Entropy-driven dynamics and robust learning procedures in games

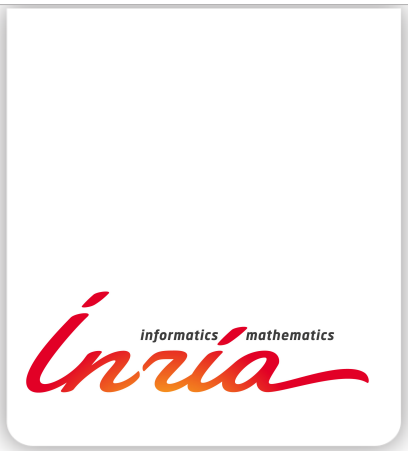
Pierre Coucheney, Bruno Gaujal, Panayotis Mertikopoulos

Entropy-driven dynamics and robust learning procedures in games

Pierre Coucheney, Bruno Gaujal, Panayotis Mertikopoulos

RESEARCH REPORT
N° 8210
Jan. 2013
Project-Team Mescal

ISSN 0249-6399 ISRN INRIA/RR--8210--FR+ENG



Entropy-driven dynamics and robust learning procedures in games

Pierre Coucheney*, Bruno Gaujal†, Panayotis Mertikopoulos ‡

Project-Team Mescal

Research Report n° 8210 — Jan. 2013 — 33 pages

Abstract: In this paper, we introduce a new class of game dynamics made of a pay-off replicator-like term modulated by an entropy barrier which keeps players away from the boundary of the strategy space. We show that these *entropy-driven* dynamics are equivalent to players computing a score as their on-going exponentially discounted cumulative payoff and then using a quantal choice model on the scores to pick an action. This dual perspective on *entropy-driven* dynamics helps us to extend the folk theorem on convergence to quantal response equilibria to this case, for potential games. It also provides the main ingredients to design a discrete time effective learning algorithm that is fully distributed and only requires partial information to converge to QRE. This convergence is resilient to stochastic perturbations and observation errors and does not require any synchronization between the players.

Key-words: Reinforcement Learning, Stochastic Approximation, Quantal Response Equilibria, Entropy

* PRiSM, University of Versailles, 45 avenue des Etats-Unis, 78035 Versailles, France
† Inria and University of Grenoble (LIG), 38330 Grenoble, France
‡ French National Center for Scientific Research (CNRS) and University of Grenoble (LIG), 38330 Grenoble, France

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Dynamiques de jeux avec un terme d'entropie et leurs propriétés de résilience

Résumé : Dans cet article, nous introduisons une nouvelle classe de dynamiques de jeux avec un terme de gain, de type réplication, modulé par une barrière entropique qui permet de maintenir les stratégies des joueurs loin des frontières du domaine. Nous montrons que ces dynamiques qui ont un terme d'entropie peuvent aussi être obtenues par des joueurs qui maintiennent un score sous la forme de leur gain cumulé actualisé et qui sélectionnent leurs actions sous la forme d'une réponse quantifiée à leur score courant. Cette double vision de la dynamique permet d'établir le théorème fondamental de convergence vers les points fixes de la réponse quantifié (qui sont proches des équilibres de Nash), dans le cas des jeux de potentiel. Elle permet aussi de mettre au point un algorithme discret effectif, complètement décentralisé et qui n'utilise que les données locales accessibles à chaque joueur, pour converger vers les points fixes de la dynamique. Cette convergence est conservée en présence de perturbations aléatoires et d'erreurs de mesure et ne nécessite pas de synchronisation entre les joueurs.

Mots-clés : Apprentissage par renforcement, approximation stochastique, équilibres de réponse quantitative, entropie

Contents

1 Introduction.

Owing to the computational complexity of Nash equilibria and related game-theoretic solution concepts, algorithms and processes for learning in games have received considerable attention over the last two decades. Such procedures can be divided into two broad categories, depending on whether they evolve in continuous or discrete time: the former class includes the numerous dynamics for learning and evolution (see e.g. Sandholm [28] for a recent survey), whereas the latter focuses on learning in infinitely iterated games, such as fictitious play and its variants – for an overview, see Fudenberg and Levine [11] and references therein.

A key challenge in these endeavors is that it is often unreasonable to assume that players are capable of monitoring the strategies of their opponents – or even of calculating the payoffs of actions that they did not play. As a result, much of the literature of learning in games revolves around payoff-based adaptive schemes which only require players to observe the stream of their *in-game* payoffs: for instance, in the framework of cumulative reinforcement learning, players use their observed payoff information to *score* their actions based on their estimated performance over time, and they then use a fixed decision model (such as logit choice) to determine their actions at the next instance of play. The convergence of such algorithms in 2-player games has been studied from a Q-learning perspective by Leslie and Collins [18] and Tuyls et al. [32] whereas, more recently, Cominetti et al. [10] and Bravo [9] took a moving-average approach for scoring actions in general N-player games and studied the long-term behavior of the resulting dynamics. Interestingly, in all these cases, when the learning process converges, it converges to a *perturbed* Nash equilibrium of the game – viz. a fixed point of a perturbed best-response correspondence (Fudenberg and Levine [11]).

Stochastic processes of this kind are usually analyzed with the ODE method of stochastic approximation which essentially compares the long-term behavior of the discrete-time process to the corresponding mean-field dynamics in continuous time – see e.g. the surveys by Benaïm [5] and Borkar [8]. Indeed, there are several sufficient conditions which guarantee that a discrete-time process and its continuous

counterpart both converge asymptotically to the same sets, so the continuous dynamics are usually derived as a limit of a discrete-time (and possibly random) process rooted in some adaptive learning scheme – cf. the aforementioned works by Leslie and Collins [18], Cominetti et al. [10], and Bravo [9].

Contrary to this approach, we proceed from the continuous to the discrete and develop two different learning processes from the same dynamical system (the actual algorithm depends crucially on whether we look at the evolution of the players' strategies or the performance scores of their actions). Accordingly, the first contribution of our paper is to derive a class of *entropy-driven* game dynamics which consist of a replicator-like term plus a barrier term that keeps players from approaching the boundary of the state space by imposing an entropic penalty to their payoffs – hence the dynamics' name. Interestingly, these dynamics are equivalent to players scoring their actions by taking an exponentially discounted (and continuously updated) aggregate of their payoffs and then using a quantal choice model to pick an action (McKelvey and Palfrey [21]); as such, entropy-driven dynamics constitute the strategy-space counterpart of the Q-learning dynamics of Leslie and Collins [18] – see also Tuyls et al. [32].

Another important feature of these dynamics is their *temperature*, a parameter which specifies the relative weight of the dynamics' entropic barrier term with respect to the game's payoffs – and also measures the weight that players attribute to past events, viz. the discount factor of their payoff aggregation scheme. These considerations allow us to derive a number of quite general results such as the dynamics' convergence to quantal response equilibria (QRE) in potential games and an extension of the well-known folk theorem of evolutionary game theory (Hofbauer and Sigmund [13]). In particular, we show that stability and convergence depend crucially on the temperature of the dynamics: at zero temperature, strict Nash equilibria are the only stable and attracting states of the dynamics, just as in the case of the replicator equation; for negative temperatures, all pure action profiles are attracting (but with vastly different basins of attraction), whereas for low positive temperatures, only QRE that are close to strict equilibria remain asymptotically stable.

The second important contribution of our paper concerns the practical implementation of entropy-driven game dynamics as learning algorithms with the following desirable properties:

1. The learning proceducre is *payoff-based*, *fully distributed* and *stateless* – players only need to observe their in-game payoffs and no knowledge of the game's structure or of the algorithm's state is required.

2. Payoffs may be subject to stochastic perturbations and observation errors; in fact, payoff observations need not even be up-to-date.

3. Updates need not be synchronized – there is no need for a global, centralized update clock used by all players.

These properties are key for the design of robust, decentralized optimization protocols in network and traffic engineering, but they also pose significant obstacles to convergence. Be that as it may, the convergence and boundary-avoidance properties of the continuous-time dynamics allow us to show that players converge to arbitrarily precise quantal approximations of strict Nash equilibria in potential games (Theorem 15 and Propositions 16, 17). Thus, thanks to the congestion characterization of such games (Monderer and Shapley [24]), we obtain a powerful distributed optimization method for a wide class of engineering problems, ranging from traffic routing (Altman et al. [1]) to wireless communications (Mertikopoulos et al. [22]).

1.1 Paper outline and structure.

After a few preliminaries in the rest of this section, our analysis proper begins in Section 2 where we introduce our cumulative reinforcement learning scheme and derive the associated entropy-driven game

dynamics. Owing to the duality between the evolution of the players' mixed strategies and the performance scores of their actions (measured by an exponentially discounted aggregate of past payofs), we obtain two equivalent formulations of the dynamics: the score-based integral equation (ERL) and the strategy-based dynamics (ED).

In Section **??**, we exploit this interplay to derive certain properties of the entropy-driven game dynamics, namely their convergence to perturbed equilibria in potential games (Proposition 7), and a variant of the folk theorem of evolutionary game theory (Theorem 10). Finally, Section 4 is devoted to the discretization of the dynamics (ERL) and (ED), yielding Algorithms 1 and 2 respectively. By using stochastic approximation techniques, we show that when the players' learning temperature is positive (corresponding to an exponential discount factor $\lambda < 1$), then the strategy-based algorithm converges almost surely to perturbed strict equilibria in potential games (Theorem 15), even in the presence of noise (Proposition 16) and/or update asynchronicities (Proposition 17).

1.2 Notational conventions.

If $\mathcal{S} = \{s_\alpha\}_{\alpha=0}^n$ is a finite set, the vector space spanned by $\mathcal{S}$ over $\mathbb{R}$ will be the set $\mathbb{R}^{\mathcal{S}}$ of all maps $x\colon \mathcal{S} \to \mathbb{R}$, $s \in \mathcal{S} \mapsto x_s \in \mathbb{R}$. The canonical basis $\{e_s\}_{s\in\mathcal{S}}$ of this space consists of the indicator functions $e_s\colon \mathcal{S} \to \mathbb{R}$ which take the value $e_s(s) = 1$ on s and vanish otherwise, so thanks to the natural identification $s \mapsto e_s$, we will make no distinction between $s \in \mathcal{S}$ and the corresponding basis vector e_s of $\mathbb{R}^{\mathcal{S}}$. Likewise, to avoid drowning in a morass of indices, we will frequently use the index α to refer interchangeably to either s_α or e_α (writing e.g. x_α instead of the more unwieldy x_{s_α}); in a similar vein, if $\{\mathcal{S}_k\}_{k\in\mathcal{K}}$ is a finite family of finite sets indexed by $k \in \mathcal{K}$, we will use the shorthands $(\alpha; \alpha_{-k})$ for the tuple $(\alpha_0, \dots, \alpha_{k-1}, \alpha, \alpha_{k+1}, \dots) \in \prod_k \mathcal{S}_k$ and $\sum_\alpha^k$ in place of $\sum_{\alpha\in\mathcal{S}_k}$.

We will also identify the set $\Delta(\mathcal{S})$ of probability measures on $\mathcal{S}$ with the unit n-dimensional simplex $\Delta(\mathcal{S}) \equiv \{x \in \mathbb{R}^{\mathcal{S}} : \sum_\alpha x_\alpha = 1 \text{ and } x_\alpha \geq 0\}$ of $\mathbb{R}^{\mathcal{S}}$. Since $\Delta(\mathcal{S})$ is a smooth submanifold-with-corners of $\mathbb{R}^{\mathcal{S}}$, by a smooth function on $\Delta(\mathcal{S})$ we will mean a C^∞ function in the smooth structure that $\Delta(\mathcal{S})$ inherits from $\mathbb{R}^{\mathcal{S}}$ (Lee [17]). Moreover, if $\mathcal{S}_0 \equiv \mathcal{S} \setminus \{s_0\}$, we will write $\operatorname{proj}_0(x) \equiv x_{-0} = (x_1, \dots, x_n)$ for the induced surjection $x \in \mathbb{R}^{\mathcal{S}} \mapsto x|_{\mathcal{S}_0} \in \mathbb{R}^{\mathcal{S}_0}$.

Regarding players and their actions, we will follow the original convention of Nash and employ Latin indices $(k, \ell, \dots)$ for players, while keeping Greek ones $(\alpha, \beta, \dots)$ for their actions (pure strategies); finally, unless otherwise mentioned, we will use $\alpha, \beta, \dots$, for indices that start at 0, and $\mu, \nu, \dots$, for those which start at 1.

1.3 Definitions from game theory.

A *finite game* $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ will be a tuple consisting of *a*) a finite set of *players* $\mathcal{N} = \{1, \dots, N\}$; *b*) a finite set $\mathcal{A}_k$ of *actions* (or *pure strategies*) for each player $k \in \mathcal{N}$; and *c*) the players' *payoff functions* $u_k\colon \mathcal{A} \to \mathbb{R}$, where $\mathcal{A} \equiv \prod_k \mathcal{A}_k$ denotes the game's *action space*, i.e. the set of all *action profiles* $(\alpha_1, \dots, \alpha_N)$, $\alpha_k \in \mathcal{A}_k$. More succinctly, if $\mathcal{A}^* \equiv \coprod_k \mathcal{A}_k = \{(\alpha, k) : \alpha \in \mathcal{A}_k\}$ is the disjoint union of the players' action sets, then the *payoff map* of $\mathfrak{G}$ will be the map $u\colon \mathcal{A} \to \mathbb{R}^{\mathcal{A}^*} \cong \prod_k \mathbb{R}^{\mathcal{A}_k}$ which sends the profile $(\alpha_1, \dots, \alpha_N) \in \mathcal{A}$ to the payoff vector $(u_k(\alpha; \alpha_{-k}))_{\alpha\in\mathcal{A}_k, k\in\mathcal{N}} \in \prod_k \mathbb{R}^{\mathcal{A}_k}$. A *restriction* of $\mathfrak{G}$ will then be a game $\mathfrak{G}' \equiv \mathfrak{G}'(\mathcal{N}, \mathcal{A}', u')$ with the same players as $\mathfrak{G}$, each with a subset $\mathcal{A}'_k \subseteq \mathcal{A}_k$ of their original actions, and with payoff functions $u'_k \equiv u_k|_{\mathcal{A}'}$ suitably restricted to the reduced action space $\mathcal{A}' = \prod_k \mathcal{A}'_k$.

Of course, players can mix their actions by taking probability distributions $x_k = (x_{k\alpha})_{\alpha\in\mathcal{A}_k} \in \Delta(\mathcal{A}_k)$ over their action sets $\mathcal{A}_k$. In that case, their expected payoffs will be

$$u_k(x) = \sum\nolimits_{\alpha_1}^1 \cdots \sum\nolimits_{\alpha_N}^N u_k(\alpha_1, \dots, \alpha_N)\, x_{1,\alpha_1} \cdots x_{N,\alpha_N}, \tag{1}$$

where $x = (x_1, \dots, x_N)$ denotes the players' (mixed) *strategy profile* and $u_k(\alpha_1, \dots, \alpha_N)$ is the payoff to player k in the (pure) action profile $(\alpha_1, \dots, \alpha_N) \in \mathcal{A}$;[1] more explicitly, if player k plays the pure strategy $\alpha \in \mathcal{A}_k$, we will use the notation $u_{k\alpha}(x) \equiv u_k(\alpha; x_{-k}) = u_k(x_1, \dots, \alpha, \dots, x_N)$. In this mixed context, the *strategy space* of player k will be the simplex $X_k \equiv \Delta(\mathcal{A}_k)$ while the strategy space of the game will be the convex polytope $X \equiv \prod_k X_k$. Together with the players' (expected) payoff functions $u_k : X \to \mathbb{R}$, the tuple $(\mathcal{N}, X, u)$ will be called the *mixed extension* of $\mathfrak{G}$ and it will also be denoted by $\mathfrak{G}$ (relying on context to resolve any ambiguities).

The most prominent solution concept in game theory is that of Nash equilibrium (NE) which characterizes profiles that are resilient against unilateral deviations. We will thus say that $q \in X$ is a *Nash equilibrium* of $\mathfrak{G}$ when

$$u_k(x_k; q_{-k}) \leq u_k(q) \quad \text{for all } x_k \in X_k \text{ and for all } k \in \mathcal{N}. \tag{NE}$$

In particular, if (NE) is strict for all $x_k \in X_k \setminus \{q_k\}$, $k \in \mathcal{N}$, then q will be called a *strict* Nash equilibrium.

An especially relevant class of finite games is obtained when the players' payoff functions satisfy the *potential property* :

$$u_{k\alpha}(x) - u_{k\beta}(x) = U(\alpha; x_{-k}) - U(\beta; x_{-k}) \tag{2}$$

for some (necessarily) multilinear function $U : X \to \mathbb{R}$. When this is the case, the game will be called a *potential game with potential function* U, and as is well known, the pure Nash equilibria of $\mathfrak{G}$ will be precisely the vertices of X that are local maximizers of U (Monderer and Shapley [24]).

2 Reinforcement learning and entropy-driven dynamics.

In this section, our aim will be to derive a class of continuous-time learning dynamics based on the following cumulative reinforcement premise: agents accumulate a long-term "performance score" for each of their actions and they then use a smooth choice function to map these scores to strategies and continue playing. More precisely:

1. The *assessment phase* (Section 2.2) will comprise the scheme with which players aggregate past payoff information in order to update their actions' performance scores.

