REGULARIZED BEST RESPONSES AND REINFORCEMENT LEARNING IN GAMES

PANAYOTIS MERTIKOPOULOS AND WILLIAM H. SANDHOLM

Abstract. We investigate a class of reinforcement learning dynamics in which each player plays a "regularized best response" to a score vector consisting of his actions' cumulative payoffs. Regularized best responses are single-valued regularizations of ordinary best responses obtained by maximizing the difference between a player's expected cumulative payoff and a (strongly) convex penalty term. In contrast to the class of smooth best response maps used in models of stochastic fictitious play, these penalty functions are not required to be infinitely steep at the boundary of the simplex; in fact, dropping this requirement gives rise to an important dichotomy between steep and nonsteep cases. In this general setting, our main results extend several properties of the replicator dynamics such as the elimination of dominated strategies, the asymptotic stability of strict Nash equilibria and the convergence of time-averaged trajectories to interior Nash equilibria in zero-sum games.

Contents

2010 *Mathematics Subject Classification.* Primary 91A26, 37N40; secondary 91A22, 90C25, 68T05.

Key words and phrases. Bregman divergence; dominated strategies; equilibrium stability; Fenchel coupling; penalty functions; projection dynamics; regularized best responses; replicator dynamics; time averages.

The authors are grateful to Josef Hofbauer, Joon Kwon, Rida Laraki, and seminar audiences in the Hausdorff Research Institute for Mathematics and the Economics Department of the University of Wisconsin–Madison for many interesting discussions.

Part of this work was carried out during the authors' visit to the Hausdorff Research Institute for Mathematics at the University of Bonn in the framework of the Trimester Program "Stochastic Dynamics in Economics and Finance". PM is grateful for financial support from the French National Research Agency under grant nos. ANR–GAGA–13–JS01–0004–01 and ANR–NETLEARN–13–INFR–004, and the CNRS grant PEPS–GATHERING–2014. WHS is grateful for financial support under NSF Grant SES–1155135.

1. Introduction

"Reinforcement learning" has become a catch-all term for learning in recurring decision processes where the agents' future choice probabilities are shaped by information about past payoffs. In game theory and online optimization, most work under this name has focused on multi-armed bandit problems, games against Nature (adversarial or otherwise), or simultaneous choices by strategically interacting players – for a panoramic introduction, see Sutton and Barto [54]. Accordingly, reinforcement learning in games typically revolves around discrete-time stochastic processes with the stochasticity arising at least in part from the agents' randomized choices – see Börgers and Sarin [10], Erev and Roth [16], Fudenberg and Levine [20], Freund and Schapire [18], Hopkins [28], Hart and Mas-Colell [23], Beggs [6], Leslie and Collins [37], Cominetti et al. [13], Coucheney et al. [14] and many others.

A key approach to analyzing these processes is the ordinary differential equation (ODE) method of stochastic approximation which relates the behavior of the stochastic model under study to that of an ODE describing its mean behavior (Benaïm [7]). Motivated by the success of this approach, we follow Sorin [53] and Hofbauer et al. [27] by specifying a reinforcement learning scheme directly in continuous time. By so doing, we are then able to focus squarely on the deep relations between reinforcement learning, convex analysis and population dynamics.[1]

Within this framework, our starting point is the following continuous-time *exponential learning* process: each player maintains a vector of performance scores representing his actions' cumulative payoffs, and these scores are converted into mixed strategies using a logit rule which assigns choice probabilities proportionally to the exponential of each action's score.[2] According to a well-known derivation, this logit rule amounts to each player maximizing the difference between his expected score and a penalty term given by the (negative) Gibbs entropy of the chosen mixed strategy. Moreover, under this learning process, it is well known (Rustichini [49]) that mixed strategies evolve according to the replicator dynamics of Taylor and Jonker [55], a fundamental model from evolutionary game theory.

We extend this approach by allowing for more general *regularized best response maps* obtained by replacing the Gibbs entropy with an arbitrary strongly convex penalty function. These maps include those considered in models of stochastic fictitious play (Fudenberg and Levine [20], Hofbauer and Sandholm [24]) which are generated by penalty functions that become infinitely steep at the boundary of the simplex. In our context, this steepness assumption ensures that mixed strategies follow an ODE contained in the interior of the simplex – in the Gibbs case, this is simply the replicator equation of Taylor and Jonker [55]. More generally, we find that the evolution of mixed strategies agrees with a class of evolutionary game dynamics studied by Hofbauer and Sigmund [25], Hopkins [29] and Harper [22], and which we study further in a companion paper (Mertikopoulos and Sandholm [41]).

[1] One could use the results developed in this paper to analyze discrete-time learning schemes as in Leslie and Collins [37], Sorin [53], Coucheney et al. [14] and others, but we do not pursue this direction here.

[2] The literature refers to versions of this procedure under a variety of names, including weighted/multiplicative majority algorithm (Littlestone and Warmuth [38], Freund and Schapire [18]), exponential weight algorithm (Sorin [53], Hofbauer et al. [27]), and Boltzmann Q-learning (Leslie and Collins [37], Tuyls et al. [57]).

Moving beyond these steep cases, our model also allows for penalty functions that are (sub)differentiable over the entire simplex without becoming infinitely steep at the boundary. The basic example here is the squared Euclidean distance under which choice probabilities are determined by a closest point projection rule (see Example 2.2 below). As opposed to logit choice, the image of this choice map is the entire simplex (rather than its interior), so trajectories of play may enter and exit the boundary of the simplex in perpetuity. As we argue below, the orbits of this *projected reinforcement learning* process are solutions (in an extended sense) of the projection dynamics of Friedman [19], another basic model of population dynamics from evolutionary game theory. However, this last equation only holds force over intervals of time where the support of the players' mixed strategies remains constant; it does not govern the times or the manner in which this support might change.

Our main results extend a variety of basic properties of the replicator dynamics to the class of reinforcement learning dynamics under study, allowing for both steep and nonsteep penalty functions. First, we show that (iteratively) dominated strategies become extinct and we compute their rates of extinction.[3] Second, we extend several stability and convergence properties of Nash equilibria under the replicator dynamics: namely, we show that *a*) limits of (interior) trajectories are Nash; *b*) Lyapunov stable states are Nash; and *c*) strict Nash equilibria are asymptotically stable. Finally, we show that the basic properties of the time-averaged replicator dynamics (convergence to interior equilibria in zero-sum games and asymptotic agreement with the long run behavior of the best response dynamics) also extend to the class of learning dynamics studied here.

In the paper most closely related to this one, Coucheney et al. [14] (see also Leslie and Collins [37] and Tuyls et al. [57]) consider a reinforcement learning process in which players choose smooth best responses to vectors of exponentially discounted payoff estimates. Focusing exclusively on steep penalty functions, they investigate the convergence of this process in a stochastic, discrete-time environment where players can only observe their realized payoffs. To do so, the authors also examine the Lyapunov and asymptotic stability properties of (perturbed) Nash equilibria under the resulting mean dynamics in continuous time. The no-discounting limit of these dynamics is precisely the steep version of the dynamics studied in the current paper, so our stability and convergence results can be seen as an extension of the analysis of Coucheney et al. [14] to the nonsteep regime. Also, Coucheney et al. [14] do not examine the elimination of dominated strategies or the long-term behavior of empirical frequencies of play, so our results here provide an indication of other properties that could be expected to hold in a discrete-time, stochastic setting.

From the point of view of convex programming, the reinforcement learning dynamics we consider here can be seen as a multi-agent, continuous-time analogue of the well-known mirror descent (MD) optimization method pioneered by Nemirovski and Yudin [44] and studied further by Beck and Teboulle [5], Alvarez et al. [2], Nesterov [45] and many others. This observation also extends to the class of online mirror descent (OMD) algorithms introduced by Shalev-Shwartz [52] for *online* convex problems: focusing on the interplay between discrete- and continuous-time

[3]Interestingly, while exponential reinforcement learning eliminates dominated strategies at an exponential rate, reinforcement learning with nonsteep penalty functions eliminates such strategies in finite time.

OMD schemes, Kwon and Mertikopoulos [33] recently showed that a unilateral variant of the reinforcement learning dynamics studied in this paper leads to no regret against any (locally integrable) stream of payoffs.

Our analysis relies heavily on tools from convex analysis and, in particular, the theory of *Bregman functions* (Bregman [11]). Returning to the case of the replicator dynamics, it is well known that the *Kullback–Leibler (KL) divergence* (an oriented distance measure between probability distributions) is a potent tool for understanding the dynamics' long-run behavior (see Weibull [59] and Hofbauer and Sigmund [26]). For other steep cases, this part can be played by the *Bregman divergence*, a distance-like function whose role in population dynamics was noted recently by Harper [22].[4] In the nonsteep regime however, the dynamics of mixed strategies under reinforcement learning depend intrinsically on the players' score vectors, which determine a strategy's support. To contend with this, we introduce the *Fenchel coupling*, an oriented measure of proximity between primal and dual variables (that is, between mixed strategies and score vectors) that provides a natural tool for proving elimination, convergence, and stability results.

Paper Outline. We begin in Section 2 with some game-theoretic preliminaries and the definition of penalty functions and regularized best responses. The class of reinforcement learning dynamics we examine is presented in Section 3, along with several examples. Our analysis proper begins in Section 4 where we study the elimination of dominated strategies. In Section 5, we derive some stability and convergence properties of Nash equilibria, while in Section 6, we examine the long-term behavior of the players' time-averaged play. Actual, non-regularized best responses are discussed in Section 7.

2. Preliminaries

2.1. **Notation.** If V is a finite-dimensional real space, its dual will be denoted by V^* and we will write $\langle x|y\rangle$ for the pairing between $x \in V$ and $y \in V^*$. Also, if $\mathcal{S} = \{s_\alpha\}_{\alpha=0}^{d}$ is a finite set, the real space spanned by $\mathcal{S}$ will be denoted by $\mathbb{R}^{\mathcal{S}}$ and its canonical basis by $\{e_s\}_{s\in\mathcal{S}}$. For concision, we use α to refer interchangeably to either s_α or e_α, writing e.g. x_α instead of x_{s_α}; likewise, we write $\delta_{\alpha\beta}$ for the Kronecker delta symbols on $\mathcal{S}$. The set $\Delta(\mathcal{S})$ of probability measures on $\mathcal{S}$ will be identified with the d-dimensional simplex $\Delta \equiv \Delta(\mathcal{S}) = \{x \in \mathbb{R}^{\mathcal{S}} : \sum_\alpha x_\alpha = 1 \text{ and } x_\alpha \geq 0\}$ of $\mathbb{R}^{\mathcal{S}}$ and the relative interior of Δ will be denoted by $\Delta^\circ \equiv \operatorname{rel\,int}(\Delta)$. Finally, if $\{\mathcal{S}_k\}_{k\in\mathcal{N}}$ is a finite family of finite sets, we will use the shorthand $(\alpha_k; \alpha_{-k})$ for the tuple $(\dots, \alpha_{k-1}, \alpha_k, \alpha_{k+1}, \dots)$ and we will write $\sum_{\alpha_k}^{k}$ instead of $\sum_{\alpha_k\in\mathcal{S}_k}$.

2.2. **Normal form games.** A *finite game in normal form* is a tuple $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ 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*) per 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* $(\alpha_1, \dots, \alpha_N)$, $\alpha_k \in \mathcal{A}_k$. The set of *mixed strategies* of player k is denoted by $\mathcal{X}_k \equiv \Delta(\mathcal{A}_k)$ and the space $\mathcal{X} \equiv \prod_k \mathcal{X}_k$ of *mixed strategy profiles* $x = (x_1, \dots, x_N)$ will be called the game's *strategy space*. Unless mentioned otherwise, we will write $V_k \equiv \mathbb{R}^{\mathcal{A}_k}$ for the real space spanned by $\mathcal{A}_k$ and $V = \prod_k V_k$ for the ambient space of $\mathcal{X}$.

[4] Each penalty function defines a Bregman divergence; in the case of the Gibbs entropy, the resulting divergence is simply the KL divergence. For a comprehensive treatment, see Kiwiel [32].

The expected payoff of player k in the mixed strategy profile $x = (x_1, \dots, x_N) \in \mathcal{X}$ is

$$u_k(x) = \sum\nolimits_{\beta_1}^{1} \cdots \sum\nolimits_{\beta_N}^{N} u_k(\beta_1, \dots, \beta_N)\; x_{1,\beta_1} \cdots x_{N,\beta_N}, \tag{2.1}$$

where $u_k(\beta_1, \dots, \beta_N)$ denotes the payoff of player k in the profile $(\beta_1, \dots, \beta_N) \in \mathcal{A}$. Accordingly, the payoff corresponding to $\alpha \in \mathcal{A}_k$ in the mixed profile $x \in \mathcal{X}$ is

$$v_{k\alpha}(x) \equiv u_k(\alpha; x_{-k}) = \sum\nolimits_{\beta_1}^{1} \cdots \sum\nolimits_{\beta_N}^{N} u_k(\beta_1, \dots, \beta_N)\; x_{1,\beta_1} \cdots \delta_{\alpha\beta_k} \cdots x_{N,\beta_N}, \tag{2.2}$$

and we will write $v_k(x) = (v_{k\alpha}(x))_{\alpha \in \mathcal{A}_k} \in V_k^*$ for the *payoff (co)vector* of player k at $x \in \mathcal{X}$. The prefix "co" above is motivated by the natural duality pairing

$$\langle v_k(x) | x_k \rangle = \sum\nolimits_{\alpha}^{k} x_{k\alpha} v_{k\alpha} = u_k(x). \tag{2.3}$$

which shows that $v_k(x)$ acts on x_k as a linear functional. Duality plays a basic role in our analysis, so mixed strategies $x_k \in V_k$ will be treated as *primal* variables and payoff vectors $v_k \in V_k^*$ as *duals*.[5]

Finally, a *restriction* of $\mathfrak{G}$ is 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 $\mathcal{A}' \equiv \prod_k \mathcal{A}'_k$ of $\mathfrak{G}'$.

2.3. **Regularized best responses.** In view of (2.3), a player's *best response* to a payoff vector $v_k \in V_k^*$ is simply

$$\mathsf{br}_k(v_k) = \operatorname*{arg\,max}_{x_k \in \mathcal{X}_k} \langle v_k | x_k \rangle \,. \tag{2.4}$$

A standard way of obtaining a single-valued analogue of (2.4) is to introduce a *penalty* term that is (at least) strictly convex in x_k. It is also customary to assume that such penalties are "infinitely steep" at the boundary of the simplex (see e.g. Fudenberg and Levine [20]), but we obtain a much richer theory by dropping this requirement. Formally:

Definition 2.1. Let Δ be the unit d-dimensional simplex of $\mathbb{R}^{d+1}$. We say that $h\colon \Delta \to \mathbb{R}$ is a *penalty function* on Δ if:

(1) h is continuous on Δ.
(2) h is smooth on the relative interior of every face of Δ.[6]
(3) h is *strongly convex* on Δ: there exists some $K > 0$ such that

$$h(tx_1 + (1-t)x_2) \leq t h(x_1) + (1-t)h(x_2) - \tfrac{1}{2} K t(1-t) \left\| x_1 - x_2 \right\|^2, \tag{2.5}$$

for all $x_1, x_2 \in \Delta$ and for all $t \in [0,1]$.

Moreover, if $\lim_{x_n \to x} \|dh(x_n)\| = +\infty$, we will say that h is steep at x – or, more simply, that h is *steep* if this holds for all $x \in \mathrm{bd}(\Delta)$. Finally, h is called *decomposable with kernel* θ if

$$h(x) = \sum\nolimits_{\beta=0}^{d} \theta(x_\beta), \tag{2.6}$$

for some continuous and strongly convex $\theta\colon [0,1] \to \mathbb{R}$ that is smooth on $(0,1]$.

[5]Even though this distinction is rarely made in game theory, it is standard in learning and optimization – see e.g. Rockafellar [48], Nemirovski and Yudin [44], Shalev-Shwartz [52].

[6]More precisely, we posit here that $h(\gamma(t))$ is smooth for every smooth curve $\gamma\colon (-\varepsilon, \varepsilon) \to \Delta$ that is entirely contained in the relative interior of a face of $\mathcal{C}$.

