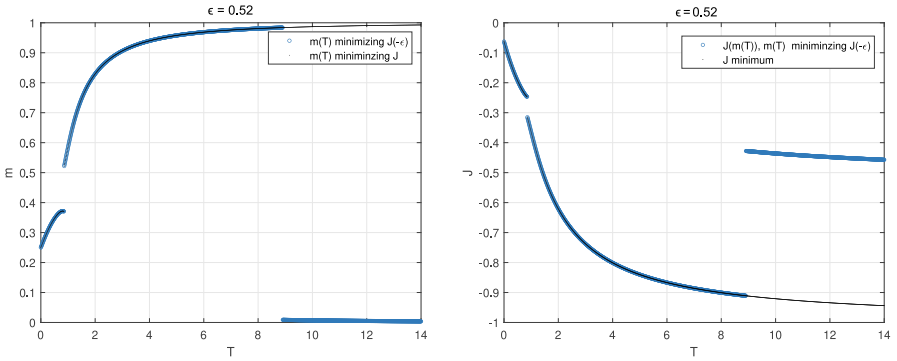

Fig. 5 Left: The value of $m(T)$ minimizing the functional $J^{(-\epsilon)}$ (bold blue circles) and minimizing J (thin black line). Right: The functional J computed for $m(T)$ minimizing $J^{(-\epsilon)}$ (bold circles) and the minimum level of J (thin black line). Here, $\epsilon = 0.52$

Table 2 Results of numerical experiments run to check the goodness of Property 3

T	m_0	ϵ	$m(T)$	$\overline{m}_N(T)$	$SD(m_N(T))$	$\lvert\overline{m}_N(T) - m(T)\rvert$	$\frac{\lvert\overline{m}_N(T)-m(T)\rvert}{SD(m_N(T))}$
1	0.1	0.42	0.8261	0.8240	0.0794	0.0021	0.0264
1	0.1	0.45	0.8126	0.8020	0.0859	0.0106	0.1234
1	0.1	0.5	0.0506	0.0390	0.0765	0.0116	0.1516
1	0.5	0.55	0.8962	0.9171	0.0553	0.0209	0.3779
1	0.5	0.6	0.8818	0.8860	0.0569	0.0042	0.0738
1	0.5	0.7	0.1276	0.1080	0.0783	0.0196	0.2503
2.3	0.2	0.5	0.8552	0.8173	0.0788	0.0379	0.4810
2.3	0.2	0.58	0.0583	0.0460	0.0818	0.0123	0.1504
2.8	0.2	0.5	0.8925	0.8827	0.0715	0.0098	0.1371
3.5	0.2	0.7	0.0203	0.0173	0.0473	0.0030	0.0634
5.5	0.2	0.7	0.0094	0.0077	0.0307	0.0017	0.0554
9	0.2	0.7	0.0039	0.0030	0.0171	0.0009	0.0526

5 Conclusions

We have studied a simple mean field game on a time interval $[0, T]$, where players can control their binary state according to a functional made of a quadratic cost and a final reward. This latter depends on two competing drivers: (i) a social component rewarding conformism, namely, being part of the majority (conformism) and (ii) a private signal favoring the coherence of individuals with respect to a personal type (stubbornness). The trade-off between these two factors, associated with the antimonotonicity of the objective functional, leads to a fairly rich phase diagram. Specifically, the presence of multiple Nash Equilibria for the mean field game has been detected; moreover, when looking at the aggregate outcome of the game, several different types of equilibria can emerge in terms of polarization (fraction of conformists) and coherence (sign of the majority at the final time T compared to the sign of the initial condition).

We have described and characterized the full phase diagram and discussed the role of all the parameters of the model with respect to the aforementioned classification of possible equilibria. We have also analyzed a N-player version of the same mean field game. It is a well-known that in this latter case, the Nash Equilibrium is necessarily unique. It becomes, therefore, interesting to identify which equilibrium is selected by the N-finite population, in case the corresponding mean field game exhibits multiple equilibria. In this respect, we detected phase transitions, in the sense that, depending on one or more parameters of the model, the equilibrium emerging in the finite-dimensional game is always coherent, but it may turn from an unpolarized one to a polarized and vice versa, depending on the length of the time horizon, T. This fact seems to be new in the mean field game literature. At a first glance, we could expect the finite dimensional population to select the equilibrium minimizing the cost associated with the equilibrium for the entire system (interpreted as the collective cost). By contrast, what emerges from our simulations is that the equilibrium prevailing in the finite dimensional game is the one converging, for N large, to the coherent equilibrium that minimizes the cost functional associated with the *ex-ante underdog* subpopulation, namely, the collection of such players whose private signal opposites the sign of the majority at time zero. Put differently, it seems that the *ex-ante* favorite subpopulation (namely, the one whose private signal is aligned with the initial condition) imposes the selection of a coherent equilibrium, whereas the *ex-ante* underdog subpopulation (namely, the one whose private signal opposites the sign of the initial condition) decides about polarization.

Funding Open access funding provided by Universitá degli Studi di Padova within the CRUI-CARE Agreement.

A Proof of Results

Proof of Proposition 3.1: Polarized Coherent Equilibria ($m > \epsilon$)

For $m > \epsilon$, (13) can be rewritten as

$$F(m, \epsilon, T, m_0) := 1 - m - \frac{1 - m_0}{\epsilon^2 T^2} \varphi\left(\frac{1 + mT}{\epsilon\, T}\right) = 0, \tag{19}$$

where

$$\varphi(y) = \frac{y^2 + 1}{(y^2 - 1)^2}.$$

Before describing the solutions of the equation $F(m, \epsilon, T, m_0)$ in $(\epsilon, 1]$, we give some simple facts that will be useful in the proof.

Fact 1. The map $m \mapsto F(m, \epsilon, T, m_0)$ is strictly concave for $m \in (\epsilon, 1]$, for each ϵ, T, m_0.

Proof This comes immediately from the fact that φ is strictly convex in $(1, +\infty)$. Indeed,

$$\varphi'(y) = -\frac{2y(3+y^2)}{(y^2-1)^3}, \quad \varphi''(y) = \frac{6(y^4+6y^2+1)}{(y^2-1)^4}. \tag{20}$$

Note also that φ is strictly decreasing. □

As a consequence of Fact 1, (19) can have at most two solutions in $(\epsilon, 1]$.

Fact 2. The map $T \mapsto F(m, \epsilon, T, m_0)$ is strictly increasing for $T \in [0, +\infty)$, for each m, ϵ, m_0, with $m > \epsilon$.