2. The *choice phase* (Section 2.3) will then describe how these scores are used to select a mixed strategy.

For simplicity, we will first derive the dynamics that correspond to this learning process in the case of a single player whose payoffs are determined at each instance of play by Nature—the case of several players involved in a finite game will be entirely similar. Furthermore, the passage from discrete to continuous time will be done here at a heuristic level and we will assume that players have perfect payoff information, that is: *a*) they are assumed able to observe or otherwise calculate the payoffs of all their actions; and *b*) unless mentioned otherwise, this payoff information will be assumed accurate and not subject to measurement errors or other exogenous perturbations. The precise interplay between discrete and continuous time and the effect of imperfect information and stochastic fluctuations will be explored in detail in Section 4.

[1] Recall that we will be using α for both elements $\alpha \in \mathcal{A}_k$ and basis vectors $e_\alpha \in \Delta(\mathcal{A}_k)$, so there is no clash of notation between payoffs to pure and mixed strategies.

2.1 The model.

In view of the above, our single-player learning model will be as follows: at time t, an agent makes a discrete choice from the elements of a finite set $\mathcal{A} = \{\alpha_0, \alpha_1, \dots, \alpha_n\}$ (representing e.g. the routes of a traffic network, different stock options, etc.). We will denote the payoff to $\alpha \in \mathcal{A}$ at time t by $u_\alpha(t)$, and the agent's assessment of his actions' performance up to instance t will be represented by the *score variables* $y_\alpha(t) \in \mathbb{R}$, $\alpha \in \mathcal{A}$. In this context, the assessment phase will describe how $y_\alpha(t)$ is updated using the payoffs $u_\alpha(s)$, $s \leq t$, of all past instances of play, whereas the choice stage will specify the choice map $Q\colon \mathbb{R}^{\mathcal{A}} \to X \equiv \Delta(\mathcal{A})$ which prescribes the agent's mixed strategy $x \in X$ given his assessment of each action $\alpha \in \mathcal{A}$ so far.

2.2 The assessment stage: memory and aggregation of past information.

Assuming for the moment that the agent plays at discrete time intervals $s = 0, 1, \dots, t$, the class of assessment schemes that we will consider is the familiar and widely used exponential model of long-term performance evaluation

$$y_\alpha(t) = \sum\nolimits_{s=0}^{t} \lambda^{t-s} u_\alpha(s), \tag{3}$$

where $\lambda \in (0, +\infty)$ is the model's discounting parameter, $u_\alpha(t)$ is the sequence of payoffs corresponding to $\alpha \in \mathcal{A}$, and we are assuming for the moment that the model is initially *unbiased*, i.e. $y_\alpha(0) = 0$. Clearly:

1. For $\lambda \in (0, 1)$ we get the standard exponential discounting model which assigns exponentially more weight to recent observations.

2. If $\lambda = 1$ we get the unweighted aggregation scheme $y_\alpha(t) = \sum_{s=1}^{t} u_\alpha(t)$ which has been examined in the context of learning by Rustichini [27], Hofbauer et al. [14], Sorin [30], Mertikopoulos and Moustakas [23] and many others.

3. For $\lambda > 1$, the scheme (3) assigns exponentially more weight to past instances; as such, this case has attracted very little interest in the literature (after all, it seems rather counter-intuitive to discount current events in favor of a possibly irrelevant past). Nevertheless, we will see that the choice $\lambda > 1$ leads to some very surprising advantages, so we will not exclude this parameter range from our analysis.

Now, if the agent plays at discrete time intervals $0, h, \dots, nh \equiv t$ with time step $h > 0$, the exponential model (3) should be replaced by the scale-invariant version:

$$y_\alpha(t) = \sum\nolimits_{s \in [0:t:h]} \lambda^{t-s} u_\alpha(s) h, \tag{3′}$$

where the factor h has been included to make the sum (3′) intensive in h,[2] the notation $[0 : t : h]$ represents the index set $\{0, h, 2h, \dots, nh = t\}$, and we plead guilty to a slight abuse of notation for not differentiating between s and s/h in the argument of u_α (and between n and $t = nh$ in the case of y_α). Accordingly, the assessment scheme (3′) yields the recursive updating rule:

$$y_\alpha(t) = u_\alpha(t)h + \lambda^h y_\alpha(t - h), \tag{4}$$

which in turn shows that the updating of (3′) does not require the agent to have infinite memory: scores are simply updated by adding the payoffs obtained over a period h to the scores of the previous period scaled by the discount (or reinforcement) factor λ^h.

[2] Note that the sum (3′) consists of $\mathcal{O}(1/h)$ terms that are $\mathcal{O}(1)$ in h so it would scale extensively with h^{-1} if not scaled by h.

In this way, letting $h \to 0^+$ (and assuming Lipschitz-continuous payoff processes $u_\alpha(t)$ for simplicity), we readily obtain the continuous-time model

$$\dot{y}_\alpha(t) = u_\alpha(t) - T y_\alpha(t), \tag{5}$$

or, in integral form:

$$y_\alpha(t) = y_\alpha(0)e^{-Tt} + \int_0^t e^{-T(t-s)} u_\alpha(s)\, ds, \tag{5'}$$

where

$$T \equiv \log(1/\lambda) \tag{6}$$

represents the *temperature* of our performance assessment scheme (see the following sections for a justification of this terminology) and the term $y_\alpha(0)e^{-Tt}$ reflects the initial bias $y_\alpha(0)$ of the agent (initially taken equal to 0).[3] In tune with our previous discussion, the standard exponential discounting regime $\lambda \in (0,1)$ will thus correspond to positive temperatures $T > 0$, unweighted aggregation will be obtained for $T \to 0$, and exponential reinforcing of past observations will be recovered for negative learning temperatures $T < 0$.

Remark 1. In our context, the scheme (5) emerges quite simply as the differential form of an exponetially discounted model for aggregating past payoffs. It is thus interesting to note that Leslie and Collins [18] and Tuyls et al. [32] obtained the dynamics (5) for $T = 1$ from a quite different viewpoint, namely as the continuous-time limit of the Q-learning estimator

$$y_\alpha(t+1) = y_\alpha(t) + \gamma(t+1)\,(u_\alpha(t) - y_\alpha(t)) \times \frac{\mathbb{1}(\alpha(t) = \alpha)}{\mathbb{P}(\alpha(t) = \alpha)}, \tag{7}$$

where $\mathbb{1}$ and $\mathbb{P}$ denote respectively the indicator and probability of having chosen α at time t, and $\gamma(t)$ is an $(\ell^2 - \ell^1)$-summable series of time steps (see also Fudenberg and Levine [11]). The exact interplay between (5) and (7) will be explored in detail in Section 4; for now we simply note that (5) can be interpreted both a model of discounting past information and also as a moving Q-average.

While on this point, we should also highlight the relation between (7) and the moving average estimator of Cominetti et al. [10] which omits the factor $\mathbb{P}(\alpha(t) = \alpha)$ (or the similar estimator of Bravo [9] which has a state-dependent step size). As a result of this difference, the mean-field dynamics of [10] are scaled by the player's mixed strategy $x_\alpha(t) = \mathbb{P}(\alpha(t) = \alpha)$, leading to the adjusted dynamics $\dot{y}_\alpha = x_\alpha(t)\,(u_\alpha(t) - y_\alpha(t))$. Given this difference in form, there will basically be no overlap between our results and those of Cominetti et al. [10], but we will endeavor to draw analogies with their results wherever possible.

2.3 The choice stage: smoothed best responses and entropy functions.

Having established the way that an agent updates his assessment vector $y \in \mathbb{R}^{\mathcal{A}}$ over time, we now turn to mapping these scores to mixed strategies $x \in X \equiv \Delta(\mathcal{A})$ in a smooth fashion. In the theory of discrete choice, this boils down to a smooth best response problem so our aim will be to give a brief overview of the related constructions suitably adapted to our purposes; for a more comprehensive account, see McFadden [20], Anderson et al. [3], or Chapter 5 in Sandholm [28].

To motivate our approach, observe that a natural choice for the agent would be to always pick the action $\alpha \in \mathcal{A}$ with the highest score; however, this "best response" approach carries several problems: First, if two scores y_α and y_β happen to be equal (e.g. if there are payoff ties), then this mapping becomes a set-valued correspondence which requires a tie-breaking rule to be resolved (and is theoretically quite

[3] Note that the difference/differential equations (4)/(5) imply that initial scores decay (or grow) exponentially with time in the absence of external forcing, commensurately to the first payoff observation $u_\alpha(0)$.

cumbersome to boot). Additionally, such a practice could lead to completely discontinuous trajectories of play in continuous time—for instance, when the payoffs $u_\alpha(t)$ are driven by an additive white Gaussian noise process, as is commonly the case in information-theoretic applications of game theory; see e.g. Altman et al. [1]. Finally, since best responding generically results in picking pure strategies, such a process precludes convergence of strategies to non-pure equilibria in finite games.

In view of the above, a common alternative to the "best response" choice $x = \arg\max_{\alpha\in\mathcal{A}}\{y_\alpha\}$ is to smooth things out using the Gibbs map $G\colon \mathbb{R}^{\mathcal{A}} \to X$ defined as[4]

$$G_\alpha(y) = \frac{\exp(y_\alpha)}{\sum_\beta \exp(y_\beta)}, \quad \alpha \in \mathcal{A} \tag{8}$$

(see e.g. Cominetti et al. [10], Fudenberg and Levine [11], Hofbauer et al. [14], Leslie and Collins [18], Marsili et al. [19], Mertikopoulos and Moustakas [23], Rustichini [27], Sorin [30] and many others for uses of this choice map in game-theoretic learning). Indeed, it is well-known that $G(y)$ is the unique solution of the (strictly concave) maximization problem

$$\begin{aligned} &\text{maximize} && \textstyle\sum_\beta x_\beta y_\beta - g(x), \\ &\text{subject to} && x_\alpha \geq 0,\ \textstyle\sum_\beta x_\beta = 1, \end{aligned} \tag{9}$$

where the Boltzmann-Gibbs entropy $g(x) = \sum_\beta x_\beta \log x_\beta$ acts as a control cost adjustment to the agent's average score $\overline{y} = \sum_\beta x_\beta y_\beta$ (Fudenberg and Levine [11], van Damme [33]). In this way, $G(y)$ can be viewed as a *smoothed best response*: if the control cost is scaled down by some small $\varepsilon > 0$ (i.e. $g(x)$ is replaced by $\varepsilon g(x)$ in (9)), then the resulting solution $x^\varepsilon = G(\varepsilon^{-1}y)$ of (9) represents a smooth approximation to the best response correspondence $y \mapsto \arg\max_{\alpha\in\mathcal{A}}\{y_\alpha\}$ as $\varepsilon \to 0$.

Interestingly, the Gibbs map can also be seen as a special case of a *quantal response function* in the sense of McKelvey and Palfrey [21]—or a *perturbed best response* in the language of Hofbauer and Sandholm [12]. To wit, assume that the agents' scores are subject to additive stochastic fluctuations of the form

$$\tilde{y}_\alpha = y_\alpha + \xi_\alpha, \tag{10}$$

where the ξ_α are independently Gumbel-distributed random variables with zero mean and scale parameter $\varepsilon > 0$ (amounting to a variance of $\varepsilon^2\pi^2/6$). Then, the *choice probability* $P_\alpha(y)$ of $\alpha \in \mathcal{A}$ (defined as the probability that $\alpha \in \mathcal{A}$ maximizes the perturbed score variable $\tilde{y}_\alpha$) will simply be

$$P_\alpha(y) \equiv \mathbb{P}\left(\tilde{y}_\alpha = \max_\beta \tilde{y}_\beta\right) = G_\alpha(\varepsilon^{-1}y). \tag{11}$$

As a result, ordinary best responses are again recovered in the limit where the magnitude of the perturbations (measured by the scale parameter $\varepsilon > 0$ of the Gumbel distribution) approaches 0.

More generally, assume that the perturbations ξ_α are not Gumbel-distributed but instead follow an arbitrary probability law with a strictly positive and smooth density function. In this context, Hofbauer and Sandholm [12] showed that the resulting choice probability vector $Q(y) \in \Delta(\mathcal{A})$ solves the associated entropy maximization problem

$$\begin{aligned} &\text{maximize} && \textstyle\sum_\beta x_\beta y_\beta - h(x), \\ &\text{subject to} && x_\alpha \geq 0,\ \textstyle\sum_\beta x_\beta = 1, \end{aligned} \tag{EP}$$

where the *deterministic representation* h of ξ h is a smooth, strictly convex function on $X \equiv \Delta(\mathcal{A})$. Specifically, the choice probabilities $Q_\alpha(y) \equiv \mathbb{P}(\tilde{y}_\alpha = \max_\beta \tilde{y}_\beta)$ determine a strictly convex potential

[4] Due to the entrenched terminology for the logit choice model, many authors call (8) the "logit" map. However, (8) actually describes the *inverse* logit (or *logistic*) distribution, so, to avoid inconsistencies, we will refer to (8) by the name of its originator.

function $h^*\colon \mathbb{R}^{\mathcal{A}} \to \mathbb{R}$ such that $Q_\alpha(y) = \frac{\partial h^*}{\partial y_\alpha}$, so h^* will be related to the deterministic representation h of ξ via the *Legendre-Fenchel transformation* (Rockafellar [26]):

$$h^*(y) = \max_{x\in X} \left\{\sum_\beta x_\beta y_\beta - h(x)\right\}, \quad y \in \mathbb{R}^{\mathcal{A}}. \tag{L-F}$$

On account of the above, the choice map Q can be viewed either as a quantal response function to some perturbation process ξ, or as a smooth approximation to $\arg\max_\alpha y_\alpha$ with respect to an admissible control cost adjustment h (if we take the strictly concave problem (EP) as our starting point).[5] Formally, we have:

Definition 1. A function $h\colon X \to \mathbb{R} \cup \{+\infty\}$ will be called a *generalized entropy function* when:

1. h is convex and finite almost everywhere, except possibly on the relative boundary $\mathrm{bd}(X)$ of X.
2. h is smooth on $\mathrm{rel\,int}(X)$ and $|dh(x)| \to \infty$ when x converges to $\mathrm{bd}(X)$.
3. The Hessian tensor $\mathrm{Hess}(h)$ of h is positive-definite on $\mathrm{rel\,int}(X)$.

The Legendre-Fenchel conjugate h^* of h as defined by (L-F) will be called the *free entropy* of h, and the map $Q\colon \mathbb{R}^{\mathcal{A}} \to X$, $y \mapsto Q(y) \equiv \arg\max_{x\in X}\{\sum_\beta x_\beta y_\beta - h(x)\}$, will be the *choice map associated to* h. Finally, a generalized entropy function $h\colon X \to \mathbb{R}$ will be called *regular* when *a*) its restriction to any subface X' of X is itself an entropy function, and *b*) the ratio $h'(q+vt)/h''(q+vt)$ vanishes as $q+vt$ approaches $\mathrm{bd}(X')$ for all interior points $q \in \mathrm{rel\,int}(X')$ and for all tangent vectors $v \in T_q X'$.

The fact that the choice map Q of an entropy functional is well-defined and single-valued is an easy consequence of the convexity and boundary behavior of h; the smoothness of Q then follows from the implicit function theorem. Thus, given that (EP) allows us to view $Q(\varepsilon^{-1}y)$ as a smooth approximation to $\arg\max_\alpha y_\alpha$ for $\varepsilon \to 0^+$, the class of choice maps that we will consider will be precisely the maps that are derived from entropy functionals in the sense of Definition 1. A few remarks are thus in order:

Remark 1. In statistical mechanics and information theory, entropy functions are *concave*, so Definition 1 actually describes *negative* entropies. Besides notational convenience, one of the main reasons for this change of sign is that entropy functions as defined above are *essentially smooth functions of Legendre type*, with the added non-degeneracy condition of having a strictly definite Hessian (Rockafellar [26]). In fact, Definition 1 with our chosen sign convention bears very close ties to the class of *Bregman functions*, a key tool in interior point and proximal methods in optimization; for a more detailed account, see e.g. Rockafellar [26], Auslender et al. [4], Alvarez et al. [2] and references therein.

Remark 2. Examples of entropy functionals abound; some of the most prominent ones are:

1. The Boltzmann-Gibbs entropy: $h(x) = \sum_\beta x_\beta \log x_\beta$. (12a)
2. The log-entropy: $h(x) = -\sum_\beta \log x_\beta$. (12b)
3. The Tsallis entropy:[6] $h(x) = (1-q)^{-1}\sum_\beta (x_\beta - x_\beta^q)$, $0 < q \le 1$. (12c)
4. The Rényi entropy:[6] $h(x) = (q-1)^{-1}\log\sum_\beta x_\beta^q$, $0 < q \le 1$. (12d)

Except for the Rényi entropy, all of the above examples can be written in the convenient form $h(x) = \sum_\beta \theta(x_\beta)$ for some function $\theta\colon [0,+\infty) \to \mathbb{R}\cup\{+\infty\}$ with the properties:

[5] These two viewpoints are not equivalent because there exist cost functions h that do not arise as deterministic representations of perturbation processes ξ—see Hofbauer and Sandholm [12] for a counterexample based on the log-entropy $h(x) = -\sum_\beta \log x_\beta$.