Remark 2.1. For differentiation purposes, it will often be convenient to assume that the domain of h is the "thick simplex" $\Delta_\varepsilon = \{x \in \mathbb{R}^S : x_\alpha \geq 0 \text{ and } 1-\varepsilon < \sum_\alpha x_\alpha < 1+\varepsilon\}$; doing so allows us to carry out certain calculations in terms of standard coordinates, but none of our results depend on this device. Also, "smooth" should be interpreted above as "C^∞-smooth"; our analysis actually requires C^2 smoothness only if the Hessian of h is involved (and no differentiability otherwise), but we will keep the C^∞ assumption for simplicity.

Given a penalty function h on $\Delta \subseteq V \equiv \mathbb{R}^{d+1}$, the concave maximization problem

$$\begin{aligned} &\text{maximize} && \langle y|x\rangle - h(x), \\ &\text{subject to} && x \in \Delta, \end{aligned} \tag{2.7}$$

admits a unique solution for all $y \in V^*$, so the operator $\arg\max_x \{\langle y|x\rangle - h(x)\}$ becomes single-valued. We thus obtain:

Definition 2.2. The *regularized best response map* (or *choice map*) $Q\colon V^* \to \Delta$ induced by a penalty function h on Δ is

$$Q(y) = \arg\max_{x\in\Delta} \{\langle y|x\rangle - h(x)\}, \quad y \in V^*. \tag{2.8}$$

Under the steepness and (strong) convexity requirements of Definition 2.1, the discussion in Rockafellar [48, Chapter 26] shows that the induced choice map Q is smooth and its image is the relative interior Δ° of Δ; otherwise, if h is nowhere steep, the image of Q is the entire simplex (cf. Remark B.1 in Appendix B). We illustrate this dichotomy with two representative examples (see also Section 3.4):

Example 2.1. The classic example of a steep penalty function is the (negative) *Gibbs entropy*

$$h(x) = \sum\nolimits_{\alpha=0}^{d} x_\alpha \log x_\alpha. \tag{2.9}$$

As is well known, the induced choice map (2.8) is the so-called *logit map*

$$G_\alpha(y) = \frac{\exp(y_\alpha)}{\sum_{\beta=0}^{d} \exp(y_\beta)}. \tag{2.10}$$

Since h is steep, $\operatorname{im} G = \Delta^\circ$.

Example 2.2. The basic example of a non-steep penalty function is the *quadratic penalty*

$$h(x) = \frac{1}{2}\sum\nolimits_{\beta=0}^{d} x_\beta^2. \tag{2.11}$$

The induced choice map (2.8) is the (Euclidean) *projection map*

$$\Pi(y) = \arg\max_{x\in\Delta} \left\{\langle y|x\rangle - \tfrac{1}{2}\|x\|_2^2\right\} = \arg\min_{x\in\Delta} \|y-x\|_2^2 = \operatorname{proj}_\Delta y, \tag{2.12}$$

where $\operatorname{proj}_\Delta$ denotes the closest point projection to Δ with respect to the standard Euclidean norm $\|\cdot\|_2$ on $\mathbb{R}^{d+1}$. Obviously, $\operatorname{im}\Pi = \Delta$.

Remark 2.2. Up to mild technical differences, penalty functions are also known as *regularizers* in online learning and as *Bregman functions* (or *prox-functions*) in convex analysis; for comprehensive treatments, see Bregman [11], Nemirovski and Yudin [44], Shalev-Shwartz [52] and references therein. The term "decomposable" is borrowed from Alvarez et al. [2].

Choice maps generated by steep penalty functions are called *smooth best response functions* by Fudenberg and Levine [20]; in a similar vein, drawing on a connection

between deterministic and stochastic payoff perturbations, Hofbauer and Sandholm [24] use the terminology "perturbed best responses" while McKelvey and Palfrey [39] call such maps "quantal response functions".

Remark 2.3. Several models of smooth fictitious play (Fudenberg and Levine [20], Hofbauer and Sandholm [24]) use a steep penalty function with positive-definite Hessian. Concerning this last condition, the strong convexity of h imposes a positive lower bound on the smallest eigenvalue of $\mathrm{Hess}(h)$, a property which in turn ensures that the associated choice map Q is Lipschitz continuous (Proposition B.1); later, we also take advantage of the fact that strong convexity provides a lower bound for the so-called *Bregman divergence* between points in Δ (Proposition B.2). Both properties simplify our presentation considerably, so we will not venture beyond strongly convex penalty functions.

3. A Class of Reinforcement Learning Dynamics

3.1. **Definition and basic examples.** The basic reinforcement learning scheme that we consider is that at each moment $t \geq 0$, every player $k \in \mathcal{N}$ of a finite game $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ plays a regularized best response $Q_k(y_k)$ to a score vector $y_k \in V_k^*$ consisting of his actions' aggregate payoffs. More precisely, this amounts to the continuous-time process:

$$\begin{aligned} y_k(t) &= y_k(0) + \int_0^t v_k(x(s))\, ds, \\ x_k(t) &= Q_k(y_k(t)), \end{aligned} \tag{RL}$$

or, in differential form:

$$\dot{y}_k = v_k(Q(y)), \tag{3.1}$$

where $Q \equiv (Q_1, \dots, Q_N) \colon V^* \equiv \prod_k V_k^* \to \mathcal{X}$ and $y = (y_1, \dots, y_N) \in V^*$ denote the players' choice and score profiles respectively. In the above, the (primal) strategy variable $x_k(t) \in \mathcal{X}_k$ describes the mixed strategy of player k at time t, the (dual) score vector $y_k(t)$ aggregates the payoffs of the pure strategies $\alpha \in \mathcal{A}_k$ of player k, Accordingly, the basic interpretation of (RL) is that each player observes the realized expected payoffs of his strategies over a short interval of time and then uses these payoffs to update his score vector.

More precisely, in the stochastic approximation language of Benaïm [7], (RL) is simply the mean field of the discrete-time stochastic process

$$\begin{aligned} Y_{k\alpha}(n+1) &= Y_{k\alpha}(n) + \hat{v}_{k\alpha}(n), \\ X_{k\alpha}(n+1) &= Q_{k\alpha}(Y_k(n+1)), \end{aligned} \tag{3.2}$$

where $X_{k\alpha}(n)$ is the probability of playing $\alpha \in \mathcal{A}_k$ at time $n = 1, 2, \dots$, and $\hat{v}_{k\alpha}(n)$ is an unbiased estimator of $v_{k\alpha}(X(n))$. If player k can observe the action profile $\alpha_{-k}(n)$ played by his opponents (or can otherwise calculate his strategies' payoffs), such an estimate is provided by $\hat{v}_{k\alpha}(n) = u_k(\alpha; \alpha_{-k}(n))$. If instead player k can only observe the payoff $\hat{u}_k(n) = u_k(\alpha_k(n); \alpha_{-k}(n))$ of his chosen action $\alpha_k(n)$, a standard choice for $\hat{v}_k(n)$ is

$$\hat{v}_{k\alpha}(n) = \hat{u}_k(n)/X_{k\alpha}(n) \quad \text{if } \alpha_k(n) = \alpha, \tag{3.3}$$

where the division by $X_{k\alpha}(n)$ compensates for the infrequency with which the score of strategy α is updated. Estimator (3.3) is sound if the penalty function of player k is steep (see Leslie and Collins [37] and Coucheney et al. [14]); otherwise, $X_{k\alpha}(n)$

may become zero, in which case the links between (3.3) and our continuous-time model are less clear.

We begin with two representative examples of the reinforcement learning dynamics (RL):

Example 3.1. If Q is the logit map (2.10) of Example 2.1, players select actions with probability proportional to the exponential of their aggregate payoffs. In this case, (RL) boils down to the *exponential* (or *logit*) *reinforcement learning process*:

$$\begin{aligned} \dot{y}_{k\alpha} &= v_{k\alpha}(x), \\ x_{k\alpha} &= \frac{\exp(y_{k\alpha})}{\sum_\beta^k \exp(y_{k\beta})}. \end{aligned} \tag{XL}$$

In a single-agent, online learning context, the discrete-time version of (XL) first appeared in the work of Vovk [58] and Littlestone and Warmuth [38] – see also Rustichini [49], Sorin [53], and Kwon and Mertikopoulos [33] for a continuous-time analysis. In a game-theoretic setting, this process has been studied by (among others) Freund and Schapire [18], Hofbauer et al. [27] and Mertikopoulos and Moustakas [40], while Leslie and Collins [37], Tuyls et al. [57] and, more recently, Coucheney et al. [14] considered a discounted variant that we describe in Section 3.2 below.

Differentiating $x_{k\alpha}$ in (XL) with respect to time and substituting yields

$$\dot{x}_{k\alpha} = \frac{\dot{y}_{k\alpha} e^{y_{k\alpha}} \sum_\beta^k e^{y_{k\beta}} - e^{y_{k\alpha}} \sum_\beta^k \dot{y}_{k\beta} e^{y_{k\beta}}}{\left(\sum_\beta^k e^{y_{k\beta}}\right)^2} = x_{k\alpha} \left[\dot{y}_{k\alpha} - \sum\nolimits_\beta^k x_{k\beta} \dot{y}_{k\beta} \right], \tag{3.4}$$

so, with $\dot{y}_{k\alpha} = v_{k\alpha}(x)$, we readily obtain:

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

This equation describes the replicator dynamics of Taylor and Jonker [55], so the evolution of mixed strategies under (XL) agrees with a fundamental model of evolutionary game theory whose long-term rationality properties are quite well understood. This basic relation between exponential reinforcement learning and the replicator dynamics was noted in a single-agent environment by Rustichini [49] and was explored further in a game-theoretic context by Hofbauer et al. [27] and Mertikopoulos and Moustakas [40].

Example 3.2. If Q is the projection map (2.12) of Example 2.2, (RL) leads to the *projected reinforcement learning* process:

$$\begin{aligned} \dot{y}_k &= v_k(x), \\ x &= \operatorname{proj}_{\mathcal{X}} y. \end{aligned} \tag{PL}$$

Of course, since $\operatorname{proj}_{\mathcal{X}} y$ is not smooth in y, we can no longer use the same approach as in (3.4) to derive the dynamics of the players' mixed strategies x_k. Instead, recall (or solve the defining convex program to show) that the closest point projection on $\mathcal{X}_k = \Delta(\mathcal{A}_k)$ takes the simple form

$$\left(\operatorname{proj}_{\mathcal{X}_k} y_k\right)_\alpha = \max\{y_{k\alpha} + \mu_k, 0\}, \tag{3.5}$$

where $\mu_k \in \mathbb{R}$ is such that $\sum_\alpha^k \max\{y_{k\alpha} + \mu_k, 0\} = 1$. We can thus write

$$x_{k\alpha}(t) = \begin{cases} y_{k\alpha}(t) + \mu_k(t) & \text{if } \alpha \in \mathcal{A}'_k, \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

Therefore, if I_k is an open time interval over which $x_k(t) = \operatorname{proj}_{\mathcal{X}_k} y_k(t)$ has constant support $\mathcal{A}'_k \subseteq \mathcal{A}_k$, differentiating (3.6) for $t \in I_k$ yields

$$\dot{x}_{k\alpha} = \dot{y}_{k\alpha} + \dot{\mu}_k = v_{k\alpha} + \dot{\mu}_k \quad \text{for all } \alpha \in \mathcal{A}'_k. \tag{3.7}$$

Since $\sum_{\alpha \in \mathcal{A}'_k} \dot{x}_{k\alpha} = 0$, summing over $\alpha \in \mathcal{A}'_k$ gives

$$0 = \sum\nolimits_{\alpha \in \mathcal{A}'_k} v_{k\alpha} + \dot{\mu}_k \left| \mathcal{A}'_k \right|, \tag{3.8}$$

so, by substituting into (3.7) and rearranging, we obtain the *projection dynamics*:

$$\dot{x}_{k\alpha} = \begin{cases} v_{k\alpha}(x) - \left|\operatorname{supp}(x_k)\right|^{-1} \sum\nolimits_{\beta \in \operatorname{supp}(x_k)} v_{k\beta}(x) & \text{if } \alpha \in \operatorname{supp}(x_k), \\ 0 & \text{if } \alpha \notin \operatorname{supp}(x_k). \end{cases} \tag{PD}$$

The dynamics (PD) were introduced in game theory by Friedman [19] as a geometric model of the evolution of play in population games.[7] The previous discussion shows that the projected orbits $x(t) = \operatorname{proj}_{\mathcal{X}} y(t)$ of the learning scheme (PL) satisfy the projection dynamics (PD) on every open interval over which the support of $x(t)$ is fixed; furthermore, as we argue below, the union of these intervals is dense in $[0, \infty)$. In this way, orbits of (PL) that begin in the relative interior $\mathcal{X}^\circ$ of the game's strategy space may attain a boundary face in finite time, then move to another boundary face or re-enter $\mathcal{X}^\circ$ (again in finite time), and so on. Thus, although $x(t)$ may fail to be differentiable when it moves from (the relative interior of) one face of $\mathcal{X}$ to another, it satisfies (PD) for all times in between.

These two examples illustrate a fundamental dichotomy between reinforcement learning processes induced by steep and nonsteep penalty functions. In the steep case, the dynamics of the strategy variable are well-posed and admit unique solutions that stay in $\mathcal{X}^\circ$ for all time. On the other hand, in the nonsteep regime, the dynamics of the strategy variable only admit solutions in an extended sense, and may enter and exit different faces of $\mathcal{X}$ in perpetuity.

Remark 3.1. In several treatments of stochastic fictitious play (Fudenberg and Levine [20], Hofbauer and Sandholm [24]), it is common to replace $h(x)$ with $\eta h(x)$ for some positive parameter $\eta > 0$ which is often called the *noise level*.[8] For instance, if h is the Gibbs entropy (2.9), this adjustment leads to the choice map

$$G^\gamma_\alpha(y) = \frac{\exp(\gamma y_\alpha)}{\sum_{\beta=0}^d \exp(\gamma y_\beta)}. \tag{3.9}$$

where $\gamma = \eta^{-1}$. If y is fixed, choices are nearly uniform for small γ; on the other hand, for large γ, almost all probability is placed on the pure strategies with the highest score.

[7] Nagurney and Zhang [43] (see also Lahkar and Sandholm [34] and Sandholm et al. [51]) introduce related projection-based dynamics for population games. The relations among the various projection dynamics are explored in a companion paper (Mertikopoulos and Sandholm [41]).

[8] The term "noise level" reflects the fact that η essentially controls the magnitude of the perturbation to the player's expected payoff in the regularized maximization problem (2.7).

In the present context, replacing $h_k(x_k)$ with $\eta_k h_k(x_k)$ and writing $\gamma_k = \eta_k^{-1}$ yields the following variant of (RL):

$$\begin{aligned} \dot{y}_k &= v_k(x), \\ x_k &= Q_k(\gamma_k y_k), \end{aligned} \tag{3.10}$$

Since a rescaled penalty function is itself a penalty function, (3.10) can be viewed as an instance of (RL); therefore, our results for the latter also apply to the former. Furthermore, because the score variables $y_k(t)$ scale with t, introducing γ has less drastic consequences under (RL) than under stochastic fictitious play: for instance, the stationary points of (RL) in $\mathcal{X}$ remain unaffected by this choice – see Theorem 5.2 below.

One can also consider a variant of (RL) under which different players adjust their score variables at different rates:

$$\begin{aligned} \dot{y}_k &= \gamma_k v_k(x), \\ x_k &= Q_k(y_k), \end{aligned} \tag{RL$_\gamma$}$$

Evidently, this process is equivalent to (3.10), but with initial conditions $y_k(0)$ scaled by $1/\gamma_k$. As we show in Propositions 4.2 and 5.4, the choice of γ affects the speed at which (3.10) evolves because it determines each player's characteristic time scale.

3.2. **Related models.** Before proceeding with our analysis of (RL), we mention a number of related models appearing in the literature.