Proof To see this, note that

$$\frac{\partial F}{\partial T}(m, \epsilon, T, m_0) = \frac{2(1-m_0)\left[m(1+Tm)^3 - 3\epsilon^2 T(1+Tm) - \epsilon^4 T^3\right]}{\left((1+Tm)^3 - \epsilon^2 T^2\right)^3},$$

which has the same sign as $h(m, \epsilon, T, m_0) := m(1+Tm)^3 - 3\epsilon^2 T(1+Tm) - \epsilon^4 T^3$. Observing that $h(\epsilon, \epsilon, T, m_0) = \epsilon$, that

$$\frac{\partial h}{\partial m}(m, \epsilon, T, m_0) = (1+Tm)^3 + 3Tm(1+Tm)^2 - 3\epsilon^2 T^2$$

is increasing in m and $\frac{\partial h}{\partial m}(\epsilon, \epsilon, T, m_0) = 4\epsilon^3 T^3 + 6\epsilon^2 T^2 + 6\epsilon T + 1 > 0$, we conclude that $\frac{\partial F}{\partial T}(m, \epsilon, T, m_0) > 0$ for $m \in (\epsilon, 1]$. This proves Fact2. □

Fact 3. $\forall m, \epsilon, m_0$, with $m > \epsilon$,

$$F(m, \epsilon, 0^+, m_0) := \lim_{T \downarrow 0} F(m, \epsilon, T, m_0) = m_0 - m$$

and

$$\lim_{T \uparrow +\infty} F(\epsilon, \epsilon, T, m_0) = \frac{1 + m_0 - 2\epsilon}{2},$$

whereas for $m \in (\epsilon, 1]$,

$$\lim_{T \uparrow +\infty} F(\epsilon, \epsilon, T, m_0) = 1 - m.$$

These are simple asymptotics, and the proof is omitted.

Now, we can prove Proposition 3.1. We begin with the case $\epsilon \leq m_0$. By Fact 3, $F(\epsilon, \epsilon, 0^+, m_0) = m_0 - \epsilon \geq 0$; thus, by Fact 2, $F(\epsilon, \epsilon, T, m_0) > 0, \forall T > 0$. Since, clearly, $F(1, \epsilon, T, m_0) < 0$ and $m \mapsto F(m, \epsilon, T, m_0)$ is strictly concave by Fact 1, the uniqueness of the solution to (19) follows readily and case (i) is proved.

Now, consider the case $\epsilon \geq \frac{1+m_0}{2}$ (which implies $\epsilon > m_0$). By Facts 2 and 3, $F(\epsilon, \epsilon, T, m_0) < 0, \forall T > 0$. Since $F(m, \epsilon, 0^+, m_0) = m_0 - m$, by continuity (19) has no solution in $(\epsilon, 1]$ for T small enough. Now, we claim that there exists a unique $T_c^{(1)} = T_c^{(1)}(\epsilon, m_0) > 0$ such that the graph of $z = F(m, \epsilon, T_c^{(1)}, m_0)$ is tangent to the line $z = 0$. Since $F(m, \epsilon, T, m_0)$ is strictly increasing in T (Fact 2), such $T_c^{(1)}$, if any, is unique. To prove the existence of $T_c^{(1)}$ by continuity, it is enough to show that there are values of $m \in (\epsilon, 1]$ and $T > 0$ such that $F(m, \epsilon, T, m_0) > 0$: this is true $\forall m \in (\epsilon, 1)$ as $F(m, \epsilon, T, m_0) \to 1 - m$ for $T \uparrow +\infty$ (Fact 3). The conclusions in case (ii) are now obvious consequences of concavity and T-monotonicity of F.

Consider, finally, the case $m_0 < \epsilon < \frac{1+m_0}{2}$. We first investigate the following equation in the unknown $T > 0$:

$$F(\epsilon, \epsilon, T, m_0) = 0, \tag{21}$$

that can be rewritten as

$$1 - \epsilon - \frac{2\epsilon^2 T^2 + 2\epsilon\, T + 1}{4\epsilon^2 T^2 + 4\epsilon\, T + 1}(1 - m_0) = 0.$$

This can be cast as a quadratic equation in T: it has real solutions if and only if $\epsilon < \frac{1+m_0}{2}$, and in this case the only positive solution is $T^*(\epsilon, m_0)$ given in (15). Note that $\epsilon > m_0$ implies that $T^*(\epsilon, m_0) > 0$. Now, T-monotonicity and m-concavity of $F(m, \epsilon, T, m_0)$ and the fact that

$$\frac{\partial}{\partial m} F(\epsilon, \epsilon, T, m_0) = -1 + 2(1 - m_0)\frac{T(\epsilon\, T + 1)(4\epsilon^2 T^2 + 2\epsilon\, T + 1)}{2\epsilon\, T + 1} \tag{22}$$

is strictly increasing in T, show that there are two alternatives:

(a) if $\frac{\partial}{\partial m} F(\epsilon, \epsilon, T^*(\epsilon, m_0), m_0) \leq 0$, then,
- for $T \leq T^*(\epsilon, m_0)$, the map $m \mapsto F(m, \epsilon, T, m_0)$ is less than or equal to 0 at $m = \epsilon$ and it is decreasing: there are no solutions to (19) in $(\epsilon, 1]$;
- for $T > T^*(\epsilon, m_0)$, the map $m \mapsto F(m, \epsilon, T, m_0)$ is strictly positive at $m = \epsilon$, concave and negative at $m = 1$, so (19) has a unique solution in $(\epsilon, 1]$;

(b) if $\frac{\partial}{\partial m} F(\epsilon, \epsilon, T^*(\epsilon, m_0), m_0) > 0$, then by continuity there exists $T_c^{(1)} = T_c^{(1)}(\epsilon, m_0) \in (0, T^*(\epsilon, m_0))$ such that the graph of $z = F(m, \epsilon, T_c, m_0)$ is tangent to the line $z = 0$; as above, this $T_c^{(1)}$ is unique. Thus,
- for $T < T_c^{(1)}(\epsilon, m_0)$, (19) has no solutions in $(\epsilon, 1]$;
- for $T = T_c^{(1)}(\epsilon, m_0)$, (19) has a unique solution in $(\epsilon, 1]$;
- for $T_c^{(1)}(\epsilon, m_0) < T < T^*(\epsilon, m_0)$, (19) has two solutions in $(\epsilon, 1]$;

– for $T \geq T^*(\epsilon, m_0)$, (19) has a unique solution in $(\epsilon, 1]$.