[6] The Tsallis and Rényi entropies are not well-defined for $q = 1$, but they both approach the standard Gibbs entropy as $q \to 1$, so we will use the definition (12a) for $q = 1$ in (12c–12d).

1. θ is finite and smooth everywhere except possibly at 0.

2. $\theta'(x) \to -\infty$ as $x \to 0^+$ and $\theta''(x) > 0$ for all $x > 0$.

When h can be decomposed in this way, we will follow Alvarez et al. [2] and say that *h is decomposable with Legendre kernel θ*.

Remark 3. The regularity requirement of Definition 1 is just a safety net to ensure that h behaves well with respect to restrictions and does not exhibit any pathologies near $\mathrm{bd}(X)$. Of the examples (12) only the log-entropy (12b) is not regular because it is identically equal to $+\infty$ on every proper subface of X.

Remark 4. Another technical point that underlies Definition 1 is that we are implicitly assuming that h is defined on an open neighborhood of X in $\mathbb{R}^{\mathcal{A}}_+$ so that the derivatives of h are well-defined. The reason that we are not making this assumption explicit is that it may be done away with as follows: Let $\mathcal{A}_0 \equiv \mathcal{A} \setminus \{\alpha_0\}$ and consider the canonical projection $\mathrm{proj}_0 \colon \mathbb{R}^{\mathcal{A}} \to \mathbb{R}^{\mathcal{A}_0}$ defined in components as $\mathrm{proj}_0(x_0, x_1, \dots, x_n) \equiv x_{-0} = (x_1, \dots, x_n)$. Then, the image $X_0 \equiv \{w \in \mathbb{R}^{\mathcal{A}_0} : w_\mu \geq 0 \text{ and } \sum_\mu w_\mu \leq 1\}$ of X under proj_0 will be homeomorphic to X and the inverse to proj_0 on X_0 will be the injective immersion $\iota_0 \colon \mathbb{R}^{\mathcal{A}_0} \to \mathbb{R}^{\mathcal{A}}$ with $\iota_0(w_1, \dots, w_n) = (1 - \sum_\mu w_\mu, w_1, \dots, w_n)$. In view of the above, the directional derivatives of h on X may be defined by means of the pullback $h_0 \equiv \iota_0^* h = h \circ \iota_0$ as $\frac{\partial h}{\partial w_\mu} \equiv \frac{\partial h_0}{\partial w_\mu}$ (and similarly for the Hessian of h).

As one would expect, if h is defined on an open neighborhood of X, we will have $\frac{\partial h}{\partial w_\mu} = \frac{\partial h}{\partial x_\mu} - \frac{\partial h}{\partial x_0}$, so the above discussion reduces to treating $x_0 = 1 - \sum_\mu w_\mu$ as a dependent variable.Conversely, any smooth function $h \colon X \to \mathbb{R}$ can be extended smoothly to all of $\mathbb{R}^{\mathcal{A}}_+$ (e.g. via mollification), in which case it is easy to see that the directional derivatives $\frac{\partial h}{\partial x_\mu} - \frac{\partial h}{\partial x_0}$ are independent of the extension and the equality $\frac{\partial h}{\partial w_\mu} = \frac{\partial h}{\partial x_\mu} - \frac{\partial h}{\partial x_0}$ still holds on X. Consequently, we lose no generality in assuming that h is in fact defined on an open neighborhood of X in $\mathbb{R}^{\mathcal{A}}_+$, and we will do so throughout the rest of the paper unless explicitly stated otherwise. However, the "reduced" coordinates $w = \mathrm{proj}_0(x)$ and the associated derivations $\frac{\partial}{\partial w_\mu}$ will be very important in our calculations, so their introduction above is not just a technical triviality.

2.4 Entropy-driven learning dynamics.

Combining the results of the previous two sections on how to assess the long-term performance of an action and how to translate these assessments into strategies, we obtain the general class of *entropy-driven learning processes*:

$$\begin{aligned} y_\alpha(t) &= y_\alpha(0)\, e^{-Tt} + \int_0^t e^{-T(t-s)} u_\alpha(s)\, ds, \\ x(t) &= Q(y(t)), \end{aligned} \tag{ERL}$$

where Q is the choice map of the driving entropy $h \colon X \to \mathbb{R}$ and T is the model's learning temperature.

From an implementation perspective, the difficulty with (ERL) is twofold: First, for a given entropy function h, it is not always practical to write the choice function Q in a closed-form expression that the agent can use to update his strategies.[7] Furthermore, even when this is possible, (ERL) is a two-step computationally intensive process which does not allow the agent to update his strategies directly. The rest of this section will thus be devoted to writing (ERL) as a continuous-time dynamical system on X that can be updated with minimal computation overhead.

To that end, it will be convenient to work with a modified set of variables which measure the long-term difference in performance between an action $\mu \in \mathcal{A}$ and a "flagged" benchmark action $\alpha_0 \in \mathcal{A}$.

[7] The case of the Boltzmann-Gibbs entropy is a shining (but, ultimately, misleading) exception to the norm.

Formally, letting $\mathcal{A}_0 = \mathcal{A} \setminus \{\alpha_0\}$ as in Remark 4 above, the *relative score* of an action $\mu \in \mathcal{A}_0$ will be the difference

$$z_\mu = y_\mu - y_0, \tag{13}$$

or, more concisely, $z = \pi_0(y)$ where $\pi_0 \colon \mathbb{R}^{\mathcal{A}} \to \mathbb{R}^{\mathcal{A}_0}$ is the submersion

$$\pi_0(y_0, y_1, \dots, y_n) = (y_1 - y_0, \dots, y_n - y_0) = (z_1, \dots, z_n). \tag{14}$$

Thereby, the evolution of z over time will be

$$\dot{z}_\mu = \dot{y}_\mu - \dot{y}_0 = \Delta u_\mu - T z_\mu, \tag{ZD}$$

where now Δu_μ denotes the associated payoff difference $\Delta u_\mu \equiv u_\mu - u_0$, $\mu \in \mathcal{A}_0$.

The main advantage of introducing the variables z is that even though the choice map $Q \colon \mathbb{R}^{\mathcal{A}} \to X$ is not injective (and thus does not admit an inverse),[8] there exists a smooth embedding $Q_0 \colon \mathbb{R}^{\mathcal{A}_0} \to X_0 = \operatorname{proj}_0(X) = \{x \in \mathbb{R}^{\mathcal{A}_0} : x_\mu \geq 0, \sum_\mu x_\mu \leq 1\}$ such that the following diagram commutes:

$$\begin{array}{ccc} \mathbb{R}^{\mathcal{A}} & \xrightarrow{Q} & X \\ {\scriptstyle \pi_0}\downarrow & & \uparrow{\scriptstyle \iota_0} \\ \mathbb{R}^{\mathcal{A}_0} & \xrightarrow{Q_0} & X_0 \end{array} \qquad \operatorname{proj}_0 \tag{15}$$

With such an embedding at our disposal, we will then be able to translate the dynamics of $z \in \mathbb{R}^{\mathcal{A}_0}$ to X_0, and thence to X via the inverse ι_0 of proj_0 on X: $\iota_0(w_1, \dots, w_n) = (1 - \sum_\mu w_\mu, w_1, \dots, w_n)$.

To construct Q_0 itself, note that (EP) may be rewritten in terms of $z_\mu = y_\mu - y_0$ as:

$$\begin{aligned} &\text{maximize} && \textstyle\sum_\mu w_\mu z_\mu - h_0(w_1, \dots, w_n), \\ &\text{subject to} && w_\mu \geq 0,\ \textstyle\sum_\mu w_\mu \leq 1, \end{aligned} \tag{EP$_0$}$$

where $h_0 \equiv \iota_0^* h = h \circ \iota_0$, i.e. $h_0(w_1, \dots, w_n) = h(1 - \sum_\mu w_\mu, w_1, \dots, w_n)$. Similarly to (EP), the (unique) solution of (EP$_0$) will lie in the interior $\operatorname{int}(X_0)$ of X_0, so we will have $x = Q(y)$ iff

$$z_\mu = \frac{\partial h_0}{\partial w_\mu} = \frac{\partial h}{\partial x_\mu} - \frac{\partial h}{\partial x_0}, \tag{16}$$

i.e. iff $z = F_0(w)$ where $F_0 \colon \operatorname{int}(X_0) \to \mathbb{R}^{\mathcal{A}_0}$ denotes the gradient $F_0(w) = \nabla h_0(w)$, $w \in \operatorname{int}(X_0)$. As it turns out, the required embedding Q_0 is simply the inverse of F_0:

Lemma 2. *Let $h \colon X \to \mathbb{R}$ be a generalized entropy function. Then, the map $F_0 \equiv \nabla h_0 \colon \operatorname{int}(X_0) \to \mathbb{R}^{\mathcal{A}_0}$ defined above is a diffeomorphism whose inverse $Q_0 \equiv F_0^{-1}$ makes the diagram (15) commute.*

Proof. Proof. The fact that F_0 is a continuous bijection with continuous inverse follows from the general theory of Legendre-type functions—see e.g. Theorem 26.5 in Rockafellar [26]; the diagram (15) then commutes on account of the equivalence $x = Q(y) \Leftrightarrow \pi_0(y) = F_0(\operatorname{proj}_0(x))$. Finally, to show that F_0 is indeed a diffeomorphism, note that the Jacobian of F_0 is just the Hessian of h_0, and with $\operatorname{Hess}(h)$ strictly positive-definite by assumption, our claim follows from the inverse function theorem (see e.g. Lee [17]). □

[8] In fact, Q is constant along $(1, \dots, 1)$: adding $c \in \mathbb{R}$ to every component of $y \in \mathbb{R}^{\mathcal{A}}$ will not change the solution $x = Q(y)$ of (EP).

Having established a diffeomorphism between the variables z_μ and w_μ, let $h_{\mu\nu}$ denote the elements of the corresponding Jacobian matrix $JF_0 = \mathrm{Hess}(h_0)$, i.e.

$$h_{\mu\nu} = \frac{\partial z_\mu}{\partial w_\nu} = \frac{\partial^2 h_0}{\partial w_\mu \partial w_\nu} = \frac{\partial^2 h}{\partial x_\mu \partial x_\nu} + \frac{\partial^2 h}{\partial x_0^2} - \frac{\partial^2 h}{\partial x_0 \partial x_\mu} - \frac{\partial^2 h}{\partial x_0 \partial x_\nu}. \tag{17}$$

Then, letting $h^{\mu\nu} = \frac{\partial w_\mu}{\partial z_\nu}$ denote the inverse of $h_{\mu\nu}$, and combining the learning scheme (ERL) with the evolution equation (ZD), we obtain the (unilateral) *entropy-driven learning dynamics*:

$$\dot{x}_\mu = \dot{w}_\mu = \sum\nolimits_\nu \frac{\partial w_\mu}{\partial z_\nu} \dot{z}_\nu = \sum\nolimits_\nu h^{\mu\nu} \left(\Delta u_\nu - T z_\nu\right), \tag{18}$$

where, as before, $\Delta u_\nu = u_\nu - u_0$ and $z_\nu = \frac{\partial h}{\partial x_\nu} - \frac{\partial h}{\partial x_0}$.[9] Therefore, if the agent's payoffs are coming from a finite game $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$, our previous discussion yields the class of *entropy-driven game dynamics*

$$\dot{x}_{k\mu} = \sum\nolimits_\nu^k h_k^{\mu\nu}(x) \left(u_{k\nu}(x) - u_{k,0}(x)\right) - T \sum\nolimits_\nu^k h_k^{\mu\nu}(x) \left(\frac{\partial h_k}{\partial x_{k\nu}} - \frac{\partial h_k}{\partial x_{k,0}}\right), \tag{ED}$$

where now $h_k : X_k \to \mathbb{R}$ is the entropy function of player k (generating the corresponding player-specific choice map $Q_k : \mathbb{R}^{\mathcal{A}_k} \to X_k$) and $h_k^{\mu\nu}$ is the inverse Hessian matrix of h_k defined as in (17).

These dynamics will be the main focus of the rest of our paper, so some remarks are in order:

Remark 1. To get some trivial book-keeping out of the way (and to keep our notation as light as possible), note that the player-specific entropy functions h_k may be encoded in the *aggregate entropy* $h : X \to \mathbb{R}$ with $h(x_1, \dots, x_N) = \sum_k h_k(x_k)$, $x_k \in X_k$. Likewise, if we set $X = \prod_k X_k$ and replace $\mathcal{A}$ with $\mathcal{A}^* \equiv \coprod_k \mathcal{A}_k$ in Definition 1, then the player-specific choice maps $Q_k : \mathbb{R}^{\mathcal{A}_k} \to X_k$ may themselves be encoded in the composite choice map $Q : \mathbb{R}^{\mathcal{A}^*} \cong \prod_k \mathbb{R}^{\mathcal{A}_k} \to X$ associated to h. Therefore, whenever we mention entropy functions and choice maps in the context of a game (and not simply in a discrete choice problem), it should be understood that we are referring to the above construction.

Remark 2. It is also important to note that the dynamics (ED) admit *global solutions*, i.e. solutions that remain in $\mathrm{int}(X_0)$ for all $t \geq 0$. This can be proven directly using the differential system (ED), but the score representation (5) probably yields a more transparent view: since the payoff streams $u_\alpha(t)$ are Lipschitz and bounded,[10] the scores $y_\alpha(t)$ will remain finite for all $t \geq 0$, so interior solutions $x(t) = Q(y(t))$ of (ED) will be defined for all $t \geq 0$ themselves.[11]

Remark 3. The previous remark brings up an important distinction between interior and non-interior orbits: strictly speaking, (ED) is only defined on $\mathrm{int}(X_0)$, so boundary initial conditions must be handled with more care. To address initial conditions $x(0) \in X$ with arbitrary support $\mathcal{A}' \equiv \mathrm{supp}(x(0)) \subseteq \mathcal{A}$, it will be convenient to assume that the entropy h is regular; in that case, by restricting (EP) to the subface $X' \equiv \Delta(\mathcal{A}')$ of X, we obtain a similarly restricted choice map $Q' : \mathbb{R}^{\mathcal{A}'} \to X'$ and the agent may proceed by updating the scores of the supported actions $\alpha \in \mathcal{A}'$ in (ERL). In this way, every subface X' of X becomes an invariant manifold of (ERL)/(ED), so entropy-driven dynamics are seen to belong to the general class of imitative dynamics introduced by Björnerstedt and Weibull [6] (see also Weibull [34]).

Remark 4. In addition to tuning their learning temperature $T > 0$, players can also try to sharpen their response model by replacing the choice stage of (ERL) with

$$x_k = Q_k(\lambda_k y_k) \tag{19}$$

[9] Note that $\dot{x}_0 = -\sum_\mu \dot{x}_\mu$, so the action $\alpha_0 \in \mathcal{A}$ is not being discriminated against).

[10] Importantly, this property remains true in the case of several agents involved in a finite game.

[11] Interestingly, for $T = 0$, this can be seen as an alternative proof of Theorem 4.1 of Alvarez et al. [2] on the existence of global solutions in Hessian–Riemannian gradient descent dynamics.

for some $\lambda_k \geq 0$. As can be seen from (EP), these choice parameters may then be viewed as (player-specific) *inverse choice temperatures*: as $\lambda_k \to \infty$, the choices of player k freeze down to a "best-responding" to the stimulus y, whereas for $\lambda_k \to 0$, player k mixes actions uniformly, without regards to their performance scores. On the other hand, the same reasoning that led to (ED) also yields the choice-adjusted dynamics

$$\dot{x}_{k\mu} = \lambda_k \sum\nolimits_{\nu}^{k} h_k^{\mu\nu}(x)\,(u_{k\nu}(x) - u_{k,0}(x)) - \lambda_k T \sum\nolimits_{\nu}^{k} h_k^{\mu\nu}(x) \left(\frac{\partial h_k}{\partial x_{k\nu}} - \frac{\partial h_k}{\partial x_{k,0}} \right). \tag{ED$_\lambda$}$$

We thus see that the inverse temperature λ_k of the player's choice model and the temperature T of their learning model (5) play very different roles on the resulting learning dynamics (ED$_\lambda$). The learning temperature T affects only the entropic correction term of (ED$_\lambda$) whereas λ_k affects all terms of (ED$_\lambda$) commensurately; in fact, λ_k can also be seen as a player-specific change of time, an observation which will be crucial in considering players with different update schedules in Section 4.

We will close this section by noting that even though the dynamics (ED) are fully updateable on X (as opposed to the otherwise equivalent process (ERL) which interweaves X and $\mathbb{R}^{\mathcal{A}}$), the RHS of (ED) still contains a non-explicit step in the calculation of the inverse matrix $h^{\mu\nu}$. Given that the computational complexity of inverting a matrix is polynomial in the size of the matrix, this does not pose much of a problem in practical applications (after all, it is the number of players that usually explodes, not the number of actions per player).