First, as an alternative to aggregating payoffs in (RL), one can consider the exponentially discounted model

$$y_k(t) = y_k(0)\lambda^t + \int_0^t \lambda^{t-s} v_k(x(s))\, ds, \tag{3.11}$$

where the discount rate $\lambda \in (0,1)$ measures the relative weight of past observations. This variant was examined by Leslie and Collins [37], Tuyls et al. [57], and Coucheney et al. [14] for smooth best response maps generated by steep penalty functions. Obviously, when h is steep, (RL) can be seen as a special case of (3.11) in the limit $\lambda \to 1^-$; in contrast to (RL) however, discounting implies that the score variable $y_k(t)$ remains bounded, thus preventing the agents' mixed strategies from approaching the boundary of $\mathcal{X}$. For instance, under the logit rule of Example 3.1, Coucheney et al. [14] showed that mixed strategies follow a differential equation that combines the replicator dynamic with an additional term which repels orbits from the boundary $\mathrm{bd}(\mathcal{X})$ of $\mathcal{X}$.

In a single-agent environment, payoffs are determined at each instance by nature so the reinforcement learning process (RL) becomes:

$$y_\alpha(t) = y_\alpha(0) + \int_0^t v_\alpha(s)\, ds, \qquad x(t) = Q(y(t)). \tag{3.12}$$

In this context, (RL) can be seen as a continuous-time analogue of the family of online learning algorithms known as online mirror descent (OMD) – for a comprehensive account, see Bubeck [12] and Shalev-Shwartz [52]. The resulting interplay between discrete and continuous time has been analyzed by Sorin [53] and Kwon and Mertikopoulos [33] who showed that (3.12) leads to no regret against any locally integrable payoff stream $v(t)$ in V^*.

In view of the above, (RL) essentially extends the discounted dynamics of Leslie and Collins [37] to the nonsteep regime and the online learning dynamics of Kwon and Mertikopoulos [33] to a multi-agent, game-theoretic setting. That being said, there are several reinforcement learning schemes that are distinct from (RL) but which still lead to the replicator equation (RD). A leading model of this kind is presented in the seminal paper of Erev and Roth [16]: after player k chooses pure strategy $\alpha \in \mathcal{A}_k$, he increments its score by the payoff he receives (assumed positive) and then updates his choice probabilities proportionally to each action's score. The continuous-time, deterministic version of this model is

$$\begin{aligned} \dot{y}_{k\alpha} &= x_{k\alpha} v_{k\alpha}(x), \\ x_{k\alpha} &= \frac{y_{k\alpha}}{\sum_\beta^k y_{k\beta}}, \end{aligned} \tag{ER}$$

where the $x_{k\alpha}$ term in the first equation reflects the fact that the score variable $y_{k\alpha}$ is only updated when α is played. A simple calculation then shows that the evolution of mixed strategies is governed by the replicator dynamics (RD) up to a player-specific multiplicative factor of $\sum_\beta^k y_{k\beta}$. Versions of this model have been studied by Posch [46], Rustichini [49], Hopkins [30], Beggs [6], and Hopkins and Posch [31]; Rustichini [49] also considers hybrids between (XL) and (ER) in nonstrategic settings.

Börgers and Sarin [10] also consider a variant of the learning model of Cross [15] where there is no separate score variable. Instead, if player k chooses pure strategy α, he increases the probability with which he plays α and decreases the probability of every other action $\beta \neq \alpha$ proportionally to the payoff $v_{k\alpha}(x) \in (0,1)$ that the player obtained. The continuous-time, deterministic version of this model is

$$\dot{x}_{k\alpha} = x_{k\alpha}(1 - x_{k\alpha}) v_{k\alpha}(x) + \sum\nolimits_{\beta \neq \alpha} x_{k\beta}(0 - x_{k\alpha}) v_{k\beta}(x), \tag{3.13}$$

so, after a trivial rearrangement, we obtain again the replicator equation (RD).

3.3. **Basic results.** We start our analysis by showing that the dynamics (RL) are well-posed even if the players' penalty functions are not steep:

Proposition 3.1. *The reinforcement learning process* (RL) *admits a unique global solution for every initial score profile* $y(0) \in V^*$.

Proof. By Proposition B.1, strong convexity implies that $Q\colon V^* \to \mathcal{X}$ is Lipschitz. Since v_k is bounded, existence and uniqueness of global solutions follows from standard arguments – e.g. Robinson [47, Chapter V]. □

We turn now to the dynamics induced by (RL) on the game's strategy space $\mathcal{X}$. To that end, if $y(t)$ is a solution orbit of (RL), we call $x(t) = Q(y(t))$ the *trajectory of play induced by* $y(t)$ – or, more simply, an *orbit of* (RL) *in* $\mathcal{X}$. Mirroring the derivation of the projection dynamics (PD) above, our next result provides a dynamical system on $\mathcal{X}$ that is satisfied by smooth segments of orbits of (RL) in $\mathcal{X}$.

To state the result, let

$$g_k^{\alpha\beta}(x_k) = \operatorname{Hess}(h_k'(x_k))^{-1}_{\alpha\beta}, \qquad \alpha, \beta \in \operatorname{supp}(x_k), \tag{3.14}$$

denote the inverse Hessian matrix of the restriction $h'_k \equiv h_k|_{\mathcal{X}'_k}$ of the penalty function of player k to the face $\mathcal{X}'_k = \Delta(\operatorname{supp}(x_k))$ of $\mathcal{X}_k$ that is spanned by $\operatorname{supp}(x_k)$.[9] Furthermore, let

$$g_k^\alpha(x_k) = \sum\nolimits_{\beta\in\operatorname{supp}(x_k)} g_k^{\alpha\beta}(x_k) \quad \text{and} \quad G_k(x_k) = \sum\nolimits_{\alpha\in\operatorname{supp}(x_k)} g_k^\alpha(x_k) \tag{3.15}$$

denote the row sums and the grand sum of $g_k^{\alpha\beta}(x_k)$ respectively. We then have:

Proposition 3.2. *Let $x(t) = Q(y(t))$ be an orbit of (RL) in $\mathcal{X}$, and let I be an open interval over which the support of $x(t)$ remains constant. Then, for all $t \in I$, $x(t)$ satisfies:*

$$\dot{x}_{k\alpha} = \sum_{\beta\in\operatorname{supp}(x_k)} \left[g_k^{\alpha\beta}(x_k) - \frac{1}{G_k(x_k)} g_k^\alpha(x_k) g_k^\beta(x_k)\right] v_{k\beta}(x), \quad \alpha \in \operatorname{supp}(x_k). \tag{RLD}$$

Corollary 3.3. *Every orbit $x(t) = Q(y(t))$ of (RL) in $\mathcal{X}$ is Lipschitz continuous and satisfies (RLD) on an open dense subset of $[0,\infty)$. Furthermore, if the players' penalty functions are steep, (RLD) is well-posed and $x(t)$ is an ordinary solution thereof.*

Proof of Proposition 3.2. Let $\mathcal{A}'_k$ denote the (constant) support of $x_k(t)$ for $t \in I$. Then, the first-order Karush–Kuhn–Tucker (KKT) conditions for the softmax problem (2.7) of player k readily yield

$$y_{k\alpha} - \frac{\partial h'_k}{\partial x_{k\alpha}} = \mu_k \quad \text{for all } \alpha \in \mathcal{A}'_k, \tag{3.16}$$

where $h'_k = h_k|_{\mathcal{X}'_k}$ is the restriction of h_k to $\mathcal{X}'_k = \Delta(\mathcal{A}'_k)$ and μ_k is the Lagrange multiplier associated to the constraint $\sum_{\alpha\in\mathcal{A}'_k} x_{k\alpha} = 1$. Differentiating (3.16) then yields

$$\dot{y}_{k\alpha} - \sum_{\beta\in\mathcal{A}'_k} \frac{\partial^2 h'_k}{\partial x_{k\alpha}\partial x_{k\beta}} \dot{x}_{k\beta} = \dot{\mu}_k, \tag{3.17}$$

and hence

$$\dot{x}_{k\alpha} = \sum_{\beta\in\mathcal{A}'_k} g_k^{\alpha\beta}(x_k)\,(v_{k\beta}(x) - \dot{\mu}_k) = \sum_{\beta\in\mathcal{A}'_k} g_k^{\alpha\beta}(x_k) v_{k\beta}(x) - g_k^\alpha(x_k)\dot{\mu}_k, \tag{3.18}$$

where we have used the fact that $\dot{y}_k = v_k$. However, since $x_k(t) \in \mathcal{X}'_k$ for all $t \in I$ by assumption, we must also have $\sum_{\alpha\in\mathcal{A}'_k} \dot{x}_{k\alpha} = 0$; accordingly, (3.18) gives

$$0 = \sum_{\alpha,\beta\in\mathcal{A}'_k} g_k^{\alpha\beta}(x_k) v_{k\beta}(x) - \dot{\mu}_k \sum_{\alpha\in\mathcal{A}'_k} g_k^\alpha(x_k) = \sum_{\beta\in\mathcal{A}'_k} g_k^\beta(x_k) v_{k\beta}(x) - G_k(x_k)\dot{\mu}_k, \tag{3.19}$$

so (RLD) is obtained by solving (3.19) for $\dot{\mu}_k$ and substituting in (3.18). □

Proof of Corollary 3.3. Lipschitz continuity follows from Proposition 3.1 and the Lipschitz continuity of Q (Proposition B.1). To establish the next claim, we must show that the union of all open intervals over which $x(t)$ has constant support is dense in $[0,\infty)$. To do so, fix some $\alpha \in \mathcal{A}_k$, $k \in \mathcal{N}$, and let $A = \{t : x_{k\alpha}(t) > 0\}$ so that $A^c = x_{k\alpha}^{-1}(0)$; then, if $B = \operatorname{int}(A^c)$, it suffices to show that $A \cup B$ is dense

[9] Strong convexity ensures that $\operatorname{Hess}(h'_k)$ is positive-definite – and, hence, invertible (cf. Remark 2.3).

in $[0,\infty)$. Indeed, if $t \notin \mathrm{cl}(A \cup B)$, we must have $t \notin A$ and hence $x_{k\alpha}(t) = 0$. Furthermore, since $t \notin \mathrm{cl}(A)$, there exists a neighborhood U of t that is disjoint from A, i.e. $x_{k\alpha} = 0$ on U. Since U is open, we get $U \subseteq \mathrm{int}(A^c) = B$, contradicting that $t \notin B$.

Finally, to prove the second part of our claim, simply note that the image $\mathrm{im}\, Q_k$ of Q_k coincides with the relative interior $\mathcal{X}_k^\circ$ of $\mathcal{X}_k$ if and only if h_k is steep (cf. Proposition B.1). □

Remark 3.2. If the players' penalty functions can be decomposed as $h_k(x_k) = \sum_\beta^k \theta_k(x_{k\beta})$ (cf. Definition 2.1), the inverse Hessian matrix of h_k may be written as

$$g_k^{\alpha\beta}(x_k) = \frac{\delta_{\alpha\beta}}{\theta_k''(x_{k\alpha})}, \qquad \alpha,\beta \in \mathrm{supp}(x_k). \tag{3.20}$$

In this case, (RLD) may be written more explicitly as

$$\dot{x}_{k\alpha} = \frac{1}{\theta_k''(x_{k\alpha})}\left[v_{k\alpha}(x) - \Theta_k''(x_k)\sum\nolimits_{\beta\in\mathrm{supp}(x_k)} v_{k\beta}(x)/\theta_k''(x_{k\beta})\right], \tag{RLD$_\theta$}$$

where Θ_k'' stands for the harmonic aggregate[10]

$$\Theta_k''(x_k) = \left[\sum\nolimits_{\beta\in\mathrm{supp}(x_k)} 1/\theta_k''(x_{k\beta})\right]^{-1}. \tag{3.21}$$

To the best of our knowledge, the separable dynamics (RLD$_\theta$) first appeared in a comparable form in the work of Harper [22] under the name "escort replicator dynamics".[11] From the perspective of convex programming, the dynamics (RLD$_\theta$) for steep θ (and, more generally, (RLD) for steep h) can be seen as a game-theoretic analogue of the Hessian Riemannian gradient flow framework developed by Alvarez et al. [2]. As such, (RLD) exhibits a deep Riemannian-geometric character which links it to class of dynamics introduced by Hofbauer and Sigmund [25] and studied further by Hopkins [28]. These geometric aspects of (RLD) are explored in detail in a companion paper (Mertikopoulos and Sandholm [41]).

3.4. **Further examples.** In Section 3.1, we introduced the exponential and projected reinforcement learning models (XL) and (PL), generated respectively by the (steep) entropic penalty (2.9) and the (nonsteep) quadratic penalty (2.11). We close this section with some further examples of penalty functions and their induced mixed strategy dynamics.

Example 3.3 (The Tsallis entropy and the q-replicator dynamics). A well known generalization of the Gibbs (negative) entropy due to Tsallis [56] is:[12]

$$h(x) = [q(1-q)]^{-1}\sum\nolimits_\beta (x_\beta - x_\beta^q), \qquad q > 0, \tag{3.22}$$

with the continuity convention $(z - z^q)/(1-q) = z\log z$ for $q = 1$ (corresponding to the Gibbs penalty of Example 2.1). In this context, some easy algebra shows

[10] We should stress here that Θ_k'' is not a second derivative; we only use this notation for visual consistency.

[11] See also Coucheney et al. [14] for a variant of (RLD$_\theta$) induced by the exponentially discounted model (3.11).

[12] In information theory, the Tsallis entropy is often referred to as the Havrda–Charvát entropy. Note that definition (3.22) uses initial constant $[q(1-q)]^{-1}$ rather than the more common $(1-q)^{-1}$.

that (RLD) reduces to the *q-replicator dynamics*:

$$\dot{x}_{k\alpha} = x_{k\alpha}^{2-q} \left[v_{k\alpha} - r_{k,q}^{q-2} \sum\nolimits_{\beta}^{k} x_{k\beta}^{2-q} v_{k\beta} \right], \tag{RD$_q$}$$

where $r_{k,q}^{2-q} = \sum_{\beta}^{k} x_{k\beta}^{2-q}$ (and $0^0 = 0$ for $q = 2$). In the context of convex programming, (RD$_q$) was derived as an example of a Hessian Riemannian gradient flow in Alvarez et al. [2]; more recently, these dynamics also appeared in Harper [22] under the name "q-deformed replicator dynamics". Obviously, for $q = 1$, (RD$_q$) is simply the replicator equation (RD), reflecting the fact that the Tsallis penalty (3.22) converges to the Gibbs penalty (2.9) as $q \to 1$. Furthermore, for $q = 2$, the Tsallis penalty (3.22) is equal to the quadratic penalty (2.11) up to an affine term; consequently, since (RLD$_\theta$) does not involve the first derivatives of h, the dynamics (RD$_q$) for $q = 2$ are the same as the projection dynamics (PD).

Of course, (3.22) is steep if and only if $0 < q \leq 1$, so (RD$_q$) may fail to be well-posed for $q > 1$; in particular, as in the case of the projection dynamics (RD), the orbits of (RD$_q$) for $q > 1$ may run into the boundary of the game's strategy space in finite time. In this way, (RD$_q$) provides a smooth interpolation between the replicator dynamics and the projection dynamics (obtained for $q = 1$ and $q = 2$ respectively), with the replicator equation defining the boundary between the well- and ill-posed regimes of (RD$_q$).

Example 3.4 (The Rényi entropy). The Rényi (negative) entropy is defined by

$$h(x) = -(1-q)^{-1} \log \sum\nolimits_{\beta} x_{\beta}^{q} \tag{3.23}$$

for $q \in (0,1)$. Just like its Tsallis counterpart, (3.23) is steep for all $q \in (0,1)$ and it approaches the Gibbs penalty (2.9) as $q \to 1^-$; unlike (3.22) though, (3.23) is not decomposable.