Note that $T_c^{(1)}$ is defined as in case (ii), and by the Implicit Function Theorem it is continuous at ϵ, $\forall \epsilon$ in its domain. To complete the proof of case (iii), we are left to show that there exists $\epsilon_*^{(1)} \in \left(m_0, \frac{1+m_0}{2}\right)$ such that

$$\frac{\partial}{\partial m} F(\epsilon, \epsilon, T^*(\epsilon, m_0), m_0) > 0 \quad \text{if and only if} \quad \epsilon \in \left(\epsilon_*^{(1)}, \frac{1+m_0}{2}\right),$$

so that $T_c^{(1)}(\epsilon, m_0)$ is defined for $\epsilon \in (\epsilon_*^{(1)}, 1]$. Moreover, if this is the case, continuity implies that

$$\lim_{\epsilon \downarrow \epsilon_*^{(1)}} T_c^{(1)}(\epsilon, m_0) = T^*(\epsilon^*, m_0).$$

To complete the proof we are left to show the existence of such $\epsilon_*^{(1)}$. This is established by proving that the map $\epsilon \mapsto \frac{\partial}{\partial m} F(\epsilon, \epsilon, T^*(\epsilon, m_0), m_0)$ is strictly increasing, negative at $\epsilon = m_0$ and diverging to $+\infty$ at $\epsilon = \frac{1+m_0}{2}$. We use the expressions (22) and (15), and the change of variable $y := 1 + 2\epsilon\, T^*(\epsilon)$. Note that, by (15),

$$\epsilon\, T^*(\epsilon) = -\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1-m_0}{1+m_0-2\epsilon}},$$

so $y = \sqrt{\frac{1-m_0}{1+m_0-2\epsilon}}$. It is easily seen that $\frac{dy}{d\epsilon} > 0$, $y(m_0) = 1$ and $\lim_{\epsilon \uparrow \frac{1+m_0}{2}} y(\epsilon) = +\infty$.

We can write, using (22),

$$G(\epsilon, m_0) := \frac{\partial}{\partial m} F(\epsilon, \epsilon, T^*(\epsilon, m_0), m_0) = -1 + (1-m_0)\frac{(y^2-1)(y^2-y+1)}{2\epsilon y}.$$

This shows that $G(m_0, m_0) = -1$ and that $G(\epsilon, m_0)$ diverges to $+\infty$ as $\epsilon \uparrow \frac{1+m_0}{2}$. So the last step is to prove that G is strictly increasing in ϵ. We have

$$\frac{\partial G}{\partial \epsilon}(\epsilon, m_0) = \frac{1-m_0}{2\epsilon}\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{(y^2-1)(y^2-y+1)}{y}\right)\frac{\mathrm{d}y}{\mathrm{d}\epsilon} - \frac{1-m_0}{2\epsilon^2}\frac{(y^2-1)(y^2-y+1)}{y}.$$

Using the facts that $\frac{\mathrm{d}y}{\mathrm{d}\epsilon} = \frac{y^3}{1-m_0}$ and $\epsilon = \frac{(1+m_0)y^2-(1-m_0)}{y^2}$, it follows that $\frac{\partial G}{\partial \epsilon}(\epsilon, m_0)$ has the same sign as

$$\begin{aligned}
&\frac{\epsilon y^3}{1-m_0}\frac{3y^4-2y^3-y^2+1}{y^2} - \frac{(y^2-1)(y^2-y+1)}{y} \\
&= \frac{\frac{1+m_0}{1-m_0}y^2-1}{y}(y^2(y-1)(3y+1)+1) - \frac{(y^2-1)(y^2-y+1)}{y}
\end{aligned}$$

$$\geq \frac{y^2-1}{y}(y^2(y-1)(3y+1)+1) - \frac{(y^2-1)(y^2-y+1)}{y}$$
$$= (y^2-1)(y-1)(3y^2+y-1) > 0, \quad \forall\, y > 1, \tag{23}$$

where we have used the fact that $m_0 \geq 0$ and the expression in the second line of (23) is increasing in m_0. This completes the proof. □

Proof of Proposition 3.2: Polarized Incoherent Equilibria ($m < -\epsilon$)

The proof repeats some of the arguments seen in the proof of Proposition 3.1. It is convenient to take advantage of the symmetry relation (14). This implies that we can equivalently find the equilibria in $(\epsilon, 1]$ after replacing m_0 by $-m_0$. For the regime $\epsilon \geq \frac{1-m_0}{2}$, the proof of part (ii) of Proposition 3.1 applies with no changes, as the assumption $m_0 \geq 0$ was not used. For the case $0 < \epsilon < \frac{1-m_0}{2}$, we can adapt the proof of part (iii) of Proposition 3.1, where the assumption $m_0 \geq 0$ was only used to prove the existence of $\epsilon_*^{(1)}$. Here, we obtain the same behavior seen for $\epsilon_*^{(1)} < \epsilon < \frac{1+m_0}{2}$ in Proposition 3.1: to repeat the same argument we need to show that

$$\frac{\partial}{\partial m} F(\epsilon, \epsilon, T^*(\epsilon, -m_0), -m_0) > 0, \quad \forall \epsilon : \ 0 < \epsilon < \frac{1-m_0}{2}.$$

Indeed, as seen in the proof of Proposition 3.1,

$$\frac{\partial}{\partial m} F(\epsilon, \epsilon, T^*(\epsilon, -m_0), -m_0) = -1 + (1+m_0)\frac{(y^2-1)(y^2-y+1)}{2\epsilon y},$$

with $y = \sqrt{\frac{1+m_0}{1-m_0-2\epsilon}}$. Note that, as $m_0 \geq 0$, $y > 1$ for all $0 < \epsilon < \frac{1+m_0}{2}$. In particular, $y^2 - y + 1 > 1$ so that, with a further simple computation we get

$$\frac{\partial}{\partial m} F(\epsilon, \epsilon, T^*(\epsilon, -m_0), -m_0) > -1 + (1+m_0)\frac{y^2-1}{2\epsilon y} = -1 + \frac{\epsilon + m_0}{\epsilon} y > 0.$$

Thus, the proof for the case $0 < \epsilon < \frac{1-m_0}{2}$ can be carried out in the same way as the proof of part (iii) of Proposition 3.1 (case $\epsilon_*^{(1)} < \epsilon < \frac{1+m_0}{2}$). □

Proof of Proposition 3.3: Unpolarized Coherent Equilibria ($0 \leq m \leq \epsilon$)

For $m \in [0, \epsilon]$, equation (13) becomes:

$$F(m, \epsilon, T, m_0) := \frac{2(\epsilon T + 1)Tm + m_0\left[(\epsilon T+1)^2 + m^2T^2\right]}{\left((\epsilon T+1)^2 - m^2T^2\right)^2} - m = 0. \tag{24}$$

We begin by observing that

$$F(m,\epsilon,T,m_0)=\frac{1}{(\epsilon\,T+1)^2}\varphi\left(\frac{T}{1+\epsilon\,T}m,m_0\right)-m, \tag{25}$$

where

$$\varphi(z,m_0):=\frac{2z+m_0(1+z^2)}{(1-z^2)^2}. \tag{26}$$

In particular, this allows to prove easily that, if $m_0\geq 0$, then $m\mapsto F(m,\epsilon,T,m_0)$ is convex.