That said, (ED) can be simplified considerably if the entropy function which is driving the process is itself decomposable. Indeed, assume that $h(x) = \sum_\beta \theta(x_\beta)$ for some non-degenerate Legendre kernel $\theta\colon [0, +\infty) \to \mathbb{R} \cup \{+\infty\}$ as in Remark 2 following Definition 1. In that case, (17) readily yields

$$h_{\mu\nu} = \frac{\partial^2 h}{\partial x_\mu \partial x_\nu} + \frac{\partial^2 h}{\partial x_0^2} - \frac{\partial^2 h}{\partial x_0 \partial x_\mu} - \frac{\partial^2 h}{\partial x_0 \partial x_\nu} = \theta''(x_\mu)\,\delta_{\mu\nu} + \theta''(x_0), \tag{20}$$

so $h^{\mu\nu}$ can be calculated from the following lemma:

Lemma 3. *Let $A_{\mu\nu} = q_\mu \delta_{\mu\nu} + q_0$ with $q_0, q_1, \dots, q_n > 0$. Then, the inverse $A^{\mu\nu}$ of $A_{\mu\nu}$ will be:*

$$A^{\mu\nu} = \frac{\delta_{\mu\nu}}{q_\mu} - \frac{q_h}{q_\mu q_\nu}, \tag{21}$$

where q_h is the harmonic aggregate: $q_h^{-1} \equiv \sum_{\alpha=0}^{n} q_\alpha^{-1}$.

Proof. Proof. By simple inspection, we have:

In view of the above inversion formula applied to (20), and setting $z_{k\mu} = \frac{\partial h}{\partial w_{k\mu}} = \theta'(x_{k\mu}) - \theta'(x_{k,0})$ in (ED), some algebra finally yields the *entropic dynamics with kernel θ*:

$$\dot{x}_{k\alpha} = \frac{1}{\theta''(x_{k\alpha})} \left[u_{k\alpha}(x) - \theta_h''(x_k) \sum\nolimits_\beta^k \frac{u_{k\beta}(x)}{\theta''(x_{k\beta})} \right] - \frac{T}{\theta''(x_{k\alpha})} \left[\theta'(x_{k\alpha}) - \theta_h''(x_k) \sum\nolimits_\beta^k \frac{\theta'(x_{k\beta})}{\theta''(x_{k\beta})} \right], \tag{ED$_\theta$}$$

where θ_h'' denotes the harmonic aggregate $\theta_h''(x_k) = \left(\sum_\beta^k 1/\theta''(x_{k\beta}) \right)^{-1}$.[12] Hence, in the important case of the Boltzmann-Gibbs kernel $\theta(x) = x \log x$, we readily obtain the *temperature-adjusted replicator dynamics*

$$\dot{x}_{k\alpha} = x_{k\alpha} \left[u_{k\alpha}(x) - \sum\nolimits_\beta^k x_{k\beta} u_{k\beta}(x) \right] - T\, x_{k\alpha} \left[\log x_{k\alpha} - \sum\nolimits_\beta^k x_{k\beta} \log x_{k\beta} \right] \tag{T-RD}$$

[12] Needless to say, θ_h'' is not a second derivative; we just use this notation for visual consistency.

which, for $T = 0$, freeze to the ordinary (asymmetric) replicator dynamics of Taylor and Jonker [31]:

$$\dot{x}_{k\alpha} = x_{k\alpha}\left[u_{k\alpha}(x) - \sum\nolimits_{\beta}^{k} x_{k\beta}u_{k\beta}(x)\right]. \tag{RD}$$

3 Entropy-driven game dynamics and rationality.

In this section, our aim will be to analyze the entropy-driven dynamics (ED) from the point of view of rational agents, mostly looking to determine their asymptotic convergence properties with respect to standard game-theoretic solution concepts. Thus, in conjunction with the notion of Nash equilibrium, we will also focus on the widely studied concept of a *quantal response equilibrium*:

Definition 4 (McKelvey and Palfrey, 1995)**.** Let $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ be a finite game in normal form and let $Q\colon \mathbb{R}^{\mathcal{A}} \to X$ be a regular choice function. We will say that $q \in X$ is a *quantal response equilibrium (QRE)* of $\mathfrak{G}$ with respect to Q (or a Q*-equilibrium* for short) when, for some $\varrho \geq 0$,

$$q = Q\left(\varrho u(q)\right), \tag{QRE}$$

where $u(q) \in \prod_k \mathbb{R}^{\mathcal{A}_k}$ denotes the payoff vector of the profile $q \in X$. More generally, if $\mathfrak{G}' \equiv \mathfrak{G}'(\mathcal{N}, \mathcal{A}', u|_{\mathcal{A}'})$ is a restriction of $\mathfrak{G}$, we will say that $q \in X$ is a *restricted QRE* of $\mathfrak{G}$ if it is a QRE of $\mathfrak{G}'$.

The scale parameter $\varrho \geq 0$ will be called the rationality level of the QRE in question. Obviously, when $\varrho = 0$, players choose actions uniformly, without any regard to their payoffs; at the other end of the spectrum, when $\varrho \to \infty$, players become fully rational and the notion of a QRE approximates smoothly that of a Nash equilibrium. Finally, one could also consider negative rationality levels, in which case players become *anti-rational*: for $\varrho < 0$, the condition $x = Q(\varrho u(x))$ characterizes the QRE of the opposite game $-\mathfrak{G} = (\mathcal{N}, \mathcal{A}, -u)$, and as $\varrho \to -\infty$, these equilibria approximate the Nash equilibria of $-\mathfrak{G}$.

To make this approximation idea more precise, let $q^* \in X$ be a Nash equilibrium of a finite game $\mathfrak{G}$ and let $\gamma\colon U \to X$ be a smooth curve on X defined on a half-infinite interval of the form $U = [a, +\infty)$, $a \in \mathbb{R}$. We will then say that γ is a Q*-path to* q^* when $\gamma(\varrho)$ is a Q-equilibrium of $\mathfrak{G}$ with rationality level ϱ and $\lim_{\varrho\to\infty}\gamma(\varrho) = q^*$; in a similar vein, we will say that $q \in X$ is a Q*-approximation* of q^* when q is itself a Q-equilibrium and there is a Q-path joining q to q^* (van Damme [33] uses the terminology *approachable*).

Example 1. By far the most widely used specification of a QRE is the *logit equilibrium* which corresponds to the Gibbs choice map (8): in particular, $q \in X$ will be a logit equilibrium of $\mathfrak{G}$ when $q_{k\alpha} = \exp(\varrho u_{k\alpha}(q))/\sum_{\beta}\exp(\varrho u_{k\beta}(q))$ for all $\alpha \in \mathcal{A}_k$, $k \in \mathcal{N}$.

Our first result links the rest points of (ED) at temperature T with the game's restricted QRE:

Proposition 5. *Let $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ be a finite game and let $h\colon X \to \mathbb{R}$ be a regular entropy function with choice map $Q\colon \mathbb{R}^{\mathcal{A}} \to X$. Then:*

1. *For $T > 0$, the rest points of the entropy-driven dynamics (ED) coincide with the restricted QRE of $\mathfrak{G}$ with rationality level $\varrho = 1/T$.*
2. *For $T = 0$, the rest points of (ED) are the restricted Nash equilibria of $\mathfrak{G}$.*
3. *For $T < 0$, the rest points of (ED) are the restricted QRE of the opposite game $-\mathfrak{G}$.*

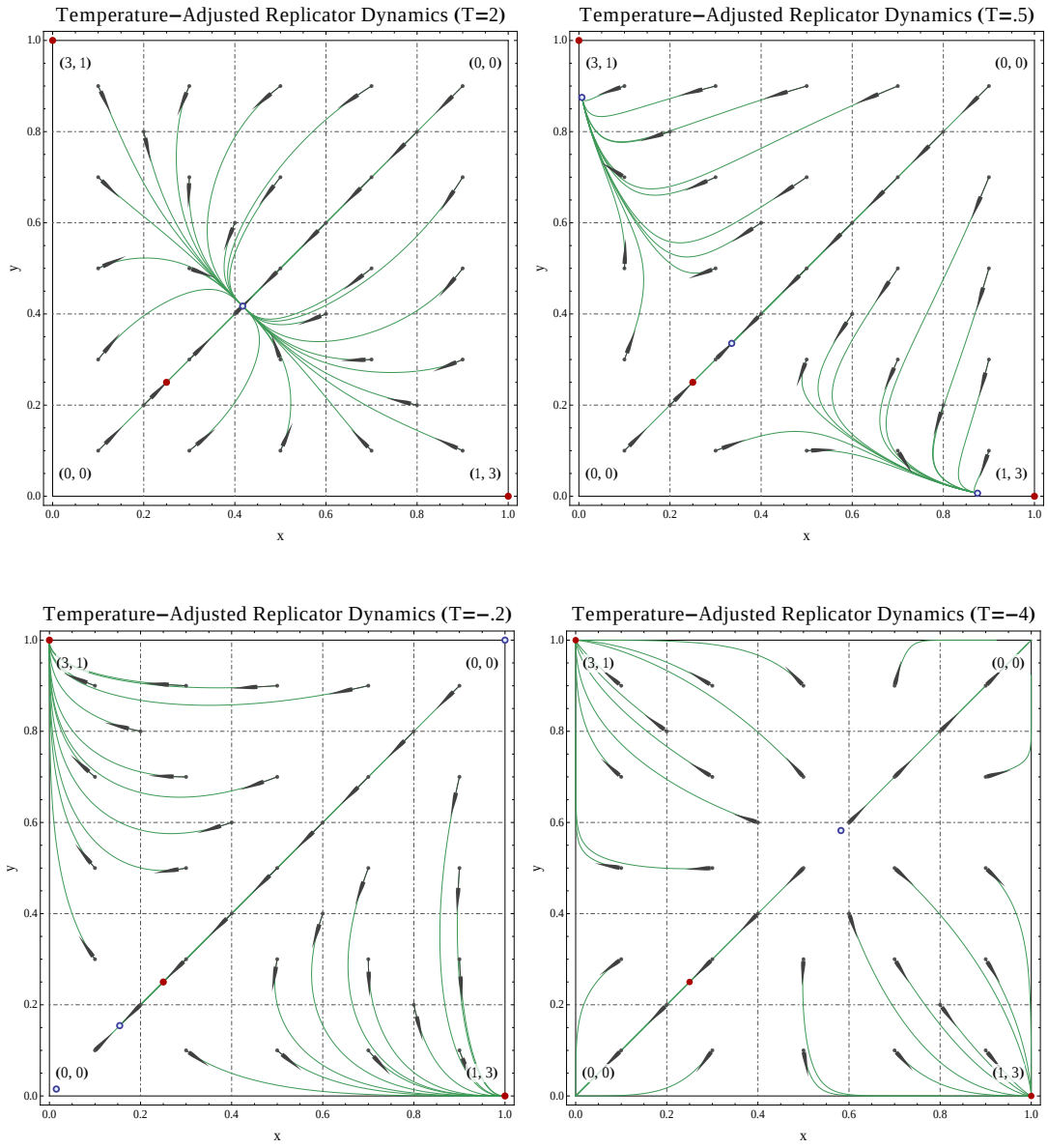


Figure 1: Phase portraits of the temperature-adjusted replicator dynamics (T-RD) in a 2×2 potential game (Nash equilibria are depicted in dark red and interior rest points in light/dark blue; see labels for the game's payoffs). For high learning temperatures $T \gg 0$, the dynamics cannot keep track of payoffs and their only rest point is a global attractor which approaches the barycenter of X as $T \to +\infty$ (corresponding to a QRE under stochastic perturbations of very high magnitude). As the temperature drops to around $T \approx 0.935$, this attractor becomes unstable and undergoes a supercritical pitchfork bifurcation (a phase transition) resulting in the appearance of two asymptotically stable QRE that converge to the strict Nash equilibria of the game as $T \to 0^+$. For negative temperatures, the non-equilibrium vertices of X become asymptotically stable (but with a very small basin of attraction), and each of them gives birth to an unstable equilibrium in a subcritical pitchfork bifurcation. Of these two equilibria, the one closer to the game's interior Nash equilibrium is annihilated with the pre-existing QRE at $T \approx -0.278$, and as $T \to -\infty$, we obtain a time-inverted image of the $T \to +\infty$ portrait with the only remaining QRE repelling all trajectories towards the vertices of X.

Proof. Proof. Since our focus is on restricted equilibria, it clearly suffices to prove the above correspondences for interior rest points; since the faces of X are forward-invariant under the dynamics (ED), the general claim then follows from passing to an appropriate restriction $\mathfrak{G}'$ of $\mathfrak{G}$.

To that end, let $T > 0$ and note that Eq. (13) on the evolution of the relative scores $z_{k\mu}$ allows us to characterize the rest points of (ED) by means of the equation $\Delta u_{k\mu}(x) - Tz_{k\mu} = 0$; equivalently, in terms of the absolute scores $y_{k\alpha}$, we will have $u_{k\alpha}(x) - Ty_{k\alpha} = u_{k\beta}(x) - Ty_{k\beta}$ for all $\alpha, \beta \in \mathcal{A}_k$, $k \in \mathcal{N}$. However, with $x = Q(y)$, we will also have $y_{k\alpha} - \frac{\partial h}{\partial x_{k\alpha}} = y_{k\beta} - \frac{\partial h}{\partial x_{k\beta}}$, so we readily obtain $u_{k\alpha}(x) - T\frac{\partial h}{\partial x_{k\alpha}} = u_{k\beta}(x) - T\frac{\partial h}{\partial x_{k\beta}}$; in turn, this shows that $x = Q(T^{-1}u(x))$, so x is a Q-equilibrium of $\mathfrak{G}$ with rationality level $\varrho = 1/T$.

For $T = 0$ the result is immediate because (13) shows that $x \in X$ is an interior rest point of (ED) if and only if $u_{k\alpha}(x) = u_{k\beta}(x)$ for all $\alpha, \beta \in \mathcal{A}_k$, $k \in \mathcal{N}$, i.e. if and only if x is an interior Nash equilibrium of $\mathfrak{G}$ (recall that the Hessian of h is non-singular). Finally, the time inversion $t \mapsto -t$ in (ED) is equivalent to the inversion $u \mapsto -u$, $T \mapsto -T$, so our claim for negative T follows from the case $T > 0$. □

Remark 1. Proposition 5 shows that the temperature T of the dynamics (ED) plays a double role: on the one hand, it determines the discount (or reinforcement) factor in the players' assessment phase (5), so it reflects the importance that they give to past events; on the other hand, T also determines the rationality level of the rest points of (ED), measuring how far the stationary points of the players' learning process are from being Nash. Perhaps unsurprisingly, this dual role of the temperature is brought to the forefront by the probabilistic/perturbed interpretation of quantal responses as choice probabilities in the case of stochastically perturbed payoffs. Indeed, recalling the relevant discussion of Section 2.3, we see that a QRE with rationality level $\varrho = T^{-1}$ corresponds to best responding in the presence of a noise process with standard deviation $\varepsilon \propto \varrho^{-1} = T$. On that account, the players' learning temperature simply measures the inherent variance (inverse rationality) of a QRE, just as the physical notion of temperature measures the variance of the random motions of the particles that make up a thermodynamic system (e.g. an ideal gas following Maxwell-Boltzmann statistics).

Of course, stationarity does not capture the long-term behavior of a dynamical system, so the rest of our analysis will be focused on the convergence properties of (ED). To that end, we begin with the special case of potential games where the players' payoff functions are aligned along a common potential function as in Eq. (2). In this setting, our first result is that for small temperatures, the game's potential function is "almost" increasing along the solution orbits of (ED):

Lemma 6. *Let $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ be a finite potential game with potential U, and let $h\colon X \to \mathbb{R}$ be a generalized entropy function. Then, the function $F(x) \equiv Th(x) - U(x)$ is Lyapunov for the entropy-driven dynamics (ED): for any interior orbit $x(t)$ of (ED), we will have $\frac{d}{dt}F(x(t)) \leq 0$ with equality if and only if $x(0)$ is a QRE of $\mathfrak{G}$ (or $-\mathfrak{G}$ for $T < 0$).*

Proof. Proof. By expressing F in the reduced coordinates $w_{k\mu} = x_{k\mu}$ that we used in the derivation of (ED) (see also Remark 4 following Definition 1), we readily obtain:

$$\frac{dF}{dt} = \sum_k \sum_\mu^k \frac{\partial F}{\partial w_{k\mu}} \dot{w}_{k\mu} = \sum_k \sum_{\mu,\nu}^k (T\frac{\partial h}{\partial w_{k\mu}} - \Delta u_{k\mu}) h_k^{\mu\nu} (\Delta u_{k\mu} - T\frac{\partial h}{\partial w_{k\mu}}), \tag{22}$$

with the second equality following from the fact that $\mathfrak{G}$ is a potential game, so $\frac{\partial u_k}{\partial x_{k\alpha}} = \frac{\partial U}{\partial x_{k\alpha}}$ by (2). Thus, with $\mathrm{Hess}(h) \succ 0$, (22) implies that $\dot{F}(x) \leq 0$ with equality if and only if $\Delta u_{k\mu}(x) = T\frac{\partial h}{\partial w_{k\mu}} = T(\frac{\partial h}{\partial x_{k\mu}} - \frac{\partial h}{\partial x_{k,0}})$, i.e. if and only if x is a QRE of $\mathfrak{G}$ (or $-\mathfrak{G}$ for $T < 0$; cf. the proof of Proposition 5). □

When the players' entropy function is regular, Lemma 6 can be easily extended to orbits lying on any subface X' of X simply by considering the restricted QRE of the game that are supported in X'. Even in that case however, Lemma 6 makes no distinction between positive and negative temperatures and simply

shows that the dynamics (ED) will tend to move towards the minimizers of F on X'. What changes with the sign of T is the relation that these minimizers have with regards to restricted QRE: for $T < 0$, the only local minimizers of F are the vertices of X (which are themselves pure restricted QRE),[13] whereas for $T > 0$, the restricted QRE of $\mathfrak{G}$ that are supported in a subface X' of X coincide with the local minimizers of $F|_{X'}$. Formally, Lemma 6 and the above reasoning give:

Proposition 7. *Let $h\colon X \to \mathbb{R}$ be a regular entropy function and let $x(t)$ be a solution orbit of the associated entropic dynamics (ED) for a potential game $\mathfrak{G}$. Then:*

1. *For $T > 0$, $x(t)$ converges to a restricted QRE of $\mathfrak{G}$ with the same support as $x(0)$.*
2. *For $T = 0$, $x(t)$ converges to a restricted Nash equilibrium whose support is contained in that of $x(0)$.*
3. *For $T < 0$, $x(t)$ converges to a vertex of X or is stationary (if $x(0)$ is a restricted QRE of $-\mathfrak{G}$).*

The above proposition will be our main result for potential games, so some remarks are in order:

Remark 1. By continuity, the phase portrait of the entropy-driven dynamics for small temperatures (positive or negative) will be broadly similar to the base case $T = 0$ (at least, in the generic case where there are no payoff ties in $\mathfrak{G}$). Accordingly, the main difference between positive and negative temperatures is that for small $T < 0$ the dynamics converge to a bona fide (pure) Nash equilibrium for most initial conditions (except for a small basin of attraction around each vertex of X which pulls the dynamics to non-equilibrium vertices), whereas for small $T > 0$, interior solutions of (ED) always converge to a Q-approximation of a Nash equilibrium (see also Fig. 1). As we shall see in Section 4, the fact that the entropy-driven dynamics always converge to the vicinity of a Nash equilibrium for small $T > 0$ will be crucial in the presence of imperfect payoff information and/or stochastic fluctuations.