Dropping momentarily the player index k, an easy calculation shows that

$$\partial_\alpha \partial_\beta h = q^2 (1-q)^{-1} \xi_\alpha \xi_\beta + q \delta_{\alpha\beta} (\xi_\alpha^2 + \xi_\alpha / x_\alpha), \tag{3.24}$$

where $\xi_\alpha = x_\alpha^{-1} (x_\alpha / r_q)^q$ and $r_q^q = \sum_\beta x_\beta^q$. It can then be shown that the inverse matrix of $\operatorname{Hess}(h)$ is $g^{\alpha\beta} = -x_\alpha x_\beta + q^{-1} \delta_{\alpha\beta} \xi_\alpha x_\beta$, so some more algebra leads to the *Rényi dynamics*:

$$\dot{x}_{k\alpha} = q^{-1} \xi_{k\alpha} x_{k\alpha} \left[v_{k\alpha} - v_{k,q} \right] + (1-q)^{-1} x_{k\alpha} (1 - \xi_{k\alpha}) \left[v_{k,q} - \bar{v}_k \right], \tag{ReD}$$

where $v_{k,q} = \sum_\beta^k x_{k\beta} \xi_{k\beta} v_{k\beta}$, $\bar{v}_k = \sum_\beta^k x_{k\beta} v_{k\beta}$ and the player index has been reinstated. Thus, with $\xi_{k\alpha} = x_{k\alpha}^{-1} (x_{k\alpha} / r_{k,q})^q \to 1$ as $q \to 1$, the replicator dynamics (RD) are immediately recovered in the limit $q \to 1$.

Example 3.5 (The log-barrier). An important nonexample of a penalty function is the logarithmic barrier:

$$h(x) = -\sum\nolimits_{\beta} \log x_\beta. \tag{3.25}$$

Obviously, (3.25) is steep, strongly convex and decomposable, but it fails to be finite at the boundary $\operatorname{bd}(\Delta)$ of Δ. Nevertheless, it can be obtained from the Tsallis penalty (3.22) in the limit $q \to 0^+$ and the recipe (RLD$_\theta$) yields the dynamics

$$\dot{x}_{k\alpha} = x_{k\alpha}^2 \left[v_{k\alpha} - r_k^{-2} \sum\nolimits_{\beta}^{k} x_{k\beta}^2 v_{k\beta} \right], \tag{LD}$$

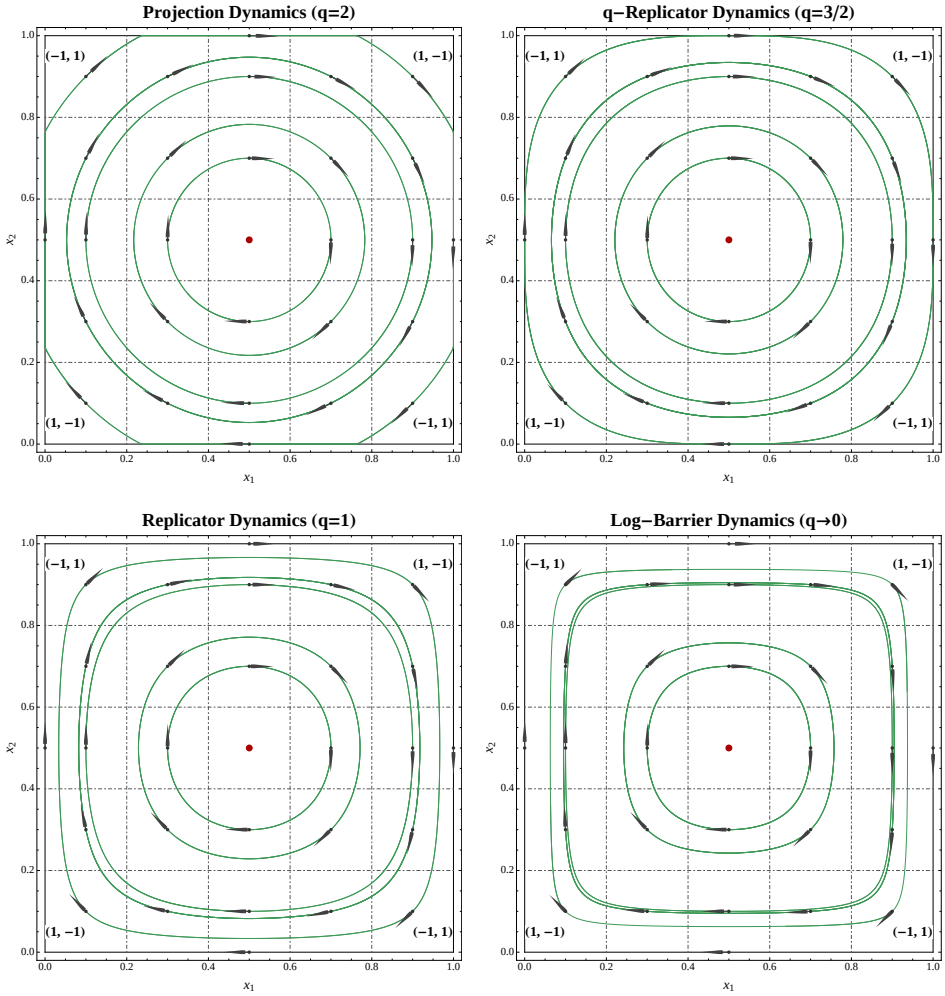


FIGURE 1. Evolution of play under the reinforcement learning dynamics (RL) in Matching Pennies (Nash equilibria are depicted in dark red and stationary points in light/dark blue; for the game's payoffs, see the vertex labels). As the deformation parameter q of (RD_q) decreases, we pass from the nonsteep regime ($q > 1$) where the orbits of (RL) in $\mathcal{X}$ collide with the boundary of $\mathcal{X}$ in finite time, to the steep regime ($q \leq 1$) where (RD_q) becomes well-posed.

with $r_k^2 = \sum_\beta^k x_{k\beta}^2$. The system (LD) is easily seen to be well-posed and it corresponds to the $q \to 0^+$ limit of (RD_q). In convex optimization, (LD) was first considered by Bayer and Lagarias [4] and it has since been studied extensively in the contexts of linear and convex programming – see e.g. Fiacco [17], Kiwiel [32], Alvarez et al. [2] and Laraki and Mertikopoulos [36]. The results that we derive in the rest of the paper for (RL) remain true in the case of (LD), but we do not provide proofs.

4. Elimination of Dominated Strategies

We begin our rationality analysis with the elimination of dominated strategies. Formally, if $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ is a finite game in normal form, we say that $q_k \in \mathcal{X}_k$ is *dominated* by $q'_k \in \mathcal{X}_k$ and we write $q_k \prec q'_k$ when

$$u_k(q_k; x_{-k}) < u_k(q'_k; x_{-k}) \quad \text{for all } x_{-k} \in \mathcal{X}_{-k} \equiv \prod\nolimits_{\ell \neq k} \mathcal{X}_\ell. \tag{4.1}$$

Thus, for pure strategies $\alpha, \beta \in \mathcal{A}_k$, we have $\alpha \prec \beta$ whenever

$$v_{k\alpha}(x) < v_{k\beta}(x) \quad \text{for all } x \in \mathcal{X}. \tag{4.2}$$

If (4.1) is strict for only some (but not all) $x \in \mathcal{X}$, we will say that q_k is *weakly dominated* by q'_k and we will write $q_k \preccurlyeq q'_k$; conversely, we will say that $q = (q_1, \dots, q_N) \in \mathcal{X}$ is *undominated* if no component $q_k \in \mathcal{X}_k$ of q is weakly dominated. Of course, if dominated strategies strategies are removed from $\mathfrak{G}$, other strategies may become dominated in the resulting restriction of $\mathfrak{G}$, leading to the notion of *iteratively dominated strategies*. Accordingly, a strategy which survives all rounds of elimination is called *iteratively undominated*.

For a given trajectory of play $x(t) \in \mathcal{X}$, $t \geq 0$, we say that the pure strategy $\alpha \in \mathcal{A}_k$ *becomes extinct along* $x(t)$ if $x_{k\alpha}(t) \to 0$ as $t \to \infty$. More generally, following Samuelson and Zhang [50], we say that the mixed strategy $q_k \in \mathcal{X}_k$ becomes extinct along $x(t)$ if $\min\{x_{k\alpha}(t) : \alpha \in \operatorname{supp}(q_k)\} \to 0$; otherwise, we say that q_k *survives*.

Extending the classic elimination results of Akin [1], Nachbar [42], and Samuelson and Zhang [50], we first show that only iteratively undominated strategies survive under (RL):

Theorem 4.1. *Let $x(t) = Q(y(t))$ be an orbit of (RL) in $\mathcal{X}$. If $q_k \in \mathcal{X}_k$ is dominated (even iteratively), then it becomes extinct along $x(t)$.*

In the case of the replicator dynamics, most proofs of elimination of dominated strategies involve some form of the Kullback–Leibler divergence $D_{\mathrm{KL}}(q_k, x_k) = \sum_\alpha^k q_{k\alpha} \log(q_{k\alpha}/x_{k\alpha})$, an asymmetric measure of the "distance" between q_k and x_k; in particular, to show that q_k is eliminated along $x_k(t)$ it suffices to show that $D_{\mathrm{KL}}(q_k, x_k(t)) \to +\infty$. Following Bregman [11], the same role for a steep penalty function $h\colon \Delta \to \mathbb{R}$ is played by the so-called *Bregman divergence*:

$$D_h(q, x) = h(q) - h(x) - \langle dh(x) | q - x \rangle, \qquad q \in \Delta,\ x \in \Delta^\circ, \tag{4.3}$$

where $dh(x)$ denotes the differential of h at x (so $D_h(q, x)$ is just the difference between $h(q)$ and the estimate of $h(q)$ based on linearization at x).[13]

On the other hand, since (RLD) may fail to be well-posed if the players' penalty functions are not steep, we must analyze the reinforcement learning dynamics (RL) directly on the dual space V^* where the score variables y evolve. To do so, we introduce the *Fenchel coupling* between $q_k \in \mathcal{X}_k$ and $y_k \in V_k^*$, defined as

$$P_k(q_k, y_k) = h_k(q_k) + h_k^*(y_k) - \langle y_k | q_k \rangle, \tag{4.4}$$

where

$$h_k^*(y_k) = \max_{x_k \in \mathcal{X}_k} \{\langle y_k | x_k \rangle - h_k(x_k)\} \tag{4.5}$$

denotes the convex conjugate of the penalty function $h_k\colon \mathcal{X}_k \to \mathbb{R}$ of player k. Our choice of terminology above simply reflects the fact that $P_k(q_k, y_k)$ collects all the

[13] One can easily verify that the Bregman divergence (4.3) of the Gibbs penalty (2.9) is simply the standard Kullback–Leibler divergence.

terms of Fenchel's inequality, so it is non-negative and (strictly) convex in both arguments. Furthermore, we show in Proposition B.3 that $P_k(q_k, y_k)$ is equal to the associated Bregman divergence between q_k and $x_k = Q_k(y_k)$ when the latter is interior and that $P_k(q_k, y_k)$ provides a proximity measure between q_k and $Q_k(y_k)$ which is applicable even when h_k is not steep.

Proof of Theorem 4.1. Assume first that $q_k \in \mathcal{X}_k$ is dominated by $q'_k \in \mathcal{X}_k$ and let $\Lambda_k = \{x_k \in \mathcal{X}_k : x_{k\alpha} = 0 \text{ for some } \alpha \in \operatorname{supp}(q_k)\}$ be the union of all faces of $\mathcal{X}_k$ that do not contain q_k. By definition, q_k becomes extinct along $x(t)$ if and only if $x_k(t) \to \Lambda_k$ as $t \to \infty$; therefore, in view of Proposition B.4, it suffices to show that $P_k(q_k, y_k(t)) \to +\infty$.

To that end, consider the "cross-coupling"

$$V_k(y_k) = P_k(q_k, y_k) - P_k(q'_k, y_k) = h_k(q_k) - h_k(q'_k) - \langle y_k | q_k - q'_k \rangle. \tag{4.6}$$

Under the dynamics (RL), we then have:

$$\frac{d}{dt} V_k(y_k(t)) = -\langle v_k(x(t)) | q_k - q'_k \rangle = u_k(q'_k; x_{-k}(t)) - u_k(q_k; x_{-k}(t)) \geq \delta_k > 0, \tag{4.7}$$

where $\delta_k = \min_{x \in \mathcal{X}} \{u_k(q'_k; x_{-k}) - u_k(q_k; x_{-k})\}$ denotes the minimum payoff difference between q_k and q'_k. Hence, with $P_k(q'_k, y_k) \geq 0$ for all $y_k \in V_k^*$ (Proposition B.3), we readily obtain

$$P_k(q_k, y_k(t)) \geq V_k(y_k(0)) + \delta_k t, \tag{4.8}$$

so every ω-limit of $x_k(t)$ belongs to Λ_k by Proposition B.4, i.e. q_k becomes extinct.

To show that iteratively dominated strategies become extinct, we proceed by induction on the rounds of elimination of dominated strategies. More precisely, let $\mathcal{X}_k^r \subseteq \mathcal{X}_k$ denote the space of mixed strategies of player k that survive r rounds of elimination so that all strategies $q_k \notin \mathcal{X}_k^r$ become extinct along $x(t)$; in particular, if $\alpha \notin \mathcal{A}_k^r \equiv \mathcal{A}_k \cap \mathcal{X}_k^r$, this implies that $x_{k\alpha}(t) \to 0$ as $t \to \infty$. Assume further that $q_k \in \mathcal{X}_k^r$ survives for r elimination rounds but dies on the subsequent one, so there exists some $q'_k \in \mathcal{X}_k^r$ with $u_k(q'_k; x_{-k}) > u_k(q_k; x_{-k})$ for all $x \in \mathcal{X}^r = \prod_\ell \mathcal{X}_\ell^r$. With this in mind, decompose $x \in \mathcal{X}$ as $x = x^r + z^r$ where x^r is the (Euclidean) projection of x on the subspace of $\mathcal{X}$ spanned by the surviving pure strategies $\mathcal{A}_\ell^r$, $\ell \in \mathcal{N}$. Our induction hypothesis implies $z^r(t) = x(t) - x^r(t) \to 0$ as $t \to \infty$ (recall that $x_{k\alpha}(t) \to 0$ for all $\alpha \notin \mathcal{A}_k^r$), so, for large enough t, we have

$$|\langle v_k(x(t)) | q'_k - q_k \rangle - \langle v_k(x^r(t)) | q'_k - q_k \rangle| < \delta_k^r / 2, \tag{4.9}$$

where $\delta_k^r = \min_{x^r \in \mathcal{X}^r} \{u_k(q'_k; x^r_{-k}) - u_k(q_k; x^r_{-k})\}$.

By combining the above, we get

$$\langle v_k(x(t)) | q'_k - q_k \rangle = u_k(q'_k; x(t)) - u_k(q_k; x(t)) > \delta_k^r / 2 > 0 \tag{4.10}$$

for large t, and our claim follows by plugging this last estimate into (4.7) and arguing as in the base case $r = 0$. □

Remark 4.1. In the projection dynamics of Nagurney and Zhang [43], dominated strategies need not be eliminated: although such strategies are selected against at interior states, the dynamics' solutions may enter and leave the boundary of $\mathcal{X}$ in perpetuity, allowing dominated strategies to survive (Sandholm et al. [51]). Similarly, there exist Carathéodory solutions of the projection dynamics (PD) that do not eliminate dominated strategies (for instance, stationary trajectories at vertices

corresponding to dominated strategies). By contrast, Theorem 4.1 shows that dominated strategies become extinct along *every* weak solution orbit $x(t) = \operatorname{proj}_{\mathcal{X}} y(t)$ of (PD) that is induced by the projected reinforcement learning scheme (PL).

Remark 4.2. Theorem 4.1 imposes no restrictions on different players' choices of regularized best response maps. For instance, dominated strategies become extinct even if some players use the exponential learning scheme (XL) while others employ the projection-driven process (PL). In fact, the proof of Theorem 4.1 shows that the elimination of a player's dominated strategies is a unilateral result: if a player follows (RL), he ceases to play dominated strategies irrespective of what other players are doing.