(i) Note first that $F(0,\epsilon,T,m_0)=\frac{m_0}{(\epsilon\,T+1)^2}>0$

(recall that here $m_0\geq\epsilon>0$). Moreover,

$$F(\epsilon,\epsilon,T,m_0)=\frac{2(\epsilon\,T+1)\epsilon\,T+m_0\left[(\epsilon\,T+1)+\epsilon^2T^2\right]}{\left((\epsilon\,T+1)^2-\epsilon^2T^2\right)^2}-\epsilon=\psi(\epsilon\,T,m_0)-\epsilon,$$

with

$$\psi(y):=\frac{2(y+1)y+m_0\left[(y+1)^2+y^2\right]}{(2y+1)^2}.$$

As $\psi'(y)=\frac{2(1-m_0)}{(2y+1)^3}$, we deduce that $F(\epsilon,\epsilon,T,m_0)$ is strictly increasing in T, except for $m_0=1$ (in this case it is constant). In all cases, we have that, since $F(\epsilon,\epsilon,0,m_0)=m_0-\epsilon\geq 0$, $F(\epsilon,\epsilon,T,m_0)>0$, $\forall T>0$. Thus, the map $m\mapsto F(m,\epsilon,T,m_0)$ is strictly positive at the endpoints of the interval $(0,\epsilon)$. We also have $F(m,\epsilon,0,m_0)=m_0-m>0\ \forall m\in(0,\epsilon)$. Moreover, it is easily seen that for each $m\in(0,\epsilon)$,

$$\lim_{T\to+\infty}F(m,\epsilon,T,m_0)=-m<0. \tag{27}$$

Then, there must be a time $T_c^{(2)}(\epsilon,m_0)$ such that for $T<T_c^{(2)}(\epsilon,m_0)$, $F(m,\epsilon,T,m_0)>0$, $\forall m\in(0,\epsilon)$, and the graph of the convex function $y=F(m,\epsilon,T_c^{(2)}(\epsilon,m_0),m_0)$ is tangent to the line $y=0$. Set

$$\mathcal{T}:=\left\{T>0:\min_{m\in(0,\epsilon)}F(m,\epsilon,T,m_0)\leq 0\right\}.$$

The proof of this point (i) is completed as we show that $\mathcal{T}=[T_c^{(2)}(\epsilon,m_0),+\infty)$. If this is not the case, by continuity, there must be a time $\hat{T}\geq T_c^{(2)}(\epsilon,m_0)$, $\hat{m}\in(0,\epsilon)$ and $\delta>0$ such that

$$F(\hat{m},\epsilon,\hat{T},m_0)=0,\quad\frac{\mathrm{d}}{\mathrm{d}m}F(\hat{m},\epsilon,\hat{T},m_0)=0, \tag{28}$$

but $F(\hat{m}, \epsilon, t, m_0) > 0$, $\forall t \in (\hat{T}, \hat{T} + \delta)$. To show that this is impossible, it is enough to prove that

$$\frac{\mathrm{d}}{\mathrm{d}T} F(\hat{m}, \epsilon, \hat{T}, m_0) < 0. \tag{29}$$

To see this, it is convenient to perform the following change of variables: $u := \frac{m}{\epsilon} \in (0, 1)$ and $r := \epsilon T$, so that

$$F(m, \epsilon, T, m_0) = G(u, \epsilon, r, m_0) := \frac{1}{(1+r)^2} \varphi\left(\frac{r}{1+r} u, m_0\right) - \epsilon u, \tag{30}$$

where φ is given in (26). Note that (29) is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}r} G(\hat{u}, \epsilon, \hat{r}, m_0) < 0, \tag{31}$$

where $\hat{u} := \frac{\hat{m}}{\epsilon}$ and $\hat{r} := \epsilon \hat{T}$.

Claim $\hat{r} > \frac{1}{2}$.

Proof To see this, note that, being $\epsilon \leq m_0$,

$$G(u, \epsilon, r, m_0) \geq H(u, r, m_0) := \frac{1}{(1+r)^2} \varphi\left(\frac{r}{1+r} u, m_0\right) - m_0 u.$$

The claim follows if we show that, $\forall r \leq \frac{1}{2}$,

$$H(u, r, m_0) > 0, \quad \forall u \in (0, 1). \tag{32}$$

Since $H(u, r, m_0)$ is linear in m_0, it is enough to prove (32) for $m_0 \in \{0, 1\}$. For $m_0 = 0$ this is obvious, so we show it for $m_0 = 1$:

$$H(u, r, 1) = \frac{\left(1 + \frac{r}{1+r} u\right)^2}{(1+r)^2 \left(1 - \frac{r^2 u^2}{(1+r)^2}\right)^2} - u = \frac{(1-u)(r^2 u^2 - (r^2 + 2r)u + 1)}{(1 + r(1-u))^2},$$

which has the same sign as $p(u) := r^2 u^2 - (r^2 + 2r)u + 1)$. If $r \leq \frac{1}{2}$ (indeed here $r < 2$ would suffice), $p'(u) = 2r^2 u - r^2 - 2r < 0$, $\forall u \in (0, 1)$, so, $p(u) \geq p(1) = 1 - 2r \geq 0$, $\forall u \in (0, 1)$ and for $r \leq \frac{1}{2}$. This completes the proof of the Claim. □

We are now left with the proof of (31). Note that

$$\frac{\mathrm{d}}{\mathrm{d}r} G(u, \epsilon, r, m_0) = -\frac{2}{(1+r)^3} \varphi\left(\frac{r}{1+r} u, m_0\right) + \frac{u}{(1+r)^4} \varphi'\left(\frac{r}{1+r} u, m_0\right), \tag{33}$$

where $\varphi'(z, m_0) = \frac{\mathrm{d}}{\mathrm{d}z}\varphi(z, m_0)$. By (28),

$$G(\hat{u}, \epsilon, \hat{r}, m_0) = 0 \;\Rightarrow\; \varphi\left(\frac{\hat{r}}{1+\hat{r}}\hat{u}, m_0\right) = \epsilon\hat{u},$$

and

$$\frac{\mathrm{d}}{\mathrm{d}u}G(\hat{u}, \epsilon, \hat{r}, m_0) = 0 \;\Rightarrow\; \frac{r}{(1+r)^3}\varphi'\left(\frac{\hat{r}}{1+\hat{r}}\hat{u}, m_0\right) = \epsilon.$$