Remark 2. Interpreting the game's potential U as the internal energy of a thermodynamic system, then, modulo a change of sign,the Lyapunov function $F(x) = Th(x) - U(x)$ is known in statistical physics as the *(Helmholtz) free energy* and measures the useful work that can be obtained from a thermodynamic system at constant temperature (Landau and Lifshitz [16]).[14] In this context, the principle of energy minimization states that the free energy of an isolated system never increases, so Lemma 6 may be viewed as a corollary of the Second Law of Thermodynamics: *under constant temperature, the free energy of the system decreases until it reaches a thermal equilibrium.*

The previous discussion establishes a fundamental qualitative difference between positive and negative learning temperatures in potential games: for $T < 0$, every vertex of X is attracting in (ED), while for $T > 0$, the dynamics can only converge to interior points. As we show below, this behavior actually applies to *any* finite game:

Proposition 8. *Let $h\colon X \to \mathbb{R}$ be a regular entropy function. Then:*

1. *For negative temperatures $T < 0$, every vertex $q \in X$ is attracting in the entropic dynamics (ED).*
2. *For positive temperatures $T > 0$, any ω-limit point of an interior solution orbit $x(t)$ is itself interior; in fact, the ω-limit set of $\mathrm{int}(X)$ and the boundary $\mathrm{bd}(X)$ of X are separated by neighborhoods.*

[13] Perhaps the easiest way to see is this is to note that F is subharmonic: $\frac{\partial^2 U}{\partial x_{k\alpha}^2} = 0$ for all $\alpha \in \mathcal{A}_k$, $k \in \mathcal{N}$, on account of U being multilinear, and $\frac{\partial^2 h}{\partial x_{k\alpha}^2} > 0$ on account of h having a positive-definite Hessian.

[14] Recall that our sign convention for the entropy is the opposite of physics and probability; furthermore, potentials are *minimized* in physics, so h and U should be replaced by $-h$ and $-U$ respectively, yielding the familiar expression $F = U - Th$ (Landau and Lifshitz [16]).

Proof. Proof. Our proof will be based on the dynamics (13) for the relative scores $z_{k\mu}$, $\mu \in \mathcal{A}_{k,0} \equiv \mathcal{A}_k \setminus \{\alpha_{k,0}\}$. In integral form, we have:

$$z_{k\mu}(t) = z_{k\mu}(0)e^{-Tt} + \int_0^t e^{-T(t-s)}\Delta u_{k\mu}(x(s))\, ds, \tag{23}$$

so, with $\Delta u_{k\mu}$ bounded on X, the last integral will be bounded in absolute value by $\frac{M_k}{T}(1 - e^{-Tt})$ for some $M_k > 0$. We will thus have

$$\left(z_{k\mu}(0) + M_k/T\right) e^{-Tt} - M_k/T \leq z_{k\mu}(t) \leq M_k/T + \left(z_{k\mu}(0) - M_k/T\right) e^{-Tt}, \tag{24}$$

and for $T > 0$, any ω-limit point of (13) will lie in the rectangle $C_T = \{z \in \prod_k \mathbb{R}^{\mathcal{A}_{k,0}} : |z_{k\mu}| \leq M_k/T\}$. However, the image of C_T under the reduced choice map $Q_0 \colon \prod \mathbb{R}^{\mathcal{A}_{k,0}} \to X_0$ will be a compact set that is wholly contained in the interior of X_0, thus proving our assertion for $T > 0$.

On the other hand, for $T < 0$, (24) shows that if we pick $z_{k\mu}(0) < -M_k/|T|$, then we will have $\lim_{t\to\infty} z_{k\mu}(t) = -\infty$ for all $\mu \in \mathcal{A}_{k,0}$, $k \in \mathcal{N}$. Since the set $U_T = \{z \in \prod_k \mathbb{R}^{\mathcal{A}_{k,0}} : z_{k\mu} < -M_k|T|^{-1}\}$ is a neighborhood of $(-\infty, \ldots, -\infty)$ in $\prod_k \mathbb{R}^{\mathcal{A}_{k,0}}$ which is mapped diffeomorphically by Q_0 to the relative interior of a neighborhood of $(0, \ldots, 0)$ in X_0, it follows that the pure vertex $q = (\alpha_{1,0}, \ldots, \alpha_{N,0})$ of X attracts all nearby interior solutions of (ED). By restriction, this property will hold on any subface of X which contains q, and with the choice of flagged actions $\alpha_{k,0} \in \mathcal{A}_k$ being arbitrary, our proof is complete. □

This dichotomy in the behavior of the entropic dynamics (ED) for positive and negative temperature ties in with the following result which is of independent interest:

Proposition 9. *Let $h \colon X \to \mathbb{R}$ be a generalized entropy function. Then, there exists a volume form* Vol_h *on* $\mathrm{int}(X)$ *such that if $U_0 \subseteq X$ is relatively open in X and* $\mathrm{cl}(U_0) \cap \mathrm{bd}(X) = \varnothing$*, then:*

$$\mathrm{Vol}_h(U_t) = \mathrm{Vol}_h(U_0)\, \exp(-A_0 T t), \tag{25}$$

where $A_0 = \mathrm{card}(\coprod_k \mathcal{A}_k) - \mathrm{card}(\mathcal{N}) = \sum_k(\mathrm{card}(\mathcal{A}_k) - 1)$*, and* $U_t \equiv \{x(t) : x(0) \in U_0\}$*. Hence, the entropy-driven game dynamics (ED) are contracting for $T > 0$, expanding for $T < 0$ and incompressible iff $T = 0$.*

Proof. Proof. Again, our proof will be based on the dynamics (13) for the relative score variables $z_{k\mu}$, $\mu \in \mathcal{A}_{k,0} \equiv \mathcal{A}_k \setminus \{\alpha_{k,0}\}$, which are related to the mixed strategy variables $x \in X$ via the commutative diagram (15): $Q_0(z) = \mathrm{proj}_0(x)$. Indeed, if V_0 is an open set of $\prod_k \mathbb{R}^{\mathcal{A}_k}$ and $W_{k\mu} = \Delta u_{k\mu}(x) - T z_{k\mu}$ denotes the RHS of (13), then, by Liouville's theorem, we will have

$$\frac{d}{dt}\,\mathrm{Vol}(V_t) = \int_{V_t} \mathrm{div}\, W\, dV, \tag{26}$$

where $dV = \bigwedge_{k,\mu} dz_{k\mu}$ is the ordinary Euclidean volume form of $\prod_k \mathbb{R}^{\mathcal{A}_{k,0}}$, Vol denotes the associated (Lebesgue) measure, and V_t is the image of V_0 at time t under (13), viz. $V_t = \{z(t) : z(0) \in V_0\}$. However, since $\Delta u_{k\mu}$ does not depend on z_k (because $u_{k\mu}$ and $u_{k,0}$ themselves do not depend on x_k), we will have $\frac{\partial W_{k\mu}}{\partial z_{k\mu}} = -T$. Hence, summing over all $\mu \in \mathcal{A}_{k,0}$, $k \in \mathcal{N}$, we obtain $\mathrm{div}\, W = -\sum_k(\mathrm{card}(\mathcal{A}_k) - 1)T = -A_0 T$, and by integrating, we obtain the volume evolution equation $\mathrm{Vol}(V_t) = \mathrm{Vol}(V_0)\exp(-A_0 T t)$.

In view of the above, let $\mathrm{Vol}_h = \mathrm{proj}_0^*(Q_0^{-1})^*\,\mathrm{Vol}$ be the push-forward of the Euclidean volume $\mathrm{Vol}(\cdot)$ to $\mathrm{int}(X)$ via the diffeomorphism $Q_0^{-1} \circ \mathrm{proj}_0 : \mathrm{int}(X) \to \prod_k \mathbb{R}^{\mathcal{A}_k}$, i.e. $\mathrm{Vol}_h(U) = \mathrm{Vol}(Q_0^{-1}(\mathrm{proj}_0(U)))$ for any relatively open set $U \in \mathrm{int}(X)$. Then, taking V_0 such that $\mathrm{proj}_0(U_0) = Q_0(V_0)$, our assertion follows from the volume evolution equation above by recalling that $\mathrm{proj}_0(x(t)) = Q_0(z(t))$. □

Interestingly, in the special case of the Boltzmann-Gibbs entropy at zero temperature, Proposition 9 yields the classical result that the asymmetric replicator dynamics (RD) are incompressible and hence do not admit interior attractors (Hofbauer and Sigmund [13], Ritzberger and Weibull [25]).[15] We thus see that incompressibility characterizes a much more general class of dynamics and, in our learning context, it is simply a consequence of the players assigning a uniform weight to their past observations (neither discounting, nor reinforcing them).

That said, in the case of the replicator dynamics, we have a significantly clearer picture for the stability and attraction properties of a game's equilibria thanks to the *folk theorem of evolutionary game theory* (Hofbauer and Sigmund [13]). In particular, it is well known that:

1. If an interior trajectory converges, its limit is Nash.
2. If a state is Lyapunov stable, then it is also Nash.
3. A state is asymptotically stable if and only if it is a strict Nash equilibrium.[16]

By comparison, in the context of our more general class of entropy-driven game dynamics we obtain:

Theorem 10. *Let $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ be a finite game and let $h\colon X \to \mathbb{R}$ be a regular entropy function with choice map $Q\colon \prod_k \mathbb{R}^{\mathcal{A}_k} \to X$. Then, the entropy-driven dynamics (ED) have the following properties:*

1. *For positive temperatures $T > 0$, if $q \in X$ is Lyapunov stable then it is also a QRE of $\mathfrak{G}$; moreover, if q is a Q-approximate strict Nash equilibrium and T is small enough, then q is also asymptotically stable.*
2. *For $T = 0$, if $q \in X$ is Lyapunov stable, then it is also a Nash equilibrium of $\mathfrak{G}$; furthermore, q is asymptotically stable if and only if it is a strict Nash equilibrium of $\mathfrak{G}$.*
3. *Finally, for $T < 0$, a profile $q \in X$ will be asymptotically stable if and only if it is pure (i.e. a vertex of X); any other rest point of (ED) is unstable.*

Proof. Proof. Our proof will be broken up in three parts based on the temperature of the dynamics (ED):

Positive temperatures. Let $T > 0$ and assume that $q \in X$ is Lyapunov stable (and, hence, stationary). Clearly, if q is interior, it will be a QRE of $\mathfrak{G}$ by Proposition 5 so there is nothing to show. Suppose therefore that $q \in \mathrm{bd}(X)$; then, by Proposition 8, we may pick a neighborhood U of q in X such that $\mathrm{cl}(U)$ does not contain any ω-limit points of the interior of X under (ED). However, since q is Lyapunov stable, any interior solution that is wholly contained in U will have an ω-limit in $\mathrm{cl}(U)$, a contradiction.

Regarding asymptotic stability, we will make the simplifying technical assumption that the entropy function h is decomposable with the same Legendre kernel θ for all players (the proof is entirely similar in the general case, but significantly more painful to write down). Assuming further that $q^* = (\alpha_{1,0}, \dots, \alpha_{N,0})$ is a strict Nash equilibrium of $\mathfrak{G}$ and let $q \equiv q(T) \in X$ be a Q-approximation of q^* with rationality level $\varrho = 1/T$. Then, letting $W_{k\mu} = \Delta u_{k\mu} - T z_{k\mu}$ denote the RHS of the dynamics (13), a simple differentiation yields:

$$\left.\frac{\partial W_{k\mu}}{\partial z_{\ell\nu}}\right|_q = \begin{cases} -T & \text{if } \ell = k, \nu = \mu, \\ 0 & \text{if } \ell = k, \nu \neq \mu, \\ \sum_\rho^\ell h_\ell^{\nu\rho}(q) \frac{\partial}{\partial w_{\ell\rho}} \Delta u_{k\mu} & \text{otherwise,} \end{cases} \tag{27}$$

[15] This does not hold in the symmetric case: there the proof breaks down because the symmetrized payoff $u_\alpha(x)$ depends on x_α.

[16] Recall that $q \in X$ is said to be *Lyapunov stable* (or *stable*) when for every neighborhood U of q in X, there exists a neighborhood V of q in X such that if $x(0) \in V$ then $x(t) \in U$ for all $t \geq 0$. Moreover, q is called *attracting* when there exists a neighborhood U of q in X such that $\lim_{t\to\infty} x(t) = q$ if $x(0) \in U$, and q is called *asymptotically stable* when it is both stable and attracting.

where $h_\ell^{\nu\rho}(q)$ is the inverse Hessian matrix of h evaluated at q and $\frac{\partial}{\partial w_{\ell\rho}} = \frac{\partial}{\partial x_{\ell\rho}} - \frac{\partial}{\partial x_{\ell,0}}$ as before. In particular, using the inversion formula of Lemma 3, we will have

$$h_\ell^{\nu\rho}(q) = \frac{\delta_{\nu\rho}}{\theta''(q_{\ell\nu})} - \frac{\theta_h''(q_\ell)}{\theta''(q_{\ell\nu})\theta''(q_{\ell\rho})}, \tag{28}$$

where $\theta_h''(q_\ell)$ denotes the harmonic aggregate: $\theta_h''(q_\ell) = \big(\sum_\beta^\ell 1/\theta''(q_{\ell\beta})\big)^{-1}$.

Since q is a Q-approximation of the strict equilibrium $q^* = (\alpha_{1,0}, \dots, \alpha_{N,0})$, we will also have $q_{k\mu} \equiv q_{k\mu}(T) \to q_{k\mu}^* = 0$ and $q_{k,0} \to q_{k,0}^* = 1$ as $T \to 0^+$. Moreover, recalling that q is a QRE of $\mathfrak{G}$ with rationality level $\varrho = 1/T$, we will also have $\Delta u_{k\mu}(q) = T\theta'(q_{k\mu}) - T\theta'(q_{k,0})$, implying in turn that $T\theta'(q_{k\mu}(T)) \to \Delta u_{k\mu}(q^*) < 0$ as $T \to 0^+$. However, with h regular, the Legendre kernel θ of h will satisfy $\theta'(x)/\theta''(x) \to 0$ as $x \to 0^+$, whence we obtain

$$\frac{1}{T\theta''(q_{k\mu}(T))} = \frac{\theta'(q_{k\mu}(T))}{\theta''(q_{k\mu}(T))} \frac{1}{T\theta'(q_{k\mu}(T))} \to \frac{0}{\Delta u_{k\mu}(q^*)} = 0. \tag{29}$$

Thus, on account of (28) and (29), the off-diagonal elements of (27) will be $o(T)$ as $T \to 0^+$, so, by continuity, the eigenvalues of the Jacobian of the vector field W evaluated at $q \equiv q(T)$ will all be negative if $T > 0$ is small enough. As a result, q will be a hyperbolic rest point of (13), so it will also be structurally stable by the Hartman-Grobman theorem, and hence asymptotically stable as well.

Zero temperature. For $T = 0$, let q be Lyapunov stable so every neighborhood U of q in X admits an interior orbit $x(t)$ that stays in U for all $t \geq 0$; we then claim that q is Nash. Indeed, assume ad abusrdum that $\alpha_{k,0} \in \operatorname{supp}(q)$ has $u_{k,0}(q) < u_{k\mu}(q)$ for some $\mu \in \mathcal{A}_k$ and let U be a neighborhood of q such that $x_{k,0} > q_{k,0}/2$ and $\Delta u_{k\mu}(x) \geq m > 0$ for all $x \in U$. Then, picking an orbit $x(t)$ that is wholly contained in U, the integral equation (23) gives $z_{k\mu}(t) \geq z_{k,0}(0) + mt$, implying in turn that $z_{k\mu}(t) \to +\infty$ as $t \to \infty$. However, with $z_{k\mu} = \frac{\partial h}{\partial x_{k\mu}} - \frac{\partial h}{x_{k,0}}$ and h regular this is only possible if $x_{k\mu}(t) \to 0$, a contradiction.