We now turn to the *rate* of elimination of dominated strategies. In the case of the replicator dynamics, this rate is known to be exponential: if $\alpha \prec \beta$, then $x_\alpha(t) = \mathcal{O}(\exp(-ct))$ for some $c > 0$ – see e.g. Weibull [59]. We show next that the rate of elimination of dominated strategies under (RL) depends crucially on the players' choice of penalty function; in fact, if the players' penalty functions are nowhere steep, dominated strategies become extinct in *finite* time. For simplicity, we focus on the decomposable case:

Proposition 4.2. *Let $x(t) = Q(y(t))$ be an orbit of the dynamics (RL$_\gamma$) and assume that the players' penalty functions are of the form $h_k(x_k) = \sum_\beta^k \theta_k(x_{k\beta})$ for some $\theta_k \colon [0,1] \to \mathbb{R}$ as in (2.6). If $\alpha \prec \beta$, then*

$$x_{k\alpha}(t) \leq [\phi_k(c_k - \gamma_k \delta_k t)]_+ , \tag{4.11}$$

where $\phi_k = (\theta_k')^{-1}$, c_k is a constant that only depends on the initial conditions of (RL) and $\delta_k = \min\{v_{k\beta}(x) - v_{k\alpha}(x) : x \in \mathcal{X}\}$ is the minimum payoff difference between α and β.

In particular, if $\theta_k'(0)$ is finite, dominated strategies become extinct in finite time.

Proof. By the definition of the reinforcement learning dynamics (RL$_\gamma$), we have:

$$\dot{y}_{k\alpha} - \dot{y}_{k\beta} = \gamma_k \left[v_{k\alpha}(x(t)) - v_{k\beta}(x(t))\right] \leq -\gamma_k \delta_k, \tag{4.12}$$

and hence:

$$y_{k\alpha}(t) - y_{k\beta}(t) \leq y_{k\alpha}(0) - y_{k\beta}(0) - \gamma_k \delta_k t. \tag{4.13}$$

On the other hand, by the KKT conditions (A.2) for the softmax problem (2.7), we obtain $\theta_k'(x_{k\alpha}) - \theta_k'(x_{k\beta}) \leq y_{k\alpha} - y_{k\beta}$ whenever $x_{k\alpha} = Q_{k\alpha}(y_k) > 0$. Since θ_k' is bounded above on $(0,1]$, (4.13) gives

$$\theta_k'(x_{k\alpha}(t)) \leq c_k - \gamma_k \delta_k t \tag{4.14}$$

for some $c_k \in \mathbb{R}$ and for all t such that $x_{k\alpha}(t) > 0$, so (4.11) follows. □

Remark 4.3. Proposition 4.2 shows that a player's penalty function can be reverse-engineered in terms of the desired rate of elimination of dominated strategies: to achieve a target extinction rate ϕ, it suffices to pick a penalty kernel θ such that $\theta' = \phi^{-1}$ (cf. Table 1). For instance, the Gibbs kernel $\theta(x) = x \log x$ of (2.9) yields the exponential extinction rate $\exp(c - \delta t)$ whereas the quadratic kernel (2.11) gives the bound $[c - \delta t]_+$ which shows that (PL) eliminates dominated strategies in finite time.

Finally, for weakly dominated strategies, we obtain the following conditional extinction result in the spirit of Weibull [59, Proposition 3.2]:

DYNAMICS		PENALTY KERNEL $\theta(x)$	DECAY RATE $\phi(-y)$
PROJECTION	(PL)	$\frac{1}{2}x^2$	$-y$
REPLICATOR	(RD)	$x \log x$	$\exp(-y)$
q-REPLICATOR	(RD_q)	$\big[q(1-q)\big]^{-1}(x - x^q)$	$\big[q^{-1}+(1-q)y\big]^{1/(q-1)}$
LOG-BARRIER	(LD)	$-\log x$	$1/y$

TABLE 1. Rates of extinction of dominated strategies and convergence to strict equilibria under the dynamics (RL_γ) for different penalty kernels θ. If $\alpha \prec \beta$ and $v_{k\beta} - v_{k\alpha} \geq \delta$, the rate function ϕ is such that $x_{k\alpha}(t) \leq [\phi(c - \gamma\delta t)]_+$ for some $c \in \mathbb{R}$. Otherwise, if $\alpha \in \mathcal{A}_k$ is the component of a strict equilibrium x^* and $v_{k,0}(x^*) - v_{k\beta}(x^*) \geq \delta > 0$ for all $\beta \in \mathcal{A}_k \setminus \{\alpha\}$, the rate function ϕ is such that $1 - x_{k\alpha}(t) \sim [\phi(-\gamma\delta t)]_+$ for large t. If the penalty kernel θ is not steep at 0 (as in the projection and q-replicator dynamics for $q > 1$), extinction of dominated strategies and/or convergence to a strict equilibrium occurs in finite time.

Proposition 4.3. *Let $x(t) = Q(y(t))$ be an orbit of (RL) in $\mathcal{X}$ and let $q_k \preccurlyeq q'_k$. Then, q_k becomes extinct along $x(t)$ or every $\alpha_{-k} \in \mathcal{A}_{-k} \equiv \prod_{\ell \neq k} \mathcal{A}_\ell$ such that $u_k(q_k; \alpha_{-k}) < u_k(q'_k; \alpha_{-k})$ becomes extinct along $x(t)$.*

Remark 4.4. The "or" above is not exclusive: the two extinction clauses could occur simultaneously – see e.g. Weibull [59, Proposition 3.2] for the case of the replicator dynamics.

Proof of Proposition 4.3. With notation as in the proof of Theorem 4.1, we have:

$$\dot{V}_k = \langle v_k(x(t)) | q'_k - q_k \rangle = \sum\nolimits_{\alpha_{-k} \in \mathcal{A}'_{-k}} [u_k(q'_k; \alpha_{-k}) - u_k(q_k; \alpha_{-k})]\, x_{\alpha_{-k}}(t), \quad (4.15)$$

where $x_{\alpha_{-k}} \equiv \prod_{\ell \neq k} x_{\ell,\alpha_\ell}$ denotes the α_{-k}-th component of x and $\mathcal{A}'_{-k} = \{\alpha_{-k} \in \mathcal{A}_k : u_k(q_k; \alpha_{-k}) < u_k(q'_k; \alpha_{-k})\}$. Integrating with respect to t, we see that $V_k(y(t))$ remains bounded if and only if the integrals $\int_0^\infty x_{\alpha_{-k}}(t)\, dt$ are all finite. However, with $\dot{x}_{\alpha_{-k}}(t)$ essentially bounded, the same argument as in the proof of Weibull [59, Prop. 3.2] shows that $\int_0^\infty x_{\alpha_{-k}}(t)\, dt$ remains finite if $x_{\alpha_{-k}}(t) \to 0$ as $t \to \infty$. If this is not the case, we have $\lim_{t\to\infty} V_k(y(t)) = +\infty$ and Proposition B.4 shows that q_k becomes extinct. □

5. EQUILIBRIUM, STABILITY AND CONVERGENCE

We now turn to the long-term stability and convergence properties of the reinforcement dynamics (RL). Our analysis focuses on *Nash equilibria*, i.e. strategy profiles $x^* = (x_1^*, \dots, x_N^*) \in \mathcal{X}$ that are unilaterally stable in the sense that

$$u_k(x^*) \geq u_k(x_k; x_{-k}^*) \quad \text{for all } x_k \in \mathcal{X}_k \text{ and for all } k \in \mathcal{N}, \quad (5.1)$$

or, equivalently:

$$v_{k\alpha}(x^*) \geq v_{k\beta}(x^*) \quad \text{for all } \alpha \in \operatorname{supp}(x_k^*) \text{ and for all } \beta \in \mathcal{A}_k,\ k \in \mathcal{N}. \quad (5.2)$$

If (5.1) is strict for all $x_k \neq x_k^*$, $k \in \mathcal{N}$, we say that x^* a *strict equilibrium*. Finally, equilibria of restrictions of $\mathfrak{G}$ are called *restricted equilibria* of $\mathfrak{G}$; in particular, x^*

is a restricted equilibrium of $\mathfrak{G}$ if (5.1) holds for every $k \in \mathcal{N}$ and for all x_k with $\operatorname{supp}(x_k) \subseteq \operatorname{supp}(x_k^*)$.

Some basic long-term stability and convergence properties of the replicator dynamics for (asymmetric) normal form games can be summarized as follows:

(1) Nash equilibria are stationary.
(2) If an interior solution orbit converges, its limit is Nash.
(3) If a point is Lyapunov stable, then it is Nash.
(4) Strict equilibria are asymptotically stable.

Our aim in this section is to establish analogous results for the reinforcement learning scheme (RL). That said, since (RL) does not evolve directly on $\mathcal{X}$ (and the induced strategy-based dynamics (RLD) are well-posed only when the players' penalty functions are steep), the standard notions of stability and stationarity must be modified accordingly.[14]

Definition 5.1. Let $x^* \in \mathcal{X}$ and let $y(t)$ be a solution orbit of (RL). We say that:

(1) x^* is *stationary* under (RL) if $x^* \in \operatorname{im} Q$ and $Q(y(t)) = x^*$ for all $t \geq 0$ whenever $Q(y(0)) = x^*$.
(2) x^* is *Lyapunov stable* under (RL) if, for every neighborhood U of x^*, there exists a neighborhood U' of x^* such that $Q(y(t)) \in U$ for all $t \geq 0$ whenever $Q(y(0)) \in U'$.
(3) x^* is *attracting* under (RL) if x^* admits a neighborhood U such that $Q(y(t)) \to x^*$ as $t \to \infty$ whenever $Q(y(0)) \in U$.
(4) x^* is *asymptotically stable* under (RL) if it is Lyapunov stable and attracting.

Remark 5.1. In Definition 5.1, stationary points $x^* \in \mathcal{X}$ are explicitly required to belong to the image of the players' choice map Q, but no such assumption is made for stable states. From a propositional point of view, this is done to ensure that points $x^* \in \mathcal{X} \setminus \operatorname{im} Q$ are not called stationary vacuously. From a dynamical standpoint, stationary points should themselves be (constant) trajectories of the dynamical system under study, whereas Lyapunov stable and attracting states only need to be *approachable* by trajectories.

Since $\operatorname{im} Q \supseteq \mathcal{X}^\circ$, any point in $\mathcal{X}$ can be a candidate for (asymptotic) stability under (RL); however, boundary points might not be suitable candidates for stationarity, so stability does *not* imply stationarity (as would be the case for a dynamical system defined on $\mathcal{X}$). In particular, recall that $x^* \notin \operatorname{im} Q$ if and only if some player's penalty function is steep at x^*. As such, in the (steep) example of exponential learning, $x^* \in \mathcal{X}$ is stationary under (XL) if and only if it is an *interior* stationary point of the replicator dynamics (RD); by contrast, in the (non-steep) projection setting of (PD), *any* point in $\mathcal{X}$ may be stationary.

With this definition at hand, we obtain:

Theorem 5.2. *Let $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ be a finite game, let $x^* \in \mathcal{X}$, and let $x(t) = Q(y(t))$ be an orbit of (RL) in $\mathcal{X}$.*

[14] The standard stability notions continue to apply in the dual space V^* where $y(t)$ evolves; however, since the mapping $Q\colon V^* \to \mathcal{X}$ which defines the trajectories of play $x(t) = Q(y(t))$ in $\mathcal{X}$ is neither injective nor surjective, this approach would not suffice to define stationarity and stability on $\mathcal{X}$.

I. *If x^* is stationary under (RL), then it is a Nash equilibrium of $\mathfrak{G}$; conversely, if x^* is a Nash equilibrium of $\mathfrak{G}$ and $x^* \in \operatorname{im} Q$, x^* is stationary under (RL).*
II. *If $\lim_{t\to\infty} x(t) = x^*$, then x^* is a Nash equilibrium of $\mathfrak{G}$.*
III. *If $x^* \in \mathcal{X}$ is Lyapunov stable under (RL), then x^* is a Nash equilibrium of $\mathfrak{G}$.*
IV. *If x^* is a strict Nash equilibrium of $\mathfrak{G}$, then it is also asymptotically stable under (RL).*

For steep and decomposable penalty functions, Parts I, III and IV of Theorem 5.2 essentially follow from Theorem 1 in Coucheney et al. [14] (see also Laraki and Mertikopoulos [35, 36] for related results in a second order setting). Our proofs mimic those of Coucheney et al. [14], but the lack of steepness means that we must work directly on the dual space V^* of the score variables y_k and rely on the properties of the Fenchel coupling.

To prove Theorem 5.2, we need the following result (which is of independent interest):

Proposition 5.3. *If every neighborhood U of $x^* \in \mathcal{X}$ admits an orbit $x_U(t) = Q(y_U(t))$ of (RL) such that $x_U(t) \in U$ for all $t \geq 0$, then x^* is a Nash equilibrium.*

Proof. Assume ad absurdum that x^* is not Nash, so $v_{k\alpha}(x^*) < v_{k\beta}(x^*)$ for some player $k \in \mathcal{N}$ and for some $\alpha \in \operatorname{supp}(x_k^*)$, $\beta \in \mathcal{A}_k$. Moreover, let U be a sufficiently small neighborhood of x^* in $\mathcal{X}$ such that $v_{k\beta}(x) - v_{k\alpha}(x^*) \geq \delta > 0$ for all $x \in U$. Then, if $x(t) = Q(y(t))$ is an orbit of (RL) in $\mathcal{X}$ that is contained in U for all $t \geq 0$, we will have

$$y_{k\alpha}(t) - y_{k\beta}(t) = y_{k\alpha}(0) - y_{k\beta}(0) + \int_0^t [v_{k\alpha}(x(s)) - v_{k\beta}(x(s))]\, ds \leq c - \delta t, \quad (5.3)$$

for some $c \in \mathbb{R}$ and for all $t \geq 0$. This shows that $y_{k\alpha}(t) - y_{k\beta}(t) \to -\infty$ so, by Proposition A.1, we must also have $\lim_{t\to\infty} x_{k\alpha}(t) = 0$. This contradicts our original assumption that $x(t)$ remains in a small enough neighborhood of x^* (recall that $\alpha \in \operatorname{supp}(x_k^*)$ by assumption), so x^* must be Nash. □

With this result at hand, we may now proceed with the proof of Theorem 5.2:

Proof of Theorem 5.2.

Part I.. If x^* is stationary under (RL), then $Q(y(t)) = x^*$ for some $y(0) \in V^*$ and for all $t \geq 0$; this shows that x^* satisfies the hypothesis of Proposition 5.3, so x^* must be a Nash equilibrium of $\mathfrak{G}$. Conversely, assume that x^* is Nash with $x^* = Q(y(0))$ for some initial $y(0) \in V^*$; we then claim that the trajectory $y(t)$ with $y_{k\alpha}(t) = y_{k\alpha}(0) + v_{k\alpha}(x^*)t$ is the unique solution of (RL) starting at $y(0)$. Indeed, since $v_{k\beta}(x^*) \leq v_{k\alpha}(x^*)$ for all $\alpha \in \operatorname{supp}(x_k^*)$ and for all $\beta \in \mathcal{A}_k$, we have $y_{k\alpha}(t) = y_{k\alpha}(0) + c_k t - d_{k\alpha} t$ where $d_{k\alpha} = 0$ if $\alpha \in \operatorname{supp}(x_k^*)$ and $d_{k\alpha} \geq 0$ otherwise. Proposition A.1 shows that $Q_k(y_k(t)) = x_k^*$, so $y(t)$ satisfies (RL) and our assertion follows by the well-posedness of (RL).

Parts II and III.. The convergence and stability assumptions of Parts II and III both imply the hypothesis of Proposition 5.3 so x^* must be a Nash equilibrium of $\mathfrak{G}$.

PART IV.. Relabeling strategies if necessary, assume that $x^* = (\alpha_{1,0}, \dots, \alpha_{N,0})$ is a strict equilibrium of $\mathfrak{G}$, let $\mathcal{A}'_k = \mathcal{A}_k \setminus \{\alpha_{k,0}\}$, and consider the relative score variables:

$$z_{k\mu} = y_{k\mu} - y_{k,0}, \quad \mu \in \mathcal{A}'_k, \tag{5.4}$$

so that

$$\dot{z}_{k\mu} = v_{k\mu}(x) - v_{k,0}(x). \tag{5.5}$$

Proposition A.1 shows that $x_{k\mu} \to 0$ whenever $z_{k\mu} \to -\infty$, so we also have $x \to x^*$ if $z_{k\mu} \to -\infty$ for all $\mu \in \mathcal{A}'_k$, $k \in \mathcal{N}$. Moreover, given that x^* is a strict equilibrium, the RHS of (5.5) is negative if x is close to x^*; as such, the main idea of our proof will be to show that the relative scores $z_{k\mu}$ escape to negative infinity when they are not too large to begin with (i.e. when $x(0)$ is close enough to x^*).