Inserting these identities in (33), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}r}G(\hat{u}, \epsilon, \hat{r}, m_0) = \frac{\epsilon\hat{u}}{1+\hat{r}}\left(-2 + \frac{1}{\hat{r}}\right) < 0, \quad \text{as } \hat{r} > \frac{1}{2}.$$

(ii) Note that $F(0, T, \epsilon, m_0) = \frac{m_0}{(\epsilon T+1)^2} > 0$. Moreover, $\lim_{T\to+\infty} F(\epsilon, T, \epsilon, m_0) = \frac{1+m_0}{2} - \epsilon \le 0$. Since, as shown in point (i), the map $T \mapsto F(\epsilon, T, \epsilon, m_0)$ is strictly increasing, then $F(\epsilon, T, \epsilon, m_0) < 0$, $\forall T > 0$. Thus, $m \mapsto F(m, T, \epsilon, m_0)$ has opposite sign at the endpoints of $(0, \epsilon)$: by convexity there is a unique solution $m = M(T, \epsilon, m_0)$ to $F(m, T, \epsilon, m_0) = 0$. The fact that $\lim_{T\to+\infty} M(T, \epsilon, m_0) = 0$ follows from the fact that $\lim_{T\to+\infty} F(0, T, \epsilon, m_0) = 0$ and that $\lim_{T\to+\infty} \frac{\mathrm{d}}{\mathrm{d}m} F(0, T, \epsilon, m_0) = -1$.

(iii) In this case, $F(0, T, \epsilon, m_0) = \frac{m_0}{(\epsilon T+1)^2} \ge 0$. Moreover, $F(\epsilon, T, \epsilon, m_0) < 0 \iff T < T^*(\epsilon, m_0)$, as already seen in the proof of Proposition 3.1, point (iii). By convexity of $m \mapsto F(m, T, \epsilon, m_0)$, for $T \le T^*(\epsilon, m_0)$ Eq. (24) has a unique solution in $[0, \epsilon)$. The fact that $m = 0$ is a solution if and only if $m_0 = 0$ is easily verified. Moreover, by (27), for T sufficiently large $F(m, T, \epsilon, m_0)$ attains negative values, it is positive at $m = 0$ and $m = \epsilon$, so (24) has two solutions in $[0, \epsilon)$. We need, however, a sharper analysis for $T > T^*(\epsilon, m_0)$. Note that, in this case $F(0, T, \epsilon, m_0) > 0$ and $F(\epsilon, T, \epsilon, m_0) > 0$. Thus, by convexity of $m \mapsto F(m, T, \epsilon, m_0)$, (24) has zero or two solutions, except for the "special times" T for which the graph of the map $m \mapsto F(m, T, \epsilon, m_0)$ is tangent to the horizontal axis. Note that these special times are identified as (T-component of the) solutions in $(0, \epsilon) \times (T^*(\epsilon, m_0), +\infty)$ of the system

$$\begin{cases} F(m, T, \epsilon, m_0) = 0 \\ \frac{\mathrm{d}}{\mathrm{d}m} F(m, T, \epsilon, m_0) = 0. \end{cases} \tag{34}$$

Note that (34) has no solutions with $T \le T^*(\epsilon, m_0)$, so we may look for solutions of (34) in $(0, \epsilon) \times (0, +\infty)$. The remaining part of the proof is based on the following Lemma.

Lemma A.1 *Let $\epsilon_*^{(2)}(m_0)$ and $\epsilon_*^{(3)}(m_0)$ be as in the statement of Proposition 3.3. Then,*

$$m_0 < \epsilon_*^{(2)}(m_0) < \epsilon_*^{(3)}(m_0) < \frac{1+m_0}{2},$$

(unless for $m_0 = 0$, where $0 = \epsilon_*^{(2)}(0) < \epsilon_*^{(3)}(0) < \frac{1}{2}$) such that

(a) for $m_0 < \epsilon \le \epsilon_*^{(2)}(m_0)$, (34) has a unique solution $(\tilde{m}(\epsilon, m_0), \tilde{T}(\epsilon, m_0))$;
(b) for $\epsilon_*^{(2)}(m_0) < \epsilon < \epsilon_*^{(3)}(m_0)$, (34) has two solutions $(\hat{m}(\epsilon, m_0), T_c^{(2)}(\epsilon, m_0))$ and $(\tilde{m}(\epsilon, m_0), T_c^{(3)}(\epsilon, m_0))$ with $T_c^{(2)}(\epsilon, m_0) < T_c^{(3)}(\epsilon, m_0)$;
(c) for $\epsilon_*^{(3)}(m_0) \le \epsilon < \frac{1+m_0}{2}$, (34) has no solutions.

The proof of this lemma is postponed after the end of this section. The desired result of the solutions of (24) readily follows from this Lemma. Indeed, in case (a), there is a unique special time $T_c^{(2)}(\epsilon, m_0)$. Since, for large T, (24) has two solutions, necessarily for $T^*(\epsilon, m_0) < T \le T_c^{(2)}(\epsilon, m_0)$ we must have $F(m, T, \epsilon, m_0) > 0$, $\forall m \in (0, \epsilon)$, so (24) has no solution.

In case (b), there are two special times $T_c^{(2)}(\epsilon, m_0) < T_c^{(3)}(\epsilon, m_0)$: the only possibility is that, for $T^*(\epsilon, m_0) < T < T_c^{(2)}(\epsilon, m_0)$, the graph of $m \mapsto F(m, T, \epsilon, m_0)$ crosses twice the horizontal axis (two solutions for (24)), for $T_c^{(2)}(\epsilon, m_0) < T < T_c^{(3)}(\epsilon, m_0)$ it stays above the horizontal axis (no solutions for (24)) and it crosses again twice the horizontal axis for $T > T_c^{(3)}(\epsilon, m_0)$. The proof is therefore completed. □

Proof of Lemma A.1

By the change of variables $u := \frac{m}{\epsilon} \in (0, 1)$ and $r := \epsilon T$ as in (30), we may replace $F(m, \epsilon, T, m_0)$ by

$$G(u, \epsilon, r, m_0) := \frac{1}{(1 + r + ru)^2}\left[m_0 + \frac{2(1 + m_0)(1 + r)ru}{(1 + r - ru)^2}\right] - \epsilon u = 0,$$

$u \in (0, 1)$, $r > 0$, and (34) is equivalent to

$$\begin{cases} G(u, \epsilon, r, m_0) = 0 \\ \frac{\mathrm{d}}{\mathrm{d}u} G(u, \epsilon, r, m_0) = 0. \end{cases} \tag{35}$$