Assume now that $q = (\alpha_{k,1}, \dots, \alpha_{k,N})$ is a strict Nash equilibrium of $\mathfrak{G}$ and let $\mathcal{A}_{k,0} \equiv \mathcal{A}_k \setminus \{\alpha_{k,0}\}$ as usual. To show first that q is Lyapunov stable, it will be again convenient to work with the relative scores $z_{k\mu}$ and show that if $m \in \mathbb{R}$ is sufficiently negative, then every trajectory $z(t)$ that starts in the open set $U_m = \{z \in \prod_k \mathbb{R}^{\mathcal{A}_{k,0}} : z_{k\mu} < m\}$ always stays in U_m; since U_m is a neighborhood of $(-\infty, \dots, -\infty)$ in $\prod_k \mathbb{R}^{\mathcal{A}_{k,0}}$, this is easily seen to imply Lyapunov stability for q in X.

In view of the above, pick $m \in \mathbb{R}$ so that $\Delta u_{k\mu}(x(z)) \leq -\varepsilon < 0$ for all $z \in U_m$ and let $\tau_m = \inf\{t : z(t) \notin U_m\}$ be the time it takes $z(t)$ to escape U_m. Then, if τ_m is finite and $t \leq \tau_m$, the integral form (23) of the relative score dynamics (13) readily yields

$$z_{k\mu}(t) = z_{k\mu}(0) + \int_0^t \Delta u_{k\mu}(Q_0(z(s)))\, ds \leq z_{k\mu}(0) - \varepsilon t < m \quad \text{for all } \mu \in \mathcal{A}_{k,0},\, k \in \mathcal{N}. \tag{30}$$

Thus, substituting τ_m for t in (30), we obtain a contradiction to the definition of τ_m and we may conclude that $z(t)$ always stays within U_m if m is chosen negative enough – i.e. q is Lyapunov stable.

To show that q is in addition attracting, it suffices to let $t \to \infty$ in (30) and recall that $Q_0(z) \to q$ when $z \to (-\infty, \dots, -\infty)$. Finally, for the converse implication, assume that q is not pure; in particular, assume that q lies in the relative interior of a non-singleton subface X' – spanned by $\operatorname{supp}(q)$. Then, with h regular, Proposition 9 shows that q cannot attract a relatively open neighborhood U' of initial conditions in X' because (ED) remains volume-preserving when restricted to any subface X' of X. In turn, this implies that q cannot be attracting in X and precludes asymptotic stability, as claimed.

Negative temperatures. For $T < 0$, the fact that every vertex of X is attracting follows from Proposition 8; Lyapunov stability then follows from (24) by noting that if $z_{k\mu}(0) < -M_k|T|^{-1}$, then we will have $z_{k\mu}(t) < z_{k\mu}(0)$ for all $t \geq 0$ (cf. the proof of Proposition 8). Conversely, assume that $q \in X$ is a non-pure Lyapunov stable state. Then, by passing to a subface of X if necessary, we may assume that q is actually interior. In that case however, if we take an interior neighborhood U of q in X, Proposition 9 shows that any neighborhood V of q that is contained in U will eventually grow to a volume larger than that of U under (ED), so there is no open set of trajectories contained in U. This shows that non-pure rest points of (ED) cannot be stable and our proof is complete. □

In conjunction with our previous results, Theorem 10 provides an interesting insight into the role of the dynamics' temperature parameter T: for small $T > 0$, the dynamics (ED) are attracted to the interior of X and they only converge to points that are *approximately* Nash; for small $T < 0$, the bona fide strict Nash equilibria of the game are indeed asymptotically stable, but so are all the vertices of X (albeit with very small basins of attractions); finally, for $T = 0$, the dynamics (ED) are attracted to strict Nash equilibria and only there (see also Fig. 1). We thus obtain the following rule of thumb: *for $T > 0$, the dynamics (ED) converge to states that are almost Nash, whereas for $T < 0$, the dynamics converge to Nash states except for a very small fraction of initial conditions.*

As such, from the point of view of control and optimization, if one seeks to reach the strict Nash equilibria of the game (e.g. as is usually the case when the game is a potential one), it would appear that the zero temperature case provides the best convergence properties. Nonetheless, there are two important caveats to keep in mind: First, if the dynamics (ED) are to be properly implemented as a discrete-time algorithm, then the results of the next section show that the positive temperature regime is much more stable – all the while allowing players to converge arbitrarily close to a strict equilibrium. On the other hand, if one is only interested in the convergence speed of the dynamics (ED), then even arbitrarily small negative temperatures yield convergence rates that are exponentially faster than the $T = 0$ case:

Proposition 11. *Let $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ be a finite game and let $h\colon X \to \mathbb{R}$ be a regular entropy function with choice map $Q\colon \prod_k \mathbb{R}^{\mathcal{A}_k} \to X$. If $q^* = (\alpha_{1,0}, \dots, \alpha_{N,0})$ is a strict equilibrium of $\mathfrak{G}$ and $x(t)$ is an interior solution of (ED) which starts sufficiently close to q^*, then, for all $T \leq 0$, we will have*

$$z_{k\mu}(t) \sim z_{k\mu}(0)e^{|T|t} + \Delta u_{k\mu}(q^*)\frac{e^{|T|t} - 1}{|T|}, \tag{31}$$

where, as before, $z_{k\mu} = \frac{\partial h}{\partial x_{k\mu}} - \frac{\partial h}{\partial x_{k,0}}$ are the relative scores of the players' equilibrium actions, $\Delta u_{k\mu} = u_{k\mu} - u_{k,0}$ are the corresponding payoff differences, and we are using the notational convention $(e^{0t} - 1)/0 = t$.

Put differently, we will have $z_{k\mu}(t) = \mathcal{O}(\exp(|T|t))$ for $T < 0$ and $z_{k\mu}(t) = \mathcal{O}(t)$ for $T = 0$: the relative scores $z_{k\mu}$ escape to negative infinity exponentially faster for $T < 0$ than for $T = 0$.

Corollary 12. *If $q^* = (\alpha_{1,0}, \dots, \alpha_{N,0})$ is a strict Nash equilibrium of $\mathfrak{G}$ and $x(t)$ is an interior solution of the temperature-adjusted replicator dynamics (T-RD) that starts close enough to q^*, then*

$$x_{k,0}(t) \sim \begin{cases} 1 - e^{-\mathcal{O}(\exp(|T|t))} & \text{for } T < 0, \\ 1 - e^{-\mathcal{O}(t)} & \text{for } T = 0. \end{cases} \tag{32}$$

By contrast, $\limsup_{t\to\infty} x_{k,0}(t) < 1$ whenever $T > 0$.

Remark 1. It is not too hard to obtain expressions similar to (32) for more general choice functions Q, but the end result is not as concise, so we omit it.

Proof. (Proposition 11.) Pick some $\varepsilon > 0$ and let $x(t)$ start close enough to q^* so that $(1+\varepsilon)\Delta u_{k\mu}(q^*) \leq \Delta u_{k\mu}(x(t)) \leq (1-\varepsilon)\Delta u_{k\mu}(q^*)$; that this is possible follows from the Lyapunov property of strict equilibria established in Theorem 10 (recall also that $\Delta u_{k\mu}(q^*) < 0$ for all $\mu \in \mathcal{A}_{k,0} \equiv \mathcal{A}_k \setminus \{\alpha_{k,0}\}$, $k \in \mathcal{N}$, because q^* is a strict equilibrium of $\mathfrak{G}$). We will thus have:

$$(1+\varepsilon)\Delta u_{k\mu}(q^*) \leq \dot{z}_{k\mu} + T z_{k\mu} \leq (1-\varepsilon)\Delta u_{k\mu}(q^*), \tag{33}$$

so if we multiply by e^{Tt} and integrate, we readily obtain:

$$(1+\varepsilon)\Delta u_{k\mu}(q^*)\frac{e^{Tt}-1}{T} \leq e^{Tt} z_{k\mu}(t) - z_{k\mu}(0) \leq (1-\varepsilon)\Delta u_{k\mu}(q^*)\frac{e^{Tt}-1}{T}, \tag{34}$$

with the convention that $(e^{0t}-1)/0 = t$. Our assertion then follows by rearranging terms in (34) above and noting that ε can be taken arbitrarily small since $z_{k\mu}(t) \to -\infty$ for all $\mu \in \mathcal{A}_{k,0}$, $k \in \mathcal{N}$. □

Proof. (Corollary 12.) For Boltzmann action selection as in (8), we will have $x_{k,0} = \left(1 + \sum_\mu^k \exp(z_{k\mu})\right)^{-1}$ so the estimate (32) follows from (31) and the Taylor expansion $1/(1+s) \sim 1 - s + \mathcal{O}(s^2)$. □

4 Discrete-time learning algorithms and stochastic approximations.

In this section, we examine how the entropy-driven dynamics (ERL) and (ED) may be used to design and implement learning algorithms in the context of finite games that are played repeatedly over time. The main challenge in this endeavor is that in practical implementations, players can only observe the payoffs that they actually received when playing the game (or even only a noisy version thereof), whereas the dynamics (ERL)/(ED) involve the expected payoff functions $u_{k\alpha}(x)$. Therefore, in the absence of perfect monitoring (or any other device permitting the calculation of expected payoffs), any discretization of the dynamics (ERL)/(ED) should involve only the players *in-game* payoff streams and no other information.

A natural way of addressing this issue is to take an Euler-like discretization of the dynamics, use the players' evolving mixed strategies to select an action at each stage, and update only those components for which payoffs were actually observed. In what follows, we will give a brief account of this approach (known as *stochastic approximation*) and then apply it directly to the dynamics (ERL) and (ED).

4.1 Stochastic approximation of continuous dynamics.

For completeness, we recall here a few general elements from the theory of stochastic approximation following Benaïm [5] and Borkar [8]. To begin with, let $\mathcal{S}$ be a finite set, and let $Z(n)$, $n \in \mathbb{N}$ be a $\mathbb{R}^{\mathcal{S}}$-valued stochastic process that satisfies the recursion

$$Z(n+1) = Z(n) + \gamma_{n+1} U(n+1), \tag{35}$$

where γ_n is a sequence of step sizes (usually assumed to vanish with n) and $U(n)$ is another $\mathbb{R}^{\mathcal{S}}$-valued process which is adapted to the filtration $\mathcal{F}$ of Z. Then, if there exists a Lipschitz-continuous vector field $f\colon \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ such that $\mathbb{E}[U(n+1)|\mathcal{F}_n] = f(Z(n))$, we will say that (35) is a *stochastic approximation* of the continuous-time dynamical system

$$\dot{z} = f(z). \tag{MD}$$

More explicitly, if we split the so-called *innovation term* $U(n)$ of (35) into its average value $f(Z(n)) = \mathbb{E}[U(n)|\mathcal{F}_n]$ and a zero-mean noise term $V(n+1) = U(n+1) - f(Z(n))$, (35) may be rewritten as

$$Z(n+1) = Z(n) + \gamma_{n+1}\left(f(Z(n)) + V(n+1)\right), \tag{SA}$$

which is easily seen to be a noisy Euler-like discretization of (MD); conversely, the equation (MD) will be referred to as the *mean dynamics* of the stochastic recursion (SA).

The main goal of the theory of stochastic approximation is to relate the process (SA) with the solution trajectories of the mean dynamics (MD). To that end, the standard assumptions that ensure that this comparison is possible are:

- (A_1) The step sequence γ_n is $(\ell^2 - \ell^1)$–summable (typically, $\gamma_n = 1/n$).
- (A_2) $V(n)$ is a difference of martingales with $\sup_n \mathbb{E}\left[\|V(n)\|^2\right] < \infty$.
- (A_3) The stochastic process $Z(n)$ is bounded: $\sup_n \|Z(n)\| < \infty$ (a.s.).

Under these assumptions, the following lemma ensures that $Z(n)$ can only converge to a connected set of rest points of the corresponding mean dynamics:

Lemma 13. *Assume that the dynamics (MD) admit a strict Lyapunov function (i.e. a real-valued function which decreases along every non-stationary solution orbit of (MD)), and assume further that the set of values of this function at the rest points of (MD) has measure zero in $\mathbb{R}$. Then, under the assumptions (A_1)–(A_3), every accumulation point of the process $Z(n)$ generated by the recursion (SA) belongs to a connected set of rest points of the mean dynamics (MD).*

Proof. Proof. Our claim is a direct consequence of the following string of results in Benaïm [5] (listed in order of successive implications): Proposition 4.2, Proposition 4.1, Theorem 5.7, and Proposition 6.4. □

As an application of the previous lemma, let us consider a game $\mathfrak{G}$ and a stochastic approximation of the entropy-driven dynamics (ED) that satisfies conditions (A_1)–(A_3). If $\mathfrak{G}$ admits a potential function U, Lemma 6 shows that the free entropy $F = Th - U$ is Lyapunov for the entropy-driven dynamics (ERL)/(ED), and Sard's theorem (Lee [17]) ensures that the set of values taken by F at its critical points has measure zero. Thus, in view of Proposition 5, we see that stochastic approximations of (ERL)/(ED) may only converge to connected sets of restricted QRE of $\mathfrak{G}$. In what follows, we will exploit this property in order to derive two entropy-driven learning algorithms based respectively on the score dynamics (ERL) and the strategy-based dynamics (ED).

Remark 1. We should note here that (SA) implicitly assumes that each component of the vector Z is updated simultaneously. In a game theoretic setting, this corresponds to complete player synchronization, an assumption which does not always hold; we will address such issues in Section 4.4.

4.2 Score-based implementation of entropy-driven learning.

We begin by constructing a discrete-time stochastic approximation of the score-based entropic dynamics (ERL) as follows: at each step, players play a smoothed best response to the performance scores of their actions (using the choice map Q defined in Section 2.3), and then update these scores depending on the payoffs they receive from the chosen action. We illustrate this process in Algorithm 1 below (presented in a synchronous version, with players selecting actions and receiving payoffs simultaneously):

Algorithm 1 Score-based learning with entropy-driven action selection.

$n \leftarrow 0$
foreach player $k \in \mathcal{N}$ and action $\alpha \in \mathcal{A}_k$ **do** initialize $Y_{k\alpha}$ and set $X_k \leftarrow Q_k(Y_k)$
Repeat
  $n \leftarrow n+1$
  foreach player $k \in \mathcal{N}$ simultaneously **do**
  select a new action $\hat{\alpha}_k$ according to the mixed strategy X_k # current action
  $\hat{u}_k \leftarrow u_k(\hat{\alpha})$ # current payoff
  $Y_{k\hat{\alpha}_k} \leftarrow Y_{k\hat{\alpha}_k} + \gamma_n(\hat{u}_k - TY_{k\hat{\alpha}_k})/X_{k\hat{\alpha}_k}$ # update current action score
  foreach action $\alpha \in \mathcal{A}_k$ **do** $X_{k\alpha} \leftarrow Q_{k\alpha}(Y_k)$ # update mixed strategy

To study the convergence properties of Algorithm 1, let $Y(n)$ denote the players' score profile at the n-th iteration of the algorithm – and similarly for $X(n)$ (strategies), $\hat{\alpha}(n)$ (actions) and $\hat{u}(n)$ (payoffs). Then, $Y(n)$ is a stochastic process adapted to the filtration $\mathcal{F}_n$ generated by X and satisfies the relation:

$$\mathbb{E}\left[(Y_{k\alpha}(n+1) - Y_{k\alpha}(n))/\gamma_{n+1} \mid \mathcal{F}_n\right] = \mathbb{E}[\hat{u}_k(n+1)|\mathcal{F}_n] - TY_{k\alpha}(n) = u_{k\alpha}(X(n)) - TY_{k\alpha}(n), \tag{36}$$

for all $\alpha \in \mathcal{A}_k$, $k \in \mathcal{N}$. Together with the selection rule $X_k(n) = Q_k(Y_k(n))$, the RHS of the above expression yields the entropy-driven score dynamics (ERL), so the strategy process $X(n)$ generated by Algorithm 1 will be a stochastic approximation of (ERL).

In the special case where $T = 1$, Algorithm 1 boils down to the Q-learning scheme of Leslie and Collins [18] which, under the assumption that $Y(n)$ remains bounded (a.s.), was proven to converge to *Nash distributions* (the analogue of a QRE with rationality level $\varrho = 1$) in various classes of 2-player games. Unfortunately, the *unconditional* convergence of this algorithm still eludes us because assumptions (A_2) and (A_3) are hard to verify: in fact, one can check that the order of the noise term $V(n)$ is

$$\mathbb{E}[\|V(n+1)\|^2|\mathcal{F}_n] = \mathcal{O}((1+\|Y(n)\|^2)e^{\|Y(n)\|}), \tag{37}$$

so (A_2) rests on first establishing the boundedness requirement (A_3).[17] However, establishing (A_3) is not trivial in itself because of the "almost surely" requirement; it seems to be possible to do so thanks to an argument by M. Faure (personal communication), but since Algorithm 1 is not the main focus of our paper, we will not venture further along this direction.