To make this precise, let $\delta_k = \min_{\mu \in \mathcal{A}'_k} \{v_{k,0}(x^*) - v_{k\mu}(x^*)\} > 0$ and let $\varepsilon > 0$ be such that $v_{k\mu}(x) - v_{k,0}(x) < -\delta_k/2$ for all x with $\|x - x^*\|^2 < \varepsilon$. Furthermore, let

$$P_h(x^*, y) = \sum\nolimits_k P_k(x^*_k, y_k) = \sum\nolimits_k [h_k(x^*_k) + h^*_k(y_k) - \langle y_k | x^*_k \rangle] \tag{5.6}$$

denote the Fenchel coupling between x^* and y (cf. Appendix B), and set

$$U^*_\varepsilon = \{y \in V^* : P_h(x^*, y) < \varepsilon K_{\min}/2\} \tag{5.7}$$

where $K_{\min} > 0$ is the smallest strong convexity constant of the players' penalty functions. Proposition B.3 gives $\|Q(y) - x^*\|^2 \leq 2K_{\min}^{-1} P_h(x^*, y) < \varepsilon$, so we will also have $v_{k\mu}(Q(y)) - v_{k,0}(Q(y)) < -\delta_k/2$ for all $y \in U^*_\varepsilon$.

In view of the above, let $y(t)$ be a solution of (RL) with $y(0) \in U^*_\varepsilon$ and let $\tau_\varepsilon = \inf\{t : y(t) \notin U^*_\varepsilon\}$ be the first exit time of $y(t)$ from U^*_ε. Then, if $\tau_\varepsilon < \infty$:

$$z_{k\mu}(\tau_\varepsilon) = z_{k\mu}(0) + \int_0^{\tau_\varepsilon} [v_{k\mu}(x(s)) - v_{k,0}(x(s))]\, ds \leq z_{k\mu}(0) - \tfrac{1}{2}\delta_k \tau_\varepsilon < z_{k\mu}(0), \tag{5.8}$$

for all $\mu \in \mathcal{A}'_k$ and for all $k \in \mathcal{N}$. Intuitively, since the score differences $y_{k,0} - y_{k\mu}$ grow with t, $x(\tau_\varepsilon)$ must be closer to x^* than $x(0)$, meaning that $y(\tau_\varepsilon) \in U^*_\varepsilon$, a contradiction. More rigorously, note that $P_k(x^*_k, y_k) = h_k(x^*_k) + h^*_k(y_k) - y_{k,0}$; since $\langle y_k | x_k \rangle = y_{k,0} + \sum_{\mu \in \mathcal{A}'_k} x_{k\mu}(y_{k\mu} - y_{k,0})$, we will also have $h^*_k(y_k) - y_{k,0} = \max_{x_k \in \mathcal{X}_k} \{\sum_{\mu \in \mathcal{A}'_k} x_{k\mu} z_{k\mu} - h_k(x_k)\}$, so $h^*_k(y'_k) - y'_{k,0} < h^*_k(y_k) - y_{k,0}$ whenever $z'_{k\mu} < z_{k\mu}$ for all $\mu \in \mathcal{A}'_k$. In this way, (5.8) yields the contradictory statement $\varepsilon \leq P_h(q, y(\tau_\varepsilon)) < P_h(q, y(0)) < \varepsilon$, so $y(t)$ must remain in U^*_ε for all $t \geq 0$. The estimate (5.8) then shows that $\lim_{t\to\infty} z_{k\mu}(t) = -\infty$, so $\lim_{t\to\infty} x_{k\mu}(t) = 0$ by Proposition A.1. Therefore, $\lim_{t\to\infty} x(t) = x^*$.

The above shows that $y(t)$ is contained in U^*_ε and $Q(y(t)) \to x^*$ whenever $y(0) \in U^*_\varepsilon$. To complete the proof, let $U_\varepsilon = \{x \in \mathcal{X} : \sum_k D_{h_k}(x^*_k, x_k) < \varepsilon K_{\min}/2\}$ where $D_{h_k}(x^*_k, x_k)$ is the Bregman divergence (B.4) between x_k and x^*_k. Propositions B.2 and B.3 show that U_ε is a neighborhood of x^* in $\mathcal{X}$ with $Q(U^*_\varepsilon) \subseteq U_\varepsilon$ and $Q^{-1}(U_\varepsilon) = U^*_\varepsilon$, so x^* is asymptotically stable under (RL). □

Remark 5.2. In the (asymmetric) replicator dynamics (RD), it is well known that only strict equilibria can be attracting – so strict Nash equilibria and asymptotically stable states coincide. One can extend this equivalence to (RL) by using restrictions of $\mathfrak{G}$ with smaller strategy sets to define the mixed strategy dynamics (RLD) on the faces of $\mathcal{X}$ – for a related discussion, see Coucheney et al. [14].

Theorem 5.2 shows that the reinforcement learning scheme (RL) exhibits essentially the same long-run properties as the benchmark replicator dynamics. That said, from a quantitative viewpoint, the situation can be quite different: as we show below, the rate of convergence of (RL) to strict equilibria depends crucially on the players' penalty functions, and convergence can occur in finite time:

Proposition 5.4. *Let $x(t) = Q(y(t))$ be an orbit of (RL$_\gamma$) in $\mathcal{X}$ and assume that the players' penalty functions are of the form $h_k(x_k) = \sum_\beta^k \theta_k(x_{k\beta})$ for some $\theta_k\colon [0,1] \to \mathbb{R}$ as in (2.6). If $x^* = (\alpha_{1,0}, \dots, \alpha_{N,0})$ is a strict Nash equilibrium and $x(0)$ is sufficiently close to x^*, then:*

$$1 - x_{k,0}(t) \sim [\phi_k(-\gamma_k \delta_k t)]_+ , \tag{5.9}$$

where $\phi_k = (\theta_k')^{-1}$ and $\delta_k = \min\{v_{k,0}(x^*) - v_{k\mu}(x^*) : \mu \in \mathcal{A}_k, \mu \neq \alpha_{k,0}\}$.

In particular, if $\theta_k'(0)$ is finite, convergence occurs in finite time.

Proof. Pick some $\varepsilon > 0$ and let U be a neighborhood of x^* in $\mathcal{X}$ such that $v_{k,0}(x) - v_{k\mu}(x)$ remains within $(1 \pm \varepsilon)$ of $v_{k,0}(x^*) - v_{k\mu}(x^*)$ for all $x \in U$ and for all $\mu \in \mathcal{A}_k' = \mathcal{A}_k \setminus \{\alpha_{k,0}\}$. If $x(0)$ is sufficiently close to x^*, $x(t)$ will remain in U for all $t \geq 0$ by Theorem 5.2; hence, the same reasoning as in the proof of Proposition 4.2 yields

$$x_{k\mu}(t) \leq [\phi_k(c_{k\mu} - (1-\varepsilon)\gamma_k \delta_{k\mu} t)]_+ , \tag{5.10}$$

where $\delta_{k\mu} = v_{k,0}(x^*) - v_{k\mu}(x^*)$ and $c_{k\mu}$ is a constant depending on initial conditions. Since ε can be taken arbitrarily small and $x(t) \to x^*$ as $t \to \infty$, the above yields $x_{k\mu}(t) \sim [\phi_k(-\gamma_k \delta_{k\mu} t)]_+$ for large t; thus, letting $\delta_k = \min_\mu \delta_{k\mu}$, we finally obtain:

$$1 - x_{k,0}(t) = \sum\nolimits_{\mu \in \mathcal{A}_k'} x_{k\mu}(t) \sim \sum\nolimits_{\mu \in \mathcal{A}_k'} [\phi_k(-\gamma_k \delta_{k\mu} t)]_+ \sim [\phi_k(-\gamma_k \delta_k t)]_+, \tag{5.11}$$

which completes our proof. □

6. Time Averages and the Best Response Dynamics

As is well known, the replicator dynamics (RD) do not converge to equilibrium when the game's only Nash equilibrium is interior (for instance, as in Matching Pennies and generic zero-sum games). On the other hand, if a replicator trajectory stays away from the boundary of the simplex in a 2-player normal form game, its time-average converges to the game's Nash set – see e.g. Hofbauer and Sigmund [26, Chap. 7].

Under the reinforcement learning dynamics (RL), trajectories of play may enter and exit the boundary of the game's strategy space in perpetuity so the "permanence" criterion of staying a bounded distance away from $\mathrm{bd}(\mathcal{X})$ is no longer natural. Instead, the conclusion about convergence of time averages can be reached by requiring that differences between the scores of each player's strategies remain bounded:

Theorem 6.1. *Let $x(t) = Q(y(t))$ be an orbit of (RL) in $\mathcal{X}$ for a 2-player game $\mathfrak{G}$. If the score differences $y_{k\alpha}(t) - y_{k\beta}(t)$ remain bounded for all $\alpha, \beta \in \mathcal{A}_k$, $k = 1, 2$, the time average $\bar{x}(t) = t^{-1} \int_0^t x(s)\, ds$ of $x(t)$ converges to the set of Nash equilibria of $\mathfrak{G}$.*

As with the classic result on time averages, our proof relies on the linearity (as opposed to multilinearity) of each player's payoff function, a property specific to 2-player games.

Proof. By the definition of the dynamics (RL) we have:

$$\begin{aligned} y_{k\alpha}(t) - y_{k\beta}(t) &= c_{\alpha\beta} + \int_0^t [v_{k\alpha}(x(s)) - v_{k\beta}(x(s))]\, ds \\ &= c_{\alpha\beta} + t \cdot [v_{k\alpha}(\bar{x}(t)) - v_{k\beta}(\bar{x}(t))]\,, \end{aligned} \tag{6.1}$$

where $c_{\alpha\beta} = y_{k\alpha}(0) - y_{k\beta}(0)$ and we used linearity to bring the integral inside the argument of $v_{k\alpha}$ and $v_{k\beta}$. Thus, dividing by t and taking the limit $t \to \infty$, we get

$$\lim_{t\to\infty} [v_{k\alpha}(\bar{x}(t)) - v_{k\beta}(\bar{x}(t))] = 0, \tag{6.2}$$

where we have used the assumption that $y_{k\alpha} - y_{k\beta}$ is bounded. Hence, if x^* is an ω-limit of $x(t)$, we will have $v_{k\alpha}(x^*) = v_{k\beta}(x^*)$ for all $\alpha, \beta \in \mathcal{A}_k$, $k = 1, 2$, so x^* must be a Nash equilibrium of $\mathfrak{G}$. Since $\mathcal{X}$ is compact, the ω-limit set of $x(t)$ is nonempty and our assertion follows. □

Remark 6.1. To see how Theorem 6.1 implies the corresponding result for the replicator dynamics, simply note that $y_{k\alpha} - y_{k\beta} = \log x_{k\alpha} - \log x_{k\beta}$ under (XL). Therefore, if $x(t)$ stays away from the boundary of the simplex, the requirement of Theorem 6.1 is fulfilled and we recover the classical result of Hofbauer and Sigmund [26].

The standard example of a 2-player game that cycles under the replicator dynamics is the zero-sum game of Matching Pennies (MP) with payoff bimatrix:

$$U_{\mathsf{MP}} = \begin{pmatrix} (1,-1) & (-1,1) \\ (-1,1) & (1,-1) \end{pmatrix}. \tag{6.3}$$

In particular, the unique equilibrium of the game is $x_1^* = x_2^* = (1/2, 1/2)$, and it can be shown that the Kullback–Leibler divergence

$$D_{\mathrm{KL}}(x^*, x) = \sum\nolimits_k \sum\nolimits_\alpha^k x_{k\alpha}^* \log\left(x_{k\alpha}^*/x_{k\alpha}\right) \tag{6.4}$$

is a constant of motion for (RD). As a result, replicator trajectories always stay away from the boundary of $\mathcal{X}$, so their time averages converge to the game's (unique) equilibrium. In our more general setting, we obtain:

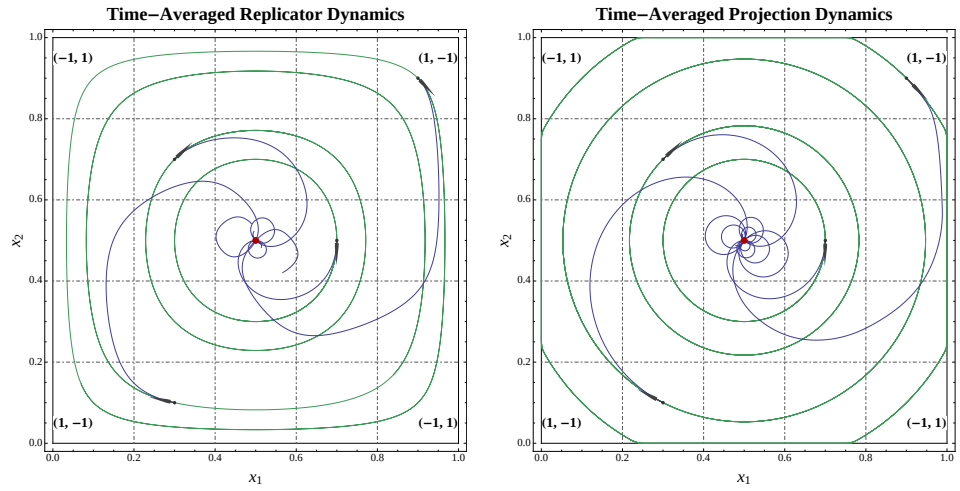
Proposition 6.2. *Let $\mathfrak{G}$ be a 2-player zero-sum game, i.e. $u_1 = -u_2$. Then, the time average $\bar{x}(t) = t^{-1}\int_0^t x(s)\, ds$ of every orbit $x(t) = Q(y(t))$ of (RL) in $\mathcal{X}$ converges to the Nash set of $\mathfrak{G}$.*

Proof. Let x^* be an interior Nash equilibrium and let $P_h(x^*, y) = \sum_{k=1,2} \left[h_k(x_k^*) + h_k^*(y_k) - \langle y_k | x_k^*\rangle\right]$ denote the Fenchel coupling between x^* and y (cf. Appendix B). Then, by Lemma B.6, we get:

$$\begin{aligned} \frac{d}{dt} P_h(x^*, y) &= \langle v_1(x) | x_1 - x_1^*\rangle + \langle v_2(x) | x_2 - x_2^*\rangle \\ &= u_1(x_1, x_2) - u_1(x_1^*, x_2) + u_2(x_1, x_2) - u_2(x_1, x_2^*) = 0 \end{aligned} \tag{6.5}$$

on account of the game being zero-sum.

The above shows that $P_h(x^*, y(t))$ remains constant along (RL). Proposition B.5 then implies that $y_\alpha(t) - y_\beta(t)$ is bounded, so our assertion follows from Theorem 6.1. □



(a) Exponential reinforcement learning.

(b) Projected reinforcement learning.

FIGURE 2. Time averages in Matching Pennies under the exponential learning scheme (XL) and the projected reinforcement learning dynamics (PL). In both cases, solution trajectories (green) cycle, either avoiding the boundary $\mathrm{bd}(\mathcal{X})$ of $\mathcal{X}$ (replicator) or attaining it infinitely often (projection); on the other hand, their time averages (blue) converge to the game's equilibrium. In both cases, the dynamics have been integrated over the same time horizon, showing that (PL) evolves significantly faster than (XL).

In the case of the replicator dynamics, a heuristic explanation for the above is that time averages of replicator trajectories in 2-player games exhibit the same long-run behavior as the best-reply dynamics of Gilboa and Matsui [21]:

$$\dot{x}_k \in \mathsf{br}_k(x_k) - x_k. \tag{BRD}$$

More precisely, Hofbauer et al. [27] showed that the ω-limit set Ω of a time-averaged replicator orbit is *internally chain transitive* under (BRD): any two points $x, y \in \Omega$ may be joined by a piecewise continuous curve (a "chain") consisting of arbitrarily long pieces of orbits in Ω broken by arbitrarily small jump discontinuities (see Benaïm et al. [8] for the precise definition).