Letting, as above, $U(u, \epsilon, r, m_0) := m_0(1 + r - ru)^2 + 2(1 + m_0)r(1 + r)u - \epsilon u\left[(1 + r)^2 - r^2u^2\right]^2 = 0$, we have that $G(u, \epsilon, r, m_0) = \frac{1}{\left[(1+r)^2 - r^2u^2\right]^2} U(u, \epsilon, r, m_0)$. It is immediately seen that (35) is equivalent to

$$\begin{cases} U(u, \epsilon, r, m_0) = 0 \\ \frac{\mathrm{d}}{\mathrm{d}u} U(u, \epsilon, r, m_0) = 0, \end{cases} \tag{36}$$

We use again the identity $U(u, \epsilon, r, m_0) = u\frac{\mathrm{d}}{\mathrm{d}u} U(u, \epsilon, r, m_0) + \left(m_0 - 4\epsilon r^2u^3\right)\left((1 + r)^2 - u^2r^2\right)$, which implies that (36) is equivalent to

$$\begin{cases} \left(m_0 - 4\epsilon r^2u^3\right) = 0 \\ \frac{\mathrm{d}}{\mathrm{d}u} U(u, \epsilon, r, m_0) = 0. \end{cases} \tag{37}$$

Summing up, the number of pairs $(m, T) \in (0, \epsilon) \times (0, +\infty)$ solving (34) equals the number of solutions r to the equation

$$W(r, \epsilon, m_0) := \frac{\mathrm{d}}{\mathrm{d}u} U\left(\left(\frac{m_0}{4\epsilon r^2}\right)^{1/3}, \epsilon, r, m_0\right) = 0 \tag{38}$$

such that $\frac{m_0}{4\epsilon r^2} < 1$, i.e., $r > \sqrt{\frac{m_0}{4\epsilon}}$. With the help of a symbolic calculator, we obtain

$$\begin{aligned} 8W(r, \epsilon, m_0) = r &\left\{16 + \left[16 + 3m_0\left(\frac{2m_0}{\epsilon r^2}\right)^{1/3}\right] r\right\} \\ &+4\epsilon(1+r)^2\left\{-2 - 4r + \left[-2 + 3\left(\frac{2m_0}{\epsilon r^2}\right)^{1/3}\right] r^2\right\}. \end{aligned}$$

Now, set $V(s, \epsilon, m_0) := W(\sqrt{s}, \epsilon, m_0)$. Thus, we are left with the problem of finding the solutions of

$$V(s, \epsilon, m_0) = 0 \tag{39}$$

with $s > \frac{m_0}{4\epsilon}$. This last change of variable is a trick to get the following claim.

Claim *The map $s \mapsto V(s, \epsilon, m_0)$ is strictly concave in $\left(\frac{m_0}{4\epsilon}, +\infty\right)$.*

Proof Using again a symbolic calculator we get, letting $k := \left(\frac{2m_0}{\epsilon}\right)^{1/3}$,

$$\begin{aligned} &-12s^{3/2}\frac{\mathrm{d}^2}{\mathrm{d}s^2}V(s, \epsilon, m_0) \\ &\quad = 6 + m_0 k s^{1/6} + \epsilon\left[-12 + 4k^2 s^{-1/6} + 36s + 5k^2 s^{1/3} + 24s^{3/2} - 8k^2 s^{5/6}\right]. \end{aligned}$$

To prove the claim, we need to show that this last expression is positive. Note that $s > \frac{m_0}{4\epsilon} = \left(\frac{k}{2}\right)^3$. Moreover, as $m_0 < \epsilon < \frac{1+m_0}{2}$, then $\frac{4m_0}{1+m_0} < k^3 < 2$. Again from $\epsilon < \frac{1+m_0}{2}$, the inequality $\frac{\mathrm{d}^2}{\mathrm{d}s^2}V(s, \epsilon, m_0) < 0$ follows if we show that

$$6 + m_0 k s^{1/6} + \frac{1+m_0}{2}\left[-12 + 4k^2 s^{-1/6} + 36s + 5k^2 s^{1/3} + 24s^{3/2} - 8k^2 s^{5/6}\right] > 0. \tag{40}$$

Using $s > \left(\frac{k}{2}\right)^3$ and $k^3 > \frac{4m_0}{1+m_0}$, we have that $\frac{1+m_0}{2}5k^2 s^{1/3} > \frac{1+m_0}{4}5k^3 > 5m_0$, so the l.h.s. of (40) is bounded from below by

$$m_0 k s^{1/6} - m_0 + (1+m_0)\left[2k^2 s^{-1/6} + 18s + 12s^{3/2} - 4k^2 s^{5/6}\right]. \tag{41}$$

Now, for $s \leq \frac{1}{4}$ the expression in (41) is bounded from below by

$$\begin{aligned} -m_0 + (1+m_0)\left[2k^2 s^{-1/6} - 4k^2 s^{5/6}\right] &= -m_0 + \frac{1+m_0}{s^{1/6}}\left(2k^2 - 4k^2 s\right) \\ &\geq -m_0 + (1+m_0)k^2 s^{-1/6} \\ &\geq -m_0 + (1+m_0)\left(\frac{4m_0}{1+m_0}\right)^{2/3} \\ &\geq -m_0 + (1+m_0)4^{2/3}\frac{m_0}{1+m_0} > 0. \end{aligned}$$

For $s > \frac{1}{4}$, the expression in (41) is bounded from below by

$$\begin{aligned} -m_0 + 18s + 12s^{3/2} - 4k^2 s^{5/6} &\geq -m_0 + s^{5/6}\left[18\left(\frac{1}{4}\right)^{1/6} + 12\left(\frac{1}{4}\right)^{2/3} - 4k^2\right] \\ &\geq -m_0 + s^{5/6}\left[\left(\frac{1}{4}\right)^{1/6} + 12\left(\frac{1}{4}\right)^{2/3} - 4\,2^{2/3}\right] \\ &\geq -m_0 + 8s^{5/6} \geq -m_0 + 2 > 0. \end{aligned}$$

This proves (40) and therefore the claim. □

Now, using concavity of $V(s,\epsilon,m_0)$ and the fact that $\lim_{s\to+\infty} V(s,\epsilon,m_0) = \lim_{r\to+\infty} W(r,\epsilon,m_0) = -\infty$, we have that (39) has a unique solution whenever $V\left(\frac{m_0}{4\epsilon},\epsilon,m_0\right) > 0$. We get

$$V\left(\frac{m_0}{4\epsilon},\epsilon,m_0\right) = \frac{m_0(1+m_0)}{2\epsilon} + (1+m_0)\sqrt{\frac{m_0}{\epsilon}} - \epsilon\left(1+e\sqrt{\frac{m_0}{\epsilon}}\right),$$

so

$$V\left(\frac{m_0}{4\epsilon},\epsilon,m_0\right)\Big|_{\epsilon=m_0} = \frac{3}{2}(1-m_0) > 0,$$

$$V\left(\frac{m_0}{4\epsilon},\epsilon,m_0\right)\Big|_{\epsilon=\frac{(1+m_0)}{2}} = -\frac{1}{2}(1-m_0) < 0 \tag{42}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}V\left(\frac{m_0}{4\epsilon},\epsilon,m_0\right) = -1 - \sqrt{\frac{m_0}{\epsilon}} - \frac{m_0(1+m_0)}{2\epsilon^2} - \frac{\sqrt{\frac{m_0}{\epsilon}}(1+m_0)}{2\epsilon} < 0.$$