There exists other conditions (replacing Assumptions A_1, A_2 and A_3) under which convergence can be proved. A possible track to prove convergence without requiring *a priori* boundedness of $V(n)$ is to truncate the stochastic approximation with expanding bounds. This approach is used in **?** [**?**], but again, the required summability conditions on $\mathbb{E}[||V(n+1)||^2|\mathcal{F}_n]$ are not satisfied in our case unless $(Y(n))$ is bounded. In any case, our focus is more on the following strategy-based algorithm, where boundedness is guaranteed.

4.3 Strategy-based implementation of entropy-driven learning.

In this section, we will focus on the strategy-based variant of the entropic game dynamics (ED), and we will implement it as a payoff-based learning procedure in discrete time. One of the advantages of this approach is that it does not rely on having a closed form expression for the choice map Q (which is hard to obtain for non-Boltzmann action selection); another is that since the algorithm is strategy-based (and hence its update variables are bounded by default), we will not need to worry too much about satisfying

[17] Note that this is also true for the weaker requirement supplied by Borkar [8], namely that there exist K such that $\mathbb{E}[||V(n+1)||^2|\mathcal{F}_n] \leq K(1+||Y_n||^2)$ for all n.

conditions (A_2) and (A_3) as in the case of Algorithm 1. As a result, the strategy-based implementation of the entropic game dynamics (ED) will be significantly easier to handle than its score-based variant.

Without further ado, we have:

Algorithm 2 Strategy-based implementation of entropy-driven learning.

$n \leftarrow 0$
foreach player $k \in \mathcal{N}$ **do** initialize X_k as a mixed strategy with full support
Repeat
  $n \leftarrow n + 1$
  foreach player $k \in \mathcal{N}$ simultaneously **do**
  select a new action $\hat{\alpha}_k$ according to the mixed strategy X_k # current action
  foreach player $k \in \mathcal{N}$ **do**
    $\hat{u}_k \leftarrow u_k(\hat{\alpha}_1, \dots, \hat{\alpha}_N)$ # current payoff
    foreach action $\alpha \in \mathcal{A}_k$ **do** # update mixed strategy

$$X_{k\alpha} \leftarrow X_{k\alpha} + \frac{\gamma_n}{\Theta''(X_{k\alpha})}\left[\frac{\hat{u}_k}{X_{k\hat{\alpha}_k}}\left(\mathbb{1}(\hat{\alpha}_k = \alpha) - \frac{\theta_h''(X_k)}{\theta''(X_{k\hat{\alpha}_k})}\right) - T g_{k\alpha}(X)\right] \tag{38}$$

    where $g_{k\alpha}(X) \equiv \theta'(X_{k\alpha}) - \theta_h''(X_k)\sum_\beta^k \theta'(X_{k\beta})/\theta''(X_{k\beta})$ is the entropy adjustment term of (ED).

As stated above, the update step of Algorithm 2 has been designed so as to track the entropic dynamics (ED). Indeed, letting $X(n)$ (resp. $\hat{\alpha}(n)$, $\hat{u}(n)$) denote the players' strategy (resp. action, payoff) profile at the n-th iteration of the algorithm, we will have for all $\alpha \in \mathcal{A}_k$, $k \in \mathcal{N}$:

$$\mathbb{E}\left[(X_{k\alpha}(n+1) - X_{k\alpha}(n))/\gamma_{n+1} \mid \mathcal{F}_n\right] = \frac{u_{k\alpha}(X(n))}{X_{k\alpha}(n)\theta''(X_{k\alpha}(n))} - \frac{\theta''(X_{k\beta}(n))}{\theta''(X_{k\alpha}(n))}\sum_\beta^k \frac{u_{k\beta}(X(n))}{\theta''(X_{k\beta}(n))} - T\frac{g_{k\alpha}(X(n))}{\theta''(X_{k\alpha}(n))}, \tag{39}$$

which is simply the RHS of the entropy-driven dynamics (ED) evaluated at $X(n)$.

Remark 1. In the case of the Boltzmann-Gibbs kernel $\theta(x) = x\log x$, the update step (38) becomes

$$X_{k\alpha} \leftarrow X_{k\alpha} + \gamma_{n+1}\left[(\mathbb{1}(\hat{\alpha}_k = \alpha) - X_{k\alpha})\cdot\hat{u}_k - T X_{k\alpha}\left(\log X_{k\alpha} - \textstyle\sum_\beta^k X_{k\beta}\log X_{k\beta}\right)\right]. \tag{40}$$

For zero temperatures, we thus obtain the reinforcement learning scheme of Sastry et al. [29] that was based on the classical replicator equation (RD).

On the other hand, unlike Algorithm 1 (which was evolving in $\mathbb{R}^{\mathcal{A}}$), Algorithm 2 is well-defined only if the iterates X are admissible mixed strategies at each update step. To check that this indeed the case, note first that the second term of the RHS of (38) vanishes when summed over $\alpha \in \mathcal{A}_k$, so $\sum_\alpha^k X_{k\alpha}(n)$ remains constant and equal to 1 (recall that $X_k(0)$ is initialized as a valid probability distribution). It thus suffices to check that $X_{k\alpha}(n) \geq 0$ for all $\alpha \in \mathcal{A}_k$; restricting ourselves to positive learning temperatures $T > 0$ and normalizing the game's payoffs to $[0, 1]$ for simplicity, we have:

Lemma 14. *Let θ be a regular entropy kernel such that $x\theta''(x) \geq m$ for some $m > 0$ and for all $x \in (0, 1)$. Then, for $T > 0$ and normalized payoffs $u_k : \mathcal{A} \to [0, 1]$, there exists a positive constant $K > 0$ (which only depends on T and θ) such that if the step sequence γ_n is bounded from above by K, then $X_{k\alpha}(n) \geq 0$ for all $\alpha \in \mathcal{A}_k$, $k \in \mathcal{N}$, and for all $n \geq 0$.*

Proof. Proof. For notational simplicity, we will only consider here the single-player case, the general case being similar. To that end, we first claim that the entropic term g_k of (38) is bounded by a constant C_θ; indeed:

- θ' is increasing, hence $\theta'(\xi) \leq \theta'(1)$ for all $\xi \in (0,1)$.
- For all $x \in X \equiv \Delta(\mathcal{A})$, $\sum_\beta 1/\theta''(x_\beta) \geq \max_\beta 1/\theta''(x_\beta)$, so $\theta''_h(x) \leq \min_\beta \theta''(x_\beta) \leq \max\{\theta''(\xi) : \operatorname{card}(\mathcal{A})^{-1} \leq \xi \leq 1\}$. Thus there exists $\theta''_{h,\max}$ such that $\theta''_{h,\max} \geq \theta''_h(x) > 0$ for all $x \in X$.
- By the regularity assumption for θ, we will have $\theta'(\xi)/\theta''(\xi) \to 0$ as $\xi \to 0^+$, so there exists $M > 0$ such that $|\theta'(\xi)/\theta''(x)| < M$ for all $\xi \in (0,1)$.

As a result, our claim follows by taking $C_\theta = \theta'(1) + \theta''_{h,\max} \operatorname{card}(\mathcal{A}) M$.

Now, letting $\hat{\alpha}$ be the chosen action at step n and $\hat{u}$ the corresponding payoff, we will have:

$$\begin{aligned}
&\frac{1}{x_\alpha \theta''(x_\alpha)} \left[T g_\alpha(x) - \hat{u}/x_{\hat{\alpha}} \left(\mathbb{1}(\hat{\alpha} = \alpha) - \theta''_h(x)/\theta''(x_{\hat{\alpha}}) \right) \right] \\
&\qquad \leq \frac{1}{x_\alpha \theta''(x_\alpha)} \left[T C_\theta + \hat{u}/x_{\hat{\alpha}} \theta''(x_{\hat{\alpha}}) \left(\theta''_h(x) - \mathbb{1}(\hat{\alpha} = \alpha) \theta''(x_{\hat{\alpha}}) \right) \right] \\
&\qquad \leq \frac{1}{x_\alpha \theta''(x_\alpha)} \left[T C_\theta + \theta''_{h,\max} \cdot \hat{u} \left(x_{\hat{\alpha}} \theta''(x_{\hat{\alpha}}) \right)^{-1} \right] \leq m^{-1} \left(T C_\theta + m^{-1} \theta''_{h,\max} \right), \quad (41)
\end{aligned}$$

where we used the assumption $T \geq 0$ for the first inequality and the normalization $\hat{u} \in [0,1]$ for the second and third one. Hence, if $\gamma_n \leq m^{-1} \left(T C_\theta + m^{-1} \theta''_{h,\max} \right) \equiv K$, we will have $X_\alpha(n+1) \geq X_\alpha(n)(1 - \gamma_{n+1} K) \geq 0$ whenever $X_\alpha(n) \geq 0$, and the induction is complete. □

Under the assumptions of Lemma 14 above, Algorithm 2 is well-defined and is no risk of crashing; we now show that if $\mathfrak{G}$ is a potential game, then Algorithm 2 converges to a QRE of $\mathfrak{G}$ almost surely:

Theorem 15. *Let $\mathfrak{G}$ be a potential game, and let θ be a regular entropy kernel such that $x\theta''(x) \geq m$ for some $m > 0$ and for all $x \in (0,1)$. Then, for $T > 0$ and sufficiently small step sizes γ_n satisfying (A_1), Algorithm 2 converges almost surely to a connected set of QRE of $\mathfrak{G}$ with rationality level $1/T$.*

Proof. Proof. By the proof of Lemma 14, assumptions (A_2) and (A_3) for the iterates $X(n)$ of Algorithm 2 are verified immediately – simply note that the innovation term of (38) is bounded by the constant K of Lemma 14. Thus, by Lemma 13 and the subsequent discussion, it follows that $X(n)$ converges to a connected set of *restricted* QRE of $\mathfrak{G}$.

We will now show that the accumulation points of $X(n)$ can only lie in the relative interior $\operatorname{rel\,int}(X)$ of X. To that end, let $z_{k\mu} = \theta'(x_{k\mu}) - \theta'(x_{k,0})$ be the reduced score variables of (13) and let $W_{k\mu} = \Delta u_{k\mu} - T z_{k\mu}$ denote the RHS of the reduced score dynamics (ZD), viz. $\dot{z} = W(z)$. Then, following Borkar and Meyn [7], the *limiting ODE* of the dynamics (ZD) will be:

$$\dot{z} = W_\infty(z) \equiv \lim_{r \to \infty} W(rz)/r = \lim_{t \to \infty} \frac{\Delta u_{k\mu}(Q_0(rz)) - Trz}{r} = -Tz, \qquad (\mathrm{ZD}_\infty)$$

where $Q_0 \colon z \mapsto x$ is the reduced choice map of Eq. (15). For $T > 0$, the origin is a global attractor of (ZD_∞), so Theorem 2.1 in Borkar and Meyn [7] implies that the discrete stochastic approximation $Z_{k\mu}(n) = \theta'(X_{k\mu}(n)) - \theta'(X_{k,0}(n))$ of (ZD) will be bounded almost surely. Moreover, given that Q_0 is a homeomorphism onto $\operatorname{rel\,int}(X)$, the image of any compact subset of $\prod_k \mathbb{R}^{\mathcal{A}_{k,0}}$ will be a compact subset of $\operatorname{rel\,int}(X)$, so any accumulation point of the process $X = Q_0(Z)$ will lie in $\operatorname{rel\,int}(X)$, as claimed. Since $X(n)$ was shown to converge to a connected set of restricted QRE of $\mathfrak{G}$, our assertion follows. □

Remark 1. It is important to note here that Theorem 15 holds for any $T > 0$, so Algorithm 2 may be tuned to converge to QRE with arbitrarily high rationality level $\varrho = 1/T$ – and hence, arbitrarily close to the game's strict Nash equilibria; cf. the discussion following Theorem 10 in Section 3. In this way, Theorem 15 is different in scope than the convergence results of Cominetti et al. [10] and Bravo [9]: instead of taking high learning temperatures to guarantee a unique QRE, players who employ low learning temperatures may converge arbitrarily close to the game's strict equilibria.

Remark 2. In view of the above, one might hope that Algorithm 2 would still converge to the game's (strict) Nash equilibria even for $T = 0$. In that case however, the limiting ODE (ZD_∞) no longer admits a global attractor at the origin, so the relative scores $Z_{k\mu}$ no longer remain bounded and we cannot discount the convergence of Algorithm 2 to non-Nash vertices of X. In fact, even in the simplest possible case of a single player game with two actions, Lamberton et al. [15] showed that the $T = 0$ version of Algorithm 2 with Boltzmann action selection fails to converge a.s. to a Nash equilibrium for step sequences of the form $\gamma_n = 1/n^h$, $0 < h < 1$.

4.4 Robustness of the strategy-based learning algorithm.

Even though Algorithm 2 only requires players to observe and record their in-game payoffs, it still relies on the following assumptions: *a*) that players have perfect measurements of their payoffs (or that the same action profile always yields the exact same payoffs); *b*) that players all play at the same time; and *c*) that there is no delay between playing and receiving payoffs. Since these assumptions are often violated in practical scenarios, we devote the rest of this section to examining the robustness of Algorithm 2 in this more general setting.

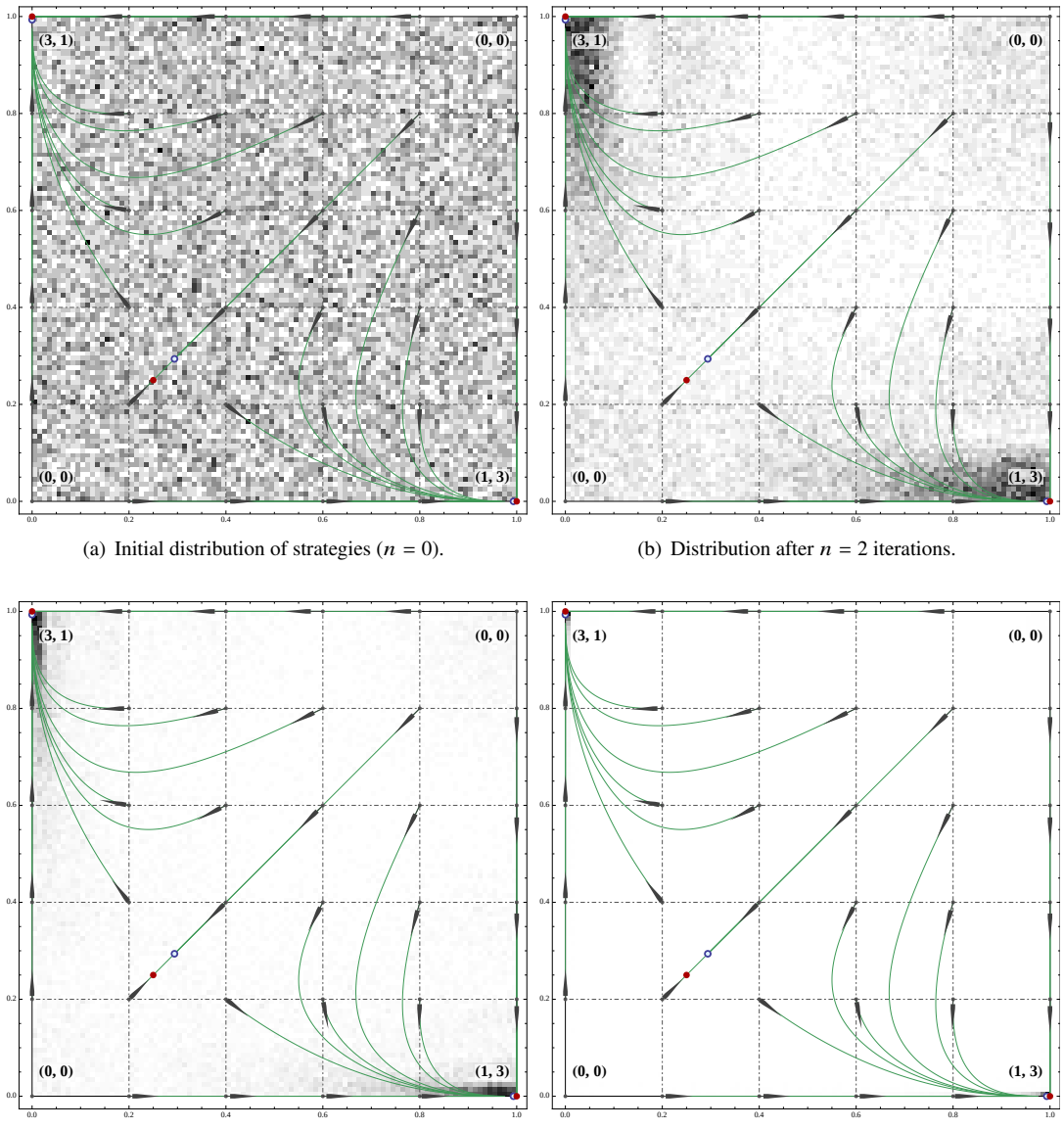
Noisy measurements and stochastic perturbations. In many real-world applications of game theory (and especially in traffic congestion games), the payoffs received by the players at each stage of the game may be subject to random shocks (Hofbauer and Sandholm [12], Mertikopoulos and Moustakas [23]) or players may only be able to get a rough measurement of their true payoffs (see e.g. **?** [**?**] for an example drawn from revenue sharing and mechanism design).