As it turns out, this property extends verbatim to the learning scheme (RL):

Theorem 6.3. *Let $x(t) = Q(y(t))$ be an orbit of (RL) in $\mathcal{X}$ for a 2-player game $\mathfrak{G}$. Then, the ω-limit set of the time average $\bar{x}(t)$ of $x(t)$ is internally chain transitive under the best reply dynamics (BRD).*

The proof of Theorem 6.3 follows closely Hofbauer et al. [27, Proposition 5.1] and relies on the following proposition – itself a generalization of (and proved in the same way as) Proposition 4.2 in Hofbauer et al. [27]:

Proposition 6.4. *Let $x(t) = Q(y(t))$ be an orbit of (RL) for a 2-player game $\mathfrak{G}$. Then, $x(t)$ lies within $\delta(t)$ of $\mathsf{br}(\bar{x}(t))$ in the uniform norm on $\mathcal{X}$ and the approximation error $\delta(t)$ tends to 0 as $t \to \infty$.*

Proof. From the definition of the dynamics (RL), and using the fact that v_k is linear in 2-player games, we readily obtain:

$$y_k(t) = y_k(0) + \int_0^t v_k(x(s))\,ds = y_k(0) + t\cdot v_k(\bar{x}(t)). \tag{6.6}$$

Hence, $x_k(t) = Q_k(y_k(t))$ is the (unique) maximizer of the strictly concave problem:

$$\begin{aligned} &\text{maximize} && \langle v_k(\bar{x}(t))|x_k'\rangle + t^{-1}\left[\langle y_k(0)|x_k'\rangle - h(x_k')\right], \\ &\text{subject to} && x_k' \in \mathcal{X}_k. \end{aligned} \tag{6.7}$$

Therefore, with $y(0)/t \to 0$ as $t\to\infty$ and h finite and continuous on $\mathcal{X}_k$, the maximum theorem of Berge [9, p. 116] shows that $x_k(t)$ lies within a vanishing distance of $\mathsf{br}_k(\bar{x}(t)) = \arg\max_{x_k'\in\mathcal{X}_k}\langle v_k(\bar{x}(t)|x_k'\rangle$, as claimed. □

Proof of Theorem 6.3. Following Hofbauer et al. [27], differentiate $\bar{x}(t)$ to obtain

$$\frac{d}{dt}\bar{x}(t) = \frac{x(t)}{t} - \frac{1}{t^2}\int_0^t x(s)\,ds = \frac{1}{t}\left(x(t) - \bar{x}(t)\right). \tag{6.8}$$

After changing time to $\tau = \log t$, this expression gives $\frac{d}{d\tau}\bar{x} = x - \bar{x}$, so Proposition 6.4 shows that $\bar{x}(\tau)$ tracks a perturbed version of the best reply dynamics (BRD) in the sense of Benaïm et al. [8, Definition II]. Our assertion then follows from Theorem 3.6 in Benaïm et al. [8]. □

We close this section with some easy corollaries of Theorem 6.3: first, if the time average of a solution orbit $x(t) = Q(y(t))$ of (RL) converges, its limit must be Nash; second, if (BRD) is globally attracted to some $x^*\in\mathcal{X}^\circ$, the time averages of (RL) also converge to x^*. These conclusions can be seen as generalizations of the corresponding statements in Hofbauer et al. [27] for *interior* replicator orbits; in our more general setting however, these conclusions hold for *every* orbit of (RL) in $\mathcal{X}$, even those that enter and then leave $\mathrm{bd}(\mathcal{X})$ – e.g. as in the Matching Pennies example of Fig. 2.

7. On Non-Regularized Best Responses

We conclude this paper with an informal discussion of the following question: what happens if regularized best responses are replaced by the *actual* best response correspondences $\mathsf{br}_k(y_k) = \arg\max_{x_k\in\mathcal{X}_k}\langle y_k|x_k\rangle$ of (2.4)?

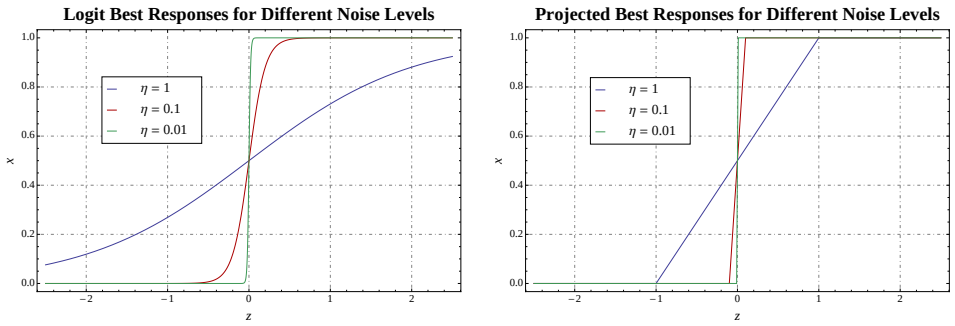
Given that $\mathsf{br}_k\colon V_k^* \rightrightarrows \mathcal{X}_k$ is multi-valued, running (RL) with ordinary best responses leads to the differential inclusion:

$$\begin{aligned} \dot{y}_k &= v_k(x), \\ x_k &\in \mathsf{br}_k(y_k), \end{aligned} \tag{7.1}$$

or, more concisely:

$$\dot{y} \in v(\mathsf{br}(y)), \tag{7.2}$$

where $\mathsf{br}\equiv(\mathsf{br}_1,\dots,\mathsf{br}_N)\colon V^*\rightrightarrows\mathcal{X}$ denotes the players' best response correspondence profile. Since this correspondence is upper hemicontinuous with (nonempty) convex values contained in a compact set, standard results imply that (7.2) admits absolutely continuous solutions that are defined for all $t\geq 0$, although multiple



(a) Logit best responses.

(b) Projected best responses.

FIGURE 3. Logit and projected best responses for different noise levels (Figs. 3(a) and 3(b) respectively). We consider a player with two actions $\mathcal{A} = \{0, 1\}$ and $z = y_1 - y_0$ denotes the score difference between the two. The player's regularized best response map $Q_1(y)$ for different noise levels $\eta = \gamma^{-1}$ is given by (3.9) for the logit case, and by $\min\{1, \max\{0, (1 + \eta^{-1}z)/2\}\}$ for the projection case.

solutions may emanate from a single initial condition (Aubin and Cellina [3]).[15] By going through the proofs of the corresponding results, it can then be shown that

(1) In generic games, any absolutely continuous solution $x(t)$ of (7.2) is piecewise constant and occupies the vertices of $\mathcal{X}$ for an open, dense set of times.
(2) Dominated strategies are eliminated under (7.2) in finite time.
(3) A point $x^* \in \mathcal{X}$ is stationary under (7.2) if and only if it is a Nash equilibrium.
(4) If x^* is a strict equilibrium, then it attracts an open set of initial conditions $y(0) \in V^*$ such that $\mathsf{br}(y(0))$ is close enough to x^*; however, since solutions of (7.2) are piecewise constant (at least, generically), this result adds little to the stationarity result above.
(5) Time-averaged trajectories of (7.2) converge to Nash equilibrium in 2-player zero-sum games: trajectories of play $x(t)$ might jump from vertex to vertex in perpetuity, but the empirical frequency of play converges to the game's set of Nash equilibria.

Perhaps the best way to understand these properties of (7.2) is to take the reinforcement learning scheme (RL) with a very small noise level $\eta \approx 0$ as in (3.10). In Fig. 3, we focus on the entropic and quadratic penalties (2.9) and (2.11), and we plot the induced regularized best response maps for different noise levels η when the player has two pure strategies. Up to a time shift, these plots also describe the behavior of the solution orbits of (3.10) for a single-player game with fixed payoff difference $v_1 - v_0 \equiv 1$ (simply note that the player's score difference z satisfies $\dot{z} = v_1 - v_0 = 1$). In the logit case (Fig. 3(a)), the weight on the dominated strategy vanishes at an exponential rate, and this rate of decay increases as the noise level η tends to 0 (cf. Table 1); however, the dominated strategy is played

[15]The same reasoning shows that the strong convexity requirement of Definition 2.1 can be replaced by strict convexity (at the cost of Lipschitz continuity for $x(t)$) or even convexity (at the cost of passing to a differential inclusion with possibly discontinuous branchings).

with positive probability for all time. In the projection case however, the player assigns positive weight on both actions only within a window whose width goes to 0 as the noise level η tends to 0 (cf. Fig. 3(b)); thus, if the player starts with a high score difference in favor of the dominated strategy, this difference will decrease linearly until it reaches that window, and the player will then transit sharply to the dominant strategy. Either way, the behavior of the dynamics (7.2) is recovered in the limit $\eta \to 0$.

Appendix A. Properties of Regularized Best Response Maps

In this appendix, we prove some intuitive properties of regularized best responses that were invoked in the main text. For simplicity, all results will be stated for the d-dimensional simplex $\Delta \equiv \Delta(\mathcal{A})$ of $V = \mathbb{R}^{d+1}$ that is spanned by the index set $\mathcal{A} = \{0, 1, \dots, d\}$. We have:

Proposition A.1. *Let h be a penalty function on Δ and let $Q\colon V^* \to \Delta$ be its induced choice map. Then:*

i. $Q(y') = Q(y)$ *for all* $y, y' \in V^*$ *such that* $y' - y \propto (1, \dots, 1)$.
ii. $Q(y - te_\beta) = Q(y)$ *for all* $t \geq 0$ *and for all* $y \in V^*$ *such that* $Q_\beta(y) = 0$.
iii. $Q_\alpha(y) \to 0$ *whenever* $y_\alpha - y_\beta \to -\infty$ *for some* $\beta \neq \alpha$.

The first part of Proposition A.1 shows that modifying all payoffs by the same amount does not change relative advantages between strategies, so regularized best responses remain invariant along $(1, \dots, 1)$. The second part shows that regularized best responses also remain unchanged if one reduces the payoff of an action that already has a strategy share of 0; finally, the last part states that the strategy share of an action becomes vanishingly small when the payoff of said action is at a great relative disadvantage to that of another action.

Proof of Proposition A.1. For the first part, assume that $y'_\alpha = y_\alpha + c$ for some $c \in \mathbb{R}$ and for all $\alpha \in \mathcal{A}$. Then:

$$\begin{aligned} Q(y') &= \arg\max_{x\in\Delta} \left\{ \langle y|x\rangle + \textstyle\sum_\beta x_\beta \cdot c - h(x) \right\} \\ &= \arg\max_{x\in\Delta} \{\langle y|x\rangle - h(x)\} = Q(y). \end{aligned} \tag{A.1}$$

For the second part, the first order Karush–Kuhn–Tucker (KKT) conditions for the regularized problem (2.7) give

$$y_\beta - \frac{\partial h}{\partial x_\beta}\bigg|_{Q(y)} = \mu - \nu_\beta \qquad (\beta = 0, \dots, m), \tag{A.2}$$

where $\mu \in \mathbb{R}$ is the Lagrange multiplier of the equality constraint $\sum_\beta x_\beta = 1$ and $\nu_\alpha \geq 0$ is the complementary slackness multiplier of the inequality constraint $x_\alpha \geq 0$ – i.e. $\nu_\alpha = 0$ for all $\alpha \in \operatorname{supp}(Q(y))$. Hence, if we let $y'_\alpha = y_\alpha - t\delta_{\alpha\beta}$ for some $\beta \notin \operatorname{supp}(Q(y))$ and set $\mu' = \mu$ and $\nu'_\alpha = \nu_\alpha + t\delta_{\alpha\beta} \geq 0$, $Q(y)$ also satisfies the KKT conditions of (2.7) for y' with Lagrange multipliers μ' and ν'_α. Since h is strictly convex, it follows that $Q(y') = Q(y)$, as claimed.

Finally, for the last part, let y_n be a sequence in V^* such that $y_{n,\alpha} - y_{n,\beta} \to -\infty$, set $x_n = Q(y_n)$, and assume (by descending to a subsequence if necessary) that $x_{n,\alpha} > \varepsilon > 0$ for all n. By definition, we then have

$$\langle y_n|x_n\rangle - h(x_n) \geq \langle y_n|x'\rangle - h(x') \tag{A.3}$$

for all $x' \in \Delta$. Therefore, if we set $x'_n = x_n + \varepsilon(e_\beta - e_\alpha)$, we readily get

$$\varepsilon(y_{n,\alpha} - y_{n,\beta}) \geq h(x_n) - h(x'_n) \geq -(h_{\max} - h_{\min}), \tag{A.4}$$

which contradicts our original assumption that $y_{n,\alpha} - y_{n,\beta} \to -\infty$. With Δ compact, the above shows that $x^*_\alpha = 0$ for any limit point x^* of x_n, i.e. $Q_\alpha(y_n) \to 0$. □

Appendix B. Bregman Divergences and the Fenchel Coupling

In this appendix, we introduce Bregman divergences and the Fenchel coupling, and we discuss their basic properties.

As before, let $h\colon \Delta \to \mathbb{R}$ be a penalty function on the unit d-dimensional simplex Δ of $V \equiv \mathbb{R}^{d+1}$. For convenience, we will treat h as an extended-real-valued function $h\colon V \to \mathbb{R}$ by setting $h(x) = +\infty$ for all $x \in V \setminus \Delta$. The *subdifferential* of h at $x \in V$ is then defined as $\partial h(x) = \{y \in V^* : h(x') \geq h(x) + \langle y | x' - x\rangle \text{ for all } x' \in V\}$ and we will say that h is *subdifferentiable at* x whenever $\partial h(x)$ is nonempty. This is always the case if $x \in \Delta^\circ$, so we have $\Delta^\circ \subseteq \operatorname{dom}\partial h \equiv \{x \in \Delta : \partial h(x) \neq \varnothing\} \subseteq \Delta$.

A key tool in our analysis is the *convex conjugate* $h^*\colon V^* \to \mathbb{R}$ of h defined as

$$h^*(y) = \max\nolimits_{x\in\Delta}\{\langle y|x\rangle - h(x)\}. \tag{B.1}$$

As it turns out, the choice map $Q\colon V^* \to \Delta$ induced by h is simply the differential of h^*:

Proposition B.1. *Let h be a penalty function on Δ. The induced choice map $Q\colon V^* \to \Delta$ is Lipschitz and $Q(y) = dh^*(y)$ for all $y \in V^*$.*

Proof. By Theorem 23.5 in Rockafellar [48], we readily obtain

$$x \in \partial h^*(y) \iff y \in \partial h(x) \iff x \in \arg\max\nolimits_{x'\in\Delta}\{\langle y|x'\rangle - h(x')\}. \tag{B.2}$$

Since the last set only contains $Q(y)$, we immediately obtain $Q(y) = dh^*(y)$. The Lipschitz property for Q then follows from the strong convexity of h – see e.g. Nesterov [45]. □

Remark B.1. Proposition B.1 is folklore in convex optimization – see e.g. Hofbauer and Sandholm [24], Nesterov [45], Shalev-Shwartz [52], Kwon and Mertikopoulos [33] and many others. Equation (B.2) also shows that the image of Q is precisely $\operatorname{dom}\partial h$, a fact which we use freely in the rest of this appendix.

Given a basepoint $q \in \Delta$, the one-sided derivative

$$h'(x; q - x) = \lim_{t\to 0^+} t^{-1}\left[h(x + t(q - x)) - h(x)\right] \tag{B.3}$$

exists for all $x \in \Delta$ and is finite whenever x lies in the relative interior of a face of Δ that also contains q. With this in mind, we define the *Bregman divergence* of h as

$$D_h(q, x) = h(q) - h(x) - h'(x; q - x) \quad \text{for all } q, x \in \Delta, \tag{B.4}$$

with $D_h(q, x)$ possibly attaining the value $+\infty$ if $h'(x; q - x) = -\infty$.[16] We then have:

[16] Usually, Bregman divergences are defined for $x \in \operatorname{dom}\partial h$ – Kiwiel [32] uses the notation D'_h to distinguish (B.4) from the original definition of Bregman [11]. The "*raison d'être*" of the more general definition (B.4) is that we often need to work with boundary points $x \in \operatorname{bd}(\Delta)$ with $\partial h(x) = \varnothing$.