Therefore, there is a unique $\epsilon_*^{(2)}(m_0)$, with $m_0 < \epsilon_*^{(2)}(m_0) < \frac{1+m_0}{2}$, such that $V\left(\frac{m_0}{4\epsilon_*^{(2)}},\epsilon_*^{(2)},m_0\right) = 0$. Moreover, for $m_0 < \epsilon < \epsilon_*^{(2)}(m_0)$, (39) has a unique solution and, since

$$\frac{\mathrm{d}}{\mathrm{d}s}V(s,\epsilon,m_0)\Big|_{s=\frac{m_0}{4\epsilon}} = \frac{2(1+m_0-2\epsilon)\left(1+\sqrt{\frac{m_0}{\alpha}}\right)}{\sqrt{\frac{m_0}{\alpha}}} > 0,$$

it follows that (39) has two solutions as ϵ crosses $\epsilon_*^{(2)}(m_0)$. This actually occurs until ϵ reaches $\epsilon_*^{(3)}(m_0)$, where $\epsilon_*^{(3)}(m_0)$ is characterized by the fact that the graph of $s \mapsto V(s, \epsilon_*^{(3)}(m_0), m_0)$ is tangent to the horizontal axis. This is a consequence of the following monotonicity property:

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}V(s,\epsilon,m_0) < 0, \quad \forall s \geq \frac{m_0}{4\epsilon}. \tag{43}$$

This suffices to characterize $\epsilon_*^{(3)}(m_0)$; to complete the proof, we need to show that $\epsilon_*^{(3)}(m_0) < \frac{1+m_0}{2}$, which is equivalent to

$$V\left(s, \frac{1+m_0}{2}, m_0\right) < 0, \quad \forall s \geq \frac{m_0}{2(1+m_0)}. \tag{44}$$

We are therefore left to prove (43) and (44). We begin with (43).

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\epsilon}V(s,\epsilon,m_0)\\ &= -\frac{1}{8\epsilon}\left[m_0\left(\frac{2m_0}{\epsilon}\right)^{1/3}s^{2/3} + 4\epsilon(1+\sqrt{s})^2\left(2+4\sqrt{s}+4s-\left(\frac{2m_0}{\epsilon}\right)^{2/3}s^{1/3}\right)\right].\end{aligned}$$

To show that this expression is negative, we just observe that, being $\frac{2m_0}{\epsilon} \leq 2$,

$$2+4\sqrt{s}+4s-\left(\frac{2m_0}{\epsilon}\right)^{2/3}s^{1/3} > 2+4\sqrt{s}+4s-2s^{1/3} > \begin{cases} 2-2s^{1/3} \geq 0 & \text{for } s \leq 1 \\ 4s-2s^{1/3} > 0 & \text{for } s > 1. \end{cases}$$

This establishes (43). Now, we show (44). We recall that $s \mapsto V\left(s, \frac{1+m_0}{2}, m_0\right)$ is strictly concave for $s \geq \frac{m_0}{2(1+m_0)}$. Moreover, again with the help of a symbolic calculator

$$\frac{\mathrm{d}}{\mathrm{d}s}V\left(s, \frac{1+m_0}{2}, m_0\right)\Big|_{s=\frac{m_0}{2(1+m_0)}} = 0.$$

So it is enough to show that

$$V\left(\frac{m_0}{2(1+m_0)}, \frac{1+m_0}{2}, m_0\right) < 0,$$

that has been seen already in (42). The proof is now complete. □

Proof of Proposition 3.4: Unpolarized Incoherent Equilibria ($-\epsilon \le m < 0$)

As done in Proposition 3.2, we use the symmetry (14). So we look for solutions $m \in (0, \epsilon]$ of the equation $F(m, \epsilon, T, -m_0) = 0$; this allows to reuse some of the ideas in Proposition 3.3.

(i) We employ here the change of variables seen in (30), so

$$\begin{aligned} F(m, \epsilon, T, -m_0) = G(u, \epsilon, r, -m_0) &:= \frac{1}{(1+r)^2}\varphi\left(\frac{r}{1+r}u, -m_0\right) - \epsilon u \\ &= \frac{2(1+r)ru - m_0\left[(1+r)^2 + r^2u^2\right]}{\left((1+r)^2 - r^2u^2\right)^2} - \epsilon u \\ &= \frac{1}{(1+r+ru)^2}\left[-m_0 + \frac{2(1-m_0)(1+r)ru}{(1+r-ru)^2}\right] - \epsilon u. \end{aligned}$$

We need to show that this last expression is strictly negative $\forall u \in (0, 1)$; it is enough to show this for $\epsilon = \frac{1-m_0}{2}$. This amounts to prove that

$$-m_0 + \frac{2(1-m_0)(1+r)ru}{(1+r-ru)^2} < \frac{1-m_0}{2}u(1+r+ru)^2,$$

which follows from

$$\begin{aligned} &\frac{2(1-m_0)(1+r)ru}{(1+r-ru)^2} \\ &< \frac{1-m_0}{2}u(1+r+ru)^2 \\ &\quad \Longleftrightarrow \quad 4r(1+r) < \left[(1+r)^2 - r^2u^2\right]^2, \end{aligned}$$

$\forall u \in (0, 1)$. This last inequality follows if we show it holds for $u = 1$, i.e., $4r(1+r) < (1+2r)^2$, that is clearly true $\forall r > 0$.

(ii) As seen in the proof of point (i) of Proposition 3.3, $F(\epsilon, \epsilon, T, -m_0)$ is strictly increasing in T. Moreover, the same computation done after (21) shows that $F(\epsilon, \epsilon, T, -m_0) > 0$ if and only if $T > T^*(\epsilon, -m_0)$. By the same change of variables used in point (i), (24) is equivalent to the equation

$$\frac{1}{(1+r+ru)^2}\left[-m_0 + \frac{2(1-m_0)(1+r)ru}{(1+r-ru)^2}\right] - \epsilon u = 0,$$

with $u \in (0, 1)$, which is also equivalent to

$$\begin{aligned} U(u, \epsilon, r, m_0) &:= -m_0(1+r-ru)^2 \\ &\quad +2(1-m_0)r(1+r)u - \epsilon u\left[(1+r)^2 - r^2u^2\right]^2 = 0. \end{aligned}$$

Our proof is based on the following claim.