We will model such stochastic perturbations by considering a random noise term $\xi_k(n)$, $k \in \mathcal{N}$, which is added to the payoff $\hat{u}_k(n)$ received by player k at the n-th iteration of Algorithm 2 (and which, in principle, might depend on the players' strategy or action profiles). Then, with notation as before, we have:

Proposition 16. *Let $\xi(n)$ be an $\mathcal{F}_n$-adapted difference of martingale with values in $\mathbb{R}^{\mathcal{N}}$ (i.e. $\mathbb{E}[\xi(n+1)|\mathcal{F}_n] = 0$); assume further that ξ_k is bounded for all $k \in \mathcal{N}$ (a.s.) and that ξ_k is stochastically independent of the chosen action $\hat{\alpha}_k$ of player k. Then, the conclusion of Theorem 15 still holds when the payoff stream $\hat{u}_k$ in Algorithm 2 is replaced by the perturbed process $\tilde{u}_k = \hat{u}_k + \xi_k$.*

Proof. Proof. Since the noise process ξ is bounded (a.s.), we can still assume that the perturbed payoff functions are normalized in $[0, 1]$; hence, by taking a sufficiently small step sequence, Algorithm 2 remains well-defined (i.e. $X(n) \in X$ for all n). It thus suffices to check that the conditional expectation of the innovation term of (38) with $\hat{u}_k$ replaced by $\tilde{u}_k = \hat{u}_k + \xi_k$ still yields the entropy-driven dynamics (ED_θ). That this is so, is an immediate consequence of the independence between ξ_k and $\hat{\alpha}_k$; indeed:

$$\begin{aligned}\mathbb{E}[\tilde{u}_k(n+1)\,\mathbb{1}(\hat{\alpha}_k(n+1) = \alpha)|\mathcal{F}_n] &\\ = \mathbb{E}\left[(\hat{u}_k(n+1) + \xi_k(n+1))\,\mathbb{1}(\hat{\alpha}_k(n+1) = \alpha)|\mathcal{F}_n\right] &\\ = u_{k\alpha}(X(n)) + \mathbb{E}\left[\xi_k(n+1)|\mathcal{F}_n\right] X_{k\alpha}(n) = u_{k\alpha}(X(n)),& \quad (42)\end{aligned}$$



(a) Initial distribution of strategies ($n = 0$).

(b) Distribution after $n = 2$ iterations.

(c) Distribution after $n = 5$ iterations.

(d) Distribution after $n = 10$ iterations.

Figure 2: Snapshots of the evolution of of Algorithm 2. In our simulations, we drew 10^4 random initial strategies in the potential game of Fig. 1 and, for each strategy allocation, we ran the Boltzmann variant of Algorithm 2 with learning temperature $T = 0.2$ and step sequence $\gamma_n = 1/(5 + n^{0.6})$. In each figure, the shades of gray represent the normalized density of states at each point of the game's strategy space and we also drew the phase portraits of the underlying mean dynamics (ED) for convenience. We see that Algorithm 2 converges to the game's QRE (which, for $T = 1/\varrho = 0.2$ are very close to the game's strict equilibria) quite rapidly: after only $n = 10$ iterations, more than 99% of the initial strategies have converged within $\varepsilon = 0.01$ of the game's equilibria.

where the second equality follows from the independence of ξ and $\hat{\alpha}$, and the last one stems from the fact that $\xi(n)$ is an $\mathcal{F}_n$-adapted difference of martingale. The entropic dynamics (ED) are then obtained as in (39). □

Remark 1. We should note here that the assumptions on the noise term $\xi(n)$ of Proposition 16 are rather mild: they encompass not only the case where the perturbations are independent of the players' choices (a case which has attracted significant interest in the literature by itself), but also scenarios where $\xi(n+1)$ might depend on the entire history of the game up to stage n, or even when the noise stems from imperfect observations of the other players' *current* actions (i.e. at the $(n+1)$-th stage of the game).

Asynchronous updates and delays. In Algorithm 2, it is assumed that players update their strategies at every iteration *simultaneously*, i.e. they adhere to a *synchronous* strategy revision process. On the other hand, in congestion games and applications (e.g. in wireless networks), it is often the case that revisions are *asynchronous*: for instance, if we consider a set of wireless users communicating with a slotted ALOHA base station [1], then each user's decision to transmit or remain silent at a given timeslot is not coordinated with other users, so updates and strategy revisions occur at different periods for each user. Furthermore, in the same scenario, variable message propagation delays often mean that the outcome of a user's choice does not depend on the choices of other users at the current timeslot, but on their choices in previous timeslots.

In view of the above, the first extension to Algorithm 2 that we will consider here is the case where only a random subset of players (possibly of cardinality 1) revises their strategies at a given iteration of the algorithm. To that end, let $R_n \subseteq 2^{\mathcal{N}}$ be the random set of players who update their strategies at the n-th iteration of the algorithm. In practice, players are not aware of the global iteration counter n but can only know the number of updates that they have carried out up to time n, as measured by the random variables $\phi_k(n) \equiv \text{card}\{m \leq n : k \in R_m\}$, $k \in \mathcal{N}$. Accordingly, the asynchronous variant of Algorithm 2 that we will consider consists of replacing the instruction "for each player $k \in \mathcal{N}$" by "for each player $k \in R_n$" and replacing "n" by "$\phi_k(n)$" in the step-size computation.

Furthermore, as noted above, another natural extension of Algorithm 2 consists of allowing the (possibly perturbed) payoffs perceived by the players to be subject to delays. Formally, let $\tau_{k\alpha}(n)$ be the (integer-valued) delay that player k experiences when playing his α-th action at step n. Then, the payoff $\hat{u}_k$ of player k in (38) at step n should be replaced by $u_k(\alpha_k(n); \alpha_{-k}(n - \tau_{k\alpha}(n)))$, with expected value conditioned on the history $\mathcal{F}_n$ given by $u_{k,\alpha_k(n)}(X(n - \tau_{k\alpha}(n)))$.

Following Chapter 7 of Borkar [8], we will make the following assumptions regarding these two extensions of Algorithm 2:

1. The step sequence is of the form $\gamma_n = K/n$, where K is a positive constant small enough to guarantee that Algorithm 2 remains well-defined for all n.

2. The strategy revision process R_n is a homogeneous ergodic Markov chain over $2^{\mathcal{N}}$. We denote by λ_k the asymptotic rate at which player k updates its strategy (if μ is the stationary distribution, then $\lambda_k = \sum_{A \subseteq 2^{\mathcal{N}} : k \in A} \mu(A)$)

3. The delays $\tau_{k\alpha}(n)$ are bounded, i.e. there exists M such that for every n, $0 \leq \tau_{k\alpha}(n) \leq M$ (a.s.). This condition ensures that the bias induced by the delay becomes negligeable in face of the step-size sequence as times goes by.

These hypotheses are rather mild in themselves, but they can be weakened even further at the expense of presentational simplicity and clarity (see e.g. Chapter 7 of Borkar [8]). Still and all, in this context, we have:

Proposition 17. *Under the previous assumptions, the conclusion of Theorem 15 still holds for the variant of Algorithm 2 with asynchronous updates and payoff delays.*

Proof. Proof. By Theorems 2 and 3 in Chapter 7 of Borkar [8], the algorithm modified to account for asynchronous strategy revisions and payoff delays as above, will be a stochastic approximation of the rate-adjusted dynamics

$$\dot{x}_k = \lambda_k \, \mathrm{ED}_\theta(x_k), \tag{43}$$

where λ_k is the mean rate at which player k updates its strategy, and ED_θ denotes the RHS of the entropy-driven dynamics (ED_θ). In general, the revision rate λ will depend on time (leading to a non-autonoumous dynamical system), but given that the revision process R_n is a homogeneous ergodic Markov chain, λ_k will be equal to the (constant) probability of including player k at the revision set R_n at the n-th iteration of the algorithm. These dynamics have the same rest points as (ED) and an easy calculation shows that the free entropy $F(x) = Th(x) - U(x)$ of 6 remains a strict Lyapunov function for (43), so the proof of Theorem 15 goes through unchanged. □

Remark 1. It is important to note here that the entropic dynamics (43) adjusted for different strategy revision rates are equivalent to the choice-adjusted dynamics (ED_λ) which correspond to players using a different inverse choice temperature (hence the identical notation). Therefore, if the players' revision process is a homogeneous ergodic Markov chain, their mean revision rates $\lambda_k \in (0, 1)$ may also be viewed as inverse choice temperatures of players who never miss a revision opportunity, but who tone down their actions' performance scores by playing the mixed strategy $x_k = Q(\lambda_k y_k)$ instead of $Q(y_k)$.[18]

4.5 Algorithm 2 in practice.

Let us give a close look at Algorithm 2 to assess its interest from an engineering perspective.

- First, it should be clear that the algorithm is highly distributed. The information needed to update one player's strategy is the payoff of its chosen action, which does not depend on any assessment of alternative choices. In particular, it does not rely on the observation of other players action, or even on the knowledge of the set of other players. Another feature is the fact that no synchronization between players is required: As shown in Section **??**, each player can choose its updates instants in a complete independent fashion.

- The algorithm is robust to imperfection in the measurement (if any) of the payoff. This measurement can be based on old actions of remote players and it can suffer from random fluctuations (for example coming from a bad instrument and/or perturbations from the environment).

- The temperature parameter T should be rather easy to tune: In potential games, it should be taken strictly positive to ensure convergence to quantal response equilibria. Of course, small temperature values give better approximation of NE because QRE gets closer and closer to Nash equilibria when the temperature goes to 0. But, small temperatures may alter the speed of convergence because the step-sizes have to be chosen very small.The optimal choice of the temperature giving a good compromise between accuracy and speed of convergence is problem dependent.

All this leads to a distributed, player centric version of Algorithm 2, given in Algorithm 3, where each player k is equipped with a clock and plays each time its clock rings. The clock rings $(\tau_n^k)_{n\in\mathbb{N}}$ form an infinite increasing sequence of integers with positive rate (*i.e.* for any n, $\frac{n}{\tau_n^k} > c \geq 0$). The corresponding revision set at time n, R_n, is defined by: $R_n = \{k | n \in (\tau_n^k)_{n\in\mathbb{N}}\}$.

[18] Note also that $\lambda_k < 1$ so players who do not update all the time tend to choose actions in a more uniform manner.

Algorithm 3 Algorithm used by one player (say player k)

$n \leftarrow 0$

Repeat

At time τ_{n+1}^k

$n \leftarrow n+1$

select a new action $\hat{\alpha}_k$ according to the mixed strategy X_k # current action

$\hat{u}_k \leftarrow u_k(\hat{\alpha}_1, \ldots, \hat{\alpha}_N)$ # compute (or measure) current payoff

foreach action $\alpha \in \mathcal{A}_k$ **do** # update mixed strategy

$$X_{k\alpha} \leftarrow X_{k\alpha} + \frac{\gamma_n}{\Theta''(X_{k\alpha})} \left[\frac{\hat{u}_k}{X_{k\hat{\alpha}_k}} \left(\mathbb{1}(\hat{\alpha}_k = \alpha) - \frac{\theta_h''(X_k)}{\theta''(X_{k\hat{\alpha}_k})} \right) - T g_{k\alpha}(X) \right]$$

References

[1] Altman, E., T. Boulogne, R. el Azouzi, T. Jiménez, and L. Wynter, 2006: A survey on networking games in telecommunications. *Computers and Operations Research*, **33 (2)**, 286–311.

[2] Alvarez, F., J. Bolte, and O. Brahic, 2004: Hessian Riemannian gradient flows in convex programming. *SIAM Journal on Control and Optimization*, **43 (2)**, 477–501.

[3] Anderson, S. P., A. de Palma, and J.-F. Thisse, 1992: *Discrete Choice Theory of Product Differentiation*. MIT Press, Cambridge, MA.

[4] Auslender, A., R. Cominetti, and M. Haddou, 1997: Asymptotic analysis for penalty and barrier methods in convex and linear programming. *Mathematics of Operations Research*, **22**, 43–62.

[5] Benaïm, M., 1999: Dynamics of stochastic approximation algorithms. *Séminaire de probabilités de Strasbourg*, **33**.

[6] Björnerstedt, J. and J. W. Weibull, 1996: Nash equilibrium and evolution by imitation. *The Rational Foundations of Economic Behavior*, K. J. Arrow, E. Colombatto, M. Perlman, and C. Schmidt, Eds., St. Martin's Press, New York, NY, 155–181.

[7] Borkar, V. and S. Meyn, 2000: The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, **38 (2)**, 447–469.

[8] Borkar, V. S., 2008: *Stochastic approximation*. Cambridge University Press and Hindustan Book Agency.

[9] Bravo, M., 2011: An adjusted payoff-based procedure for normal form games, http://arxiv.org/pdf/1106.5596.pdf.

[10] Cominetti, R., E. Melo, and S. Sorin, 2010: A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior*, **70**, 71–83.

[11] Fudenberg, D. and D. K. Levine, 1998: *The Theory of Learning in Games*, Economic learning and social evolution, Vol. 2. The MIT Press, Cambridge, MA.

[12] Hofbauer, J. and W. H. Sandholm, 2002: On the global convergence of stochastic fictitious play. *Econometrica*, **70 (6)**, 2265–2294.

[13] Hofbauer, J. and K. Sigmund, 1998: *Evolutionary Games and Population Dynamics*. Cambridge University Press.

[14] Hofbauer, J., S. Sorin, and Y. Viossat, 2009: Time average replicator and best reply dynamics. *Mathematics of Operations Research*, **34 (2)**, 263–269.

[15] Lamberton, D., G. Pagès, and P. Tarrès, 2004: When can the two-armed bandit algorithm be trusted? *The Annals of Applied Probability*, **14 (3)**, 1424–1454.

[16] Landau, L. D. and E. M. Lifshitz, 1976: Statistical physics. *Course of Theoretical Physics*, Pergamon Press, Oxford, Vol. 5.

[17] Lee, J. M., 2003: *Introduction to Smooth Manifolds*. No. 218 in Graduate Texts in Mathematics, Springer-Verlag, New York, NY.

[18] Leslie, D. S. and E. J. Collins, 2005: Individual Q-learning in normal form games. *SIAM Journal on Control and Optimization*, **44 (2)**, 495–514.

[19] Marsili, M., D. Challet, and R. Zecchina, 2000: Exact solution of a modified El Farol's bar problem: Efficiency and the role of market impact. *Physica A*, **280**, 522–553.

[20] McFadden, D. L., 1981: Econometric models of probabilistic choice. *Structural Analysis of Discrete Data with Econometric Applications*, C. F. Manski and D. L. McFadden, Eds., MIT Press, Cambridge, MA, 198–272.

[21] McKelvey, R. D. and T. R. Palfrey, 1995: Quantal response equilibria for normal form games. *Games and Economic Behavior*, **10 (6)**, 6–38.

[22] Mertikopoulos, P., E. V. Belmega, and A. L. Moustakas, 2012: Matrix exponential learning: Distributed optimization in MIMO systems. *ISIT '12: Proceedings of the 2012 IEEE International Symposium on Information Theory*.

[23] Mertikopoulos, P. and A. L. Moustakas, 2010: The emergence of rational behavior in the presence of stochastic perturbations. *The Annals of Applied Probability*, **20 (4)**, 1359–1388.

[24] Monderer, D. and L. S. Shapley, 1996: Potential games. *Games and Economic Behavior*, **14 (1)**, 124 – 143.

[25] Ritzberger, K. and J. W. Weibull, 1995: Evolutionary selection in normal-form games. *Econometrica*, **63**, 1371–99.

[26] Rockafellar, R. T., 1970: *Convex Analysis*. Princeton University Press, Princeton, NJ.

[27] Rustichini, A., 1999: Optimal properties of stimulus-response learning models. *Games and Economic Behavior*, **29**, 230–244.

[28] Sandholm, W. H., 2011: *Population Games and Evolutionary Dynamics*. Economic learning and social evolution, MIT Press, Cambridge, MA.

[29] Sastry, P. S., V. V. Phansalkar, and M. A. L. Thathachar, 1994: Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information. *IEEE Trans. Syst., Man, Cybern.*, **24 (5)**, 769–777.

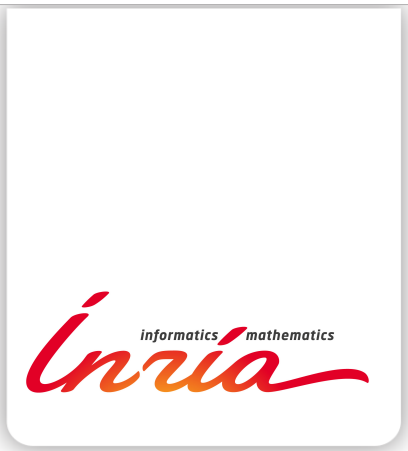
[30] Sorin, S., 2009: Exponential weight algorithm in continuous time. *Mathematical Programming*, **116 (1)**, 513–528.

[31] Taylor, P. D. and L. B. Jonker, 1978: Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, **40 (1-2)**, 145–156.

[32] Tuyls, K., P. J. 't Hoen, and B. Vanschoenwinkel, 2006: An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems*, **12**, 115–153.

[33] van Damme, E., 1987: *Stability and perfection of Nash equilibria*. Springer-Verlag, Berlin.

[34] Weibull, J. W., 1995: *Evolutionary Game Theory*. MIT Press, Cambridge, MA.



RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399