Proposition B.2. *Let h be a K-strongly convex penalty function on Δ and let Δ_q be the union of the relative interiors of the faces of Δ that contain q, i.e.*

$$\Delta_q = \{x \in \Delta : \operatorname{supp}(x) \supseteq \operatorname{supp}(q)\} = \{x \in \Delta : x_\alpha > 0 \text{ whenever } q_\alpha > 0\}. \quad \text{(B.5)}$$

Then:

i. $D_h(q,x) < +\infty$ *whenever* $x \in \Delta_q$.

ii. $D_h(q,x) \geq 0$ *for all* $x \in \Delta$ *and* $D_h(q,x) = 0$ *if and only if* $q = x$*; in particular:*

$$D_h(q,x) \geq \frac{1}{2} K \|x - q\|^2 \quad \text{for all } x \in \Delta. \quad \text{(B.6)}$$

iii. $D_h(q,x_n) \to D_h(q,x)$ *whenever* $x_n \to x$ *in* Δ_q.

Proof. Let $z = q - x$. If $x \in \Delta_q$, $h(x+tz)$ is finite and smooth for all t in a neighborhood of 0 so $h'(x; q-x)$ – and hence $D_h(q,x)$ – must be finite as well.

For part (ii), positive-definiteness is a trivial consequence of strict convexity; on the other hand, *strong* convexity yields:

$$h(x+tz) \leq th(q) + (1-t)h(x) - \frac{1}{2} K t(1-t) \|z\|^2. \quad \text{(B.7)}$$

Moreover, with $h(x)$ finite, we also get $h(x+tz) \geq h(x) + th'(x;z)$, so (B.7) gives:

$$h(x) + th'(x;z) \leq th(q) + (1-t)h(x) - \frac{1}{2} K t(1-t) \|z\|^2. \quad \text{(B.8)}$$

After rearranging and dividing by t, the above becomes

$$h(q) - h(x) - h'(x;z) \geq \frac{1}{2} K (1-t) \|z\|^2, \quad \text{(B.9)}$$

so (B.6) is obtained by letting $t \to 0$.

Finally, if $x_n \to q$, part (iii) follows from Kiwiel [32, Lemma 8.2]. Otherwise, if $\lim x_n \neq q$, let $z_n = q - x_n$ and take $\varepsilon > 0$ such that $x_n + tz_n \in \Delta$ for all $t \in (-\varepsilon, \varepsilon)$ and for all sufficiently large n (recall that Δ_q is relatively open in Δ so $x_n \in \Delta_q$ for large enough n). Furthermore, let $f_n(t) = h(x_n + tz_n)$ and let $f(t) = h(x+tz)$ for $t \in (-\varepsilon, \varepsilon)$; since f_n and f are smooth, convex and $f_n \to f$ pointwise, we obtain $h'(x_n; q - x_n) = f_n'(0) \to f'(0) = h'(x; q-x)$ by Theorem 25.7 in Rockafellar [48]. □

Dually to the above, h also induces a convex coupling $P_h : \Delta \times V^* \to \mathbb{R}$ with

$$P_h(q,y) = h(q) + h^*(y) - \langle y | q \rangle \quad \text{for all } q \in \Delta,\ y \in V^*. \quad \text{(B.10)}$$

This primal-dual coupling is (strictly) convex in both arguments and positive-semidefinite by Fenchel's inequality, so we call it the *Fenchel coupling* between q and y. The next proposition establishes the duality relation between D_h and P_h:

Proposition B.3. *Let h be a K-strongly convex penalty function on Δ and let $q \in \Delta$. Then, $P_h(q,y) \geq \frac{1}{2} K \|Q(y) - q\|^2$ for all $y \in V^*$ and $P_h(q,y) \to 0$ if and only if $Q(y) \to q$; furthermore:*

$$P_h(q,y) = D_h(q,x) \quad \text{whenever } Q(y) = x \text{ and } x \in \Delta_q. \quad \text{(B.11)}$$

Remark B.2. The first part of Proposition B.3 is not implied by the second because $\operatorname{im} Q = \operatorname{dom} \partial h$ is not necessarily contained in Δ_q.

Proof. For the first part of the proposition, let $x = Q(y)$ so that $h^*(y) = \langle y|x\rangle - h(x)$. Then:

$$P_h(q, y) = h(q) - h(x) - \langle y|q - x\rangle\,. \tag{B.12}$$

With $y \in \partial h(x)$, we also get

$$h(x + t(q - x)) \geq h(x) + t\,\langle y|q - x\rangle \quad \text{for all } t \in [0, 1], \tag{B.13}$$

so, by combining (B.13) with (B.7) as in the proof of Prop. B.2, we obtain

$$h(q) - h(x) - \langle y|q - x\rangle \geq \frac{1}{2} K \left\|x - q\right\|^2, \tag{B.14}$$

and our claim follows.

As for (B.11), if $x \in \Delta_q$, the function $h(x + t(q - x))$ is finite and smooth for all t in a neighborhood of 0 (simply note that $x + t(q - x)$ lies in the relative interior of a face of Δ for small t). Thus, since $y \in \partial h(x)$ and h admits a two-sided derivative at x along $q - x$, we will also have $h'(x; q - x) = \langle y|q - x\rangle$, so (B.11) follows from (B.12). □

Intuitively, Proposition B.3 shows that the Fenchel coupling $P_h(q, y)$ between q and y measures the proximity of $Q(y)$ to q; as such, if $P_h(q, y)$ grows large, $Q(y)$ must be moving away from q. We formalize this as follows:

Proposition B.4. *If $P_h(q, y_n) \to +\infty$ for some sequence $y_n \in V^*$, the sequence $x_n = Q(y_n)$ has no limit points in Δ_q; in particular, $\liminf_{n\to\infty}\{x_{n,\alpha} : \alpha \in \operatorname{supp}(q)\} = 0$.*

Proof. By descending to a subsequence of $x_n = Q(y_n)$ if necessary, assume that $\lim x_n = x^* \in \Delta_q$. Since Δ_q is relatively open in Δ, we must eventually have $x_n \in \Delta_q$, so Proposition B.2 gives $D_h(q, x_n) \to D_h(q, x^*) < +\infty$. However, with $x_n \in \Delta_q$, Proposition B.3 also yields $D_h(q, x_n) = P_h(q, y_n) \to +\infty$, a contradiction. □

The following result may be seen as a weak partial converse to the above:

Proposition B.5. *Let $q \in \Delta^\circ$. If $P_h(q, y_n)$ is bounded, then $y_{n,\alpha} - y_{n,\beta}$ is also bounded for all $\alpha, \beta = 0, 1, \dots, d$.*

Proof. Assume ad absurdum that $P_h(q, y_n)$ is bounded but $y_{n,\alpha} - y_{n,\beta} \to +\infty$ for some $\alpha, \beta \in \mathcal{A} \equiv \{0, 1, \dots, d\}$. Then, by relabeling indices and passing to a subsequence if necessary, we may assume that *a)* $y_{n,\alpha} \geq y_{n,\kappa} \geq y_{n,\beta}$ for all $\kappa \in \mathcal{A}$; and *b)* the index set $\mathcal{A}$ can be partitioned into two nonempty sets, $\mathcal{A}^+$ and $\mathcal{A}^-$, such that $y_{n,\alpha} - y_{n,\kappa}$ is bounded for all $\kappa \in \mathcal{A}^+$ while $y_{n,\alpha} - y_{n,\kappa} \to +\infty$ for all $\kappa \in \mathcal{A}^-$ (obviously, $\alpha \in \mathcal{A}^+$ and $\beta \in \mathcal{A}^-$). Hence, letting $x_n = Q(y_n)$, we readily obtain:

$$\begin{aligned}\langle y_n|q - x_n\rangle &= \sum_{\kappa\in\mathcal{A}} y_{n,\kappa}(q_\kappa - x_{n,\kappa}) = \sum_{\kappa\in\mathcal{A}} (y_{n,\kappa} - y_{n,\alpha})(q_\kappa - x_{n,\kappa}) \\ &= \sum_{\kappa\in\mathcal{A}^+} (y_{n,\kappa} - y_{n,\alpha})(q_\kappa - x_{n,\kappa}) + \sum_{\kappa\in\mathcal{A}^-} (y_{n,\kappa} - y_{n,\alpha})(q_\kappa - x_{n,\kappa}).\end{aligned} \tag{B.15}$$

The first sum above is bounded by assumption. As for the second one, Proposition A.1 gives $x_{n,\kappa} \to 0$ as $n \to \infty$ for all $\kappa \in \mathcal{A}^-$, so $\liminf_n (q_\kappa - x_{n,\kappa}) > 0$ (recall that $q \in \Delta^\circ$). We thus obtain $\sum_{\kappa\in\mathcal{A}^-}(y_{n,\kappa} - y_{n,\alpha})(q_\kappa - x_{n,\kappa}) \to -\infty$ and, hence, $\langle y_n|q - x_n\rangle = h(q) - h(x_n) - P_h(q, y_n) \to -\infty$, a contradiction. □

Our final result shows that the Fenchel coupling evolves in a particularly simple fashion under the reinforcement learning dynamics (RL):

Lemma B.6. *Let $y(t)$ be a solution orbit of (RL). Then, for all $q_k \in \mathcal{X}_k$, we have:*

$$\frac{d}{dt} P_{h_k}(q_k, y_k(t)) = \langle v_k(x(t)) | x_k(t) - q_k \rangle \,. \tag{B.16}$$

Proof. By the differential formulation (3.1) of the dynamics (RL), we readily obtain:

$$\begin{aligned} \frac{d}{dt} P_{h_k}(q_k, y_k) &= \frac{d}{dt} h_k^*(y_k) - \langle \dot{y}_k | q_k \rangle \\ &= \langle \dot{y}_k | dh_k^*(y_k) \rangle - \langle v_k | q_k \rangle = \langle v_k | Q_k(y_k) - q_k \rangle = \langle v_k | x_k - q_k \rangle \,, \end{aligned} \tag{B.17}$$

where the first equality in the second line follows from the chain rule and the penultimate one from the fact that $Q_k = dh_k^*$ (Proposition B.1). □

References

[1] Akin, Ethan. 1980. Domination or equilibrium. *Mathematical Biosciences* **50** 239–250.

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

[3] Aubin, Jean-Pierre, Arrigo Cellina. 1984. *Differential Inclusions*. Springer, Berlin.

[4] Bayer, D. A., J. C. Lagarias. 1989. The nonlinear geometry of linear programming I. Affine and projective scaling trajectories. *Transactions of the American Mathematical Society* **314** 499–526.

[5] Beck, Amir, Marc Teboulle. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* **31**(3) 167–175.

[6] Beggs, Alan W. 2005. On the convergence of reinforcement learning. *Journal of Economic Theory* **122** 1–36.

[7] Benaïm, Michel. 1999. Dynamics of stochastic approximation algorithms. *Séminaire de probabilités de Strasbourg* **33**.

[8] Benaïm, Michel, Josef Hofbauer, Sylvain Sorin. 2005. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization* **44** 328–348.

[9] Berge, Claude. 1997. *Topological Spaces*. Dover, New York.

[10] Börgers, T., R. Sarin. 1997. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory* **77** 1–14.

[11] Bregman, Lev M. 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* **7**(3) 200–217.

[12] Bubeck, Sébastien. 2011. Introduction to online optimization. Lecture Notes.

[13] Cominetti, Roberto, Emerson Melo, Sylvain Sorin. 2010. A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior* **70** 71–83.

[14] Coucheney, Pierre, Bruno Gaujal, Panayotis Mertikopoulos. 2014. Penalty-regulated dynamics and robust learning procedures in games. *Mathematics of Operations Research* .

[15] Cross, J. G. 1973. A stochastic learning model of economic behavior. *Quarterly Journal of Economics* **87** 239–266.

[16] Erev, Ido, Alvin E. Roth. 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* **88** 848–881.

[17] Fiacco, Anthony V. 1990. Perturbed variations of penalty function methods. Example: Projective SUMT. *Annals of Operations Research* **27** 371–380.

[18] Freund, Yoav, Robert E. Schapire. 1999. Adaptive game playing using multiplicative weights. *Games and Economic Behavior* **29** 79–103.

[19] Friedman, Daniel. 1991. Evolutionary games in economics. *Econometrica* **59**(3) 637–666.

[20] Fudenberg, Drew, David K. Levine. 1998. *The Theory of Learning in Games*, *Economic learning and social evolution*, vol. 2. MIT Press, Cambridge, MA.

[21] Gilboa, Itzhak, Akihiko Matsui. 1991. Social stability and equilibrium. *Econometrica* **59**(3) 859–867.

[22] Harper, Marc. 2011. Escort evolutionary game theory. *Physica D: Nonlinear Phenomena* **240**(18) 1411–1415.

[23] Hart, Sergiu, Andreu Mas-Colell. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* **68**(5) 1127–1150.

[24] Hofbauer, Josef, William H. Sandholm. 2002. On the global convergence of stochastic fictitious play. *Econometrica* **70**(6) 2265–2294.

[25] Hofbauer, Josef, Karl Sigmund. 1990. Adaptive dynamics and evolutionary stability. *Applied Mathematics Letters* **3** 75–79.

[26] Hofbauer, Josef, Karl Sigmund. 1998. *Evolutionary Games and Population Dynamics*. Cambridge University Press.

[27] Hofbauer, Josef, Sylvain Sorin, Yannick Viossat. 2009. Time average replicator and best reply dynamics. *Mathematics of Operations Research* **34**(2) 263–269.

[28] Hopkins, Ed. 1999. Learning, matching, and aggregation. *Games and Economic Behavior* **26** 79–110.

[29] Hopkins, Ed. 1999. A note on best response dynamics. *Games and Economic Behavior* **29** 138–150.

[30] Hopkins, Ed. 2002. Two competing models of how people learn in games. *Econometrica* **70**(6) 2141–2166.

[31] Hopkins, Ed, Martin Posch. 2005. Attainability of boundary points under reinforcement learning. *Games and Economic Behavior* **53**(1) 110–125.

[32] Kiwiel, Krzysztof C. 1997. Free-steering relaxation methods for problems with strictly convex costs and linear constraints. *Mathematics of Operations Research* **22**(2) 326–349.

[33] Kwon, Joon, Panayotis Mertikopoulos. 2014. A continuous-time approach to online optimization. http://arxiv.org/abs/1401.6956.

[34] Lahkar, Ratul, William H. Sandholm. 2008. The projection dynamic and the geometry of population games. *Games and Economic Behavior* **64** 565–590.

[35] Laraki, Rida, Panayotis Mertikopoulos. 2013. Higher order game dynamics. *Journal of Economic Theory* **148**(6) 2666–2695.

[36] Laraki, Rida, Panayotis Mertikopoulos. 2013. Inertial game dynamics and applications to constrained optimization. http://arxiv.org/abs/1305.0967.

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

[38] Littlestone, Nick, Manfred K. Warmuth. 1994. The weighted majority algorithm. *Information and Computation* **108**(2) 212–261.

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

[40] Mertikopoulos, Panayotis, Aris L. Moustakas. 2010. The emergence of rational behavior in the presence of stochastic perturbations. *The Annals of Applied Probability* **20**(4) 1359–1388.

[41] Mertikopoulos, Panayotis, William H. Sandholm. 2014. Riemannian game dynamics. Forthcoming.

[42] Nachbar, John H. 1990. Evolutionary selection dynamics in games. *International Journal of Game Theory* **19** 59–89.

[43] Nagurney, A., D. Zhang. 1997. Projected dynamical systems in the formulation, stability analysis, and computation of fixed demand traffic network equilibria. *Transportation Science* **31** 147–158.

[44] Nemirovski, Arkadi Semen, David Berkovich Yudin. 1983. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, NY.

[45] Nesterov, Yurii. 2009. Primal-dual subgradient methods for convex problems. *Mathematical Programming* **120**(1) 221–259.

[46] Posch, Martin. 1