Claim *Let* $u^* \in (0,1)$ *be such that* $U(u^*, \epsilon, r, m_0) = 0$. *Then,* $\frac{\mathrm{d}}{\mathrm{d}u} U(u^*, \epsilon, r, m_0) > 0$.

Proof We use the identity

$$\begin{aligned} U(u, \epsilon, r, m_0) &= u \frac{\mathrm{d}}{\mathrm{d}u} U(u, \epsilon, r, m_0) + \left(-m_0 - 4\epsilon r^2 u^3\right)\left((1+r)^2 - u^2 r^2\right) \\ &< u \frac{\mathrm{d}}{\mathrm{d}u} U(u, \epsilon, r, m_0) \end{aligned}$$

which proves the claim. □

This clearly implies that such u^* exists if and only if $U(1, \epsilon, r, m_0) > 0$ (as $U(0, \epsilon, r, m_0) = -m_0(1+r)^2 \leq 0$), i.e., if and only if $r > \epsilon T^*(\epsilon, -m_0)$ and in this case it is unique. This completes the proof. □

References

1. Bardi, M., Fischer, M.: On non-uniqueness and uniqueness of solutions in finite-horizon mean field games. ESAIM Control Optim. Calc. Var. **25**, 44 (2019). https://doi.org/10.1051/cocv/2018026
2. Bauso, D., Pesenti, R., Tolotti, M.: Opinion dynamics and stubbornness via multi-population mean-field games. J. Optim. Theory Appl. **170**(1), 266–293 (2016). https://doi.org/10.1007/s10957-016-0874-5
3. Bayraktar, E., Zhang, X.: On non-uniqueness in mean field games. Proc. Am. Math. Soc. **148**(9), 4091–4106 (2020). https://doi.org/10.1090/proc/15046
4. Blume, L., Durlauf, S.: Equilibrium concepts for social interaction models. Int. Game Theory Rev. **5**(03), 193–209 (2003). https://doi.org/10.1142/S021919890300101X
5. Brock, W., Durlauf, S.: Discrete choice with social interactions. Rev. Econ. Stud. **68**(2), 235–260 (2001). https://doi.org/10.1111/1467-937X.00168
6. Cardaliaguet, P.: Notes on mean field games. Université de Paris—Dauphine, Tech. rep. (2013)
7. Carmona, R., Delarue, F.: Probabilistic theory of mean field games with applications. In: Probability Theory and Stochastic Modelling, vol. 83–84. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-58920-6
8. Carmona, R., Graves, C.V.: Jet lag recovery: synchronization of circadian oscillators as a mean field game. Dyn. Games Appl. **10**(1), 79–99 (2020). https://doi.org/10.1007/s13235-019-00315-1
9. Cecchin, A., Fischer, M.: Probabilistic approach to finite state mean field games. Appl. Math. Optim. **81**(2), 253–300 (2020). https://doi.org/10.1007/s00245-018-9488-7
10. Cecchin, A., Pra, P.D., Fischer, M., Pelino, G.: On the convergence problem in mean field games: a two state model without uniqueness. SIAM J. Control Optim. **57**(4), 2443–2466 (2019). https://doi.org/10.1137/18M1222454
11. Dai Pra, P., Sartori, E., Tolotti, M.: Climb on the bandwagon: consensus and periodicity in a lifetime utility model with strategic interactions. Dyn. Games Appl. **9**(4), 1061–1075 (2019). https://doi.org/10.1007/s13235-019-00299-y
12. Delarue, F., Tchuendom, R.F.: Selection of equilibria in a linear quadratic mean-field game. Stoch. Process. Appl. **130**(2), 1000–1040 (2020). https://doi.org/10.1016/j.spa.2019.04.005
13. Dockner, E., Jergensen, S., Van Long, N., Sorger, G.: Differential Games in Economics and Management Science. Cambridge University Press, Cambridge (2000). https://doi.org/10.1017/CBO9780511805127
14. Gomes, D.A., Mohr, J., Souza, R.R.: Discrete time, finite state mean field games. J. de Mathématiques Pures et Appliquées **93**(3), 308–328 (2010). https://doi.org/10.1016/j.matpur.2009.10.010
15. Gomes, D.A., Mohr, J., Souza, R.R.: Continuous time finite state mean field games. Appl. Math. Optim. **68**(1), 99–143 (2013). https://doi.org/10.1007/s00245-013-9202-8
16. Gomes, D.A., Saúde, J.: Numerical methods for finite-state mean-field games satisfying a monotonicity condition. Appl. Math. Optim. **83**(1), 51–82 (2021). https://doi.org/10.1007/s00245-018-9510-0

17. Gomes, D.A., Velho, R.M., Wolfram, M.T.: Socio-economic applications of finite state mean field games. In: Philosophical Transactions of the Royal Society A, p. 372: 20130405. Royal Society (2014). https://doi.org/10.1098/rsta.2013.0405
18. Hajek, B., Livesay, M.: On non-unique solutions in mean field games. In: 2019 IEEE 58th Conference on Decision and Control (CDC), pp. 1219–1224. IEEE (2019). https://doi.org/10.1109/CDC40024.2019.9029906
19. Huang, M., Malhamé, R.P., Caines, P.E.: Large population stochastic dynamic games: closed-loop McKean–Vlasov systems and the Nash certainty equivalence principle. Commun. Inf. Syst. **6**(3), 221–252 (2006). https://doi.org/10.4310/CIS.2006.v6.n3.a5
20. Lasry, J.M., Lions, P.L.: Mean field games. Jpn. J. Math. **2**(1), 229–260 (2007). https://doi.org/10.1007/s11537-007-0657-8
21. Nutz, M., San Martin, J., Tan, X.: Convergence to the mean field game limit: a case study. Ann. Appl. Probab. **30**(1), 259–286 (2020). https://doi.org/10.1214/19-AAP1501
22. Yin, H., Mehta, P.G., Meyn, S.P., Shanbhag, U.V.: Synchronization of coupled oscillators is a game. IEEE Trans. Autom. Control **57**(4), 920–935 (2012). https://doi.org/10.1109/TAC.2011.2168082
23. Yin, H., Mehta, P.G., Meyn, S.P., Shanbhag, U.V.: On the efficiency of equilibria in mean-field oscillator games. Dyn. Games Appl. **4**(2), 177–207 (2014). https://doi.org/10.1007/s13235-013-0100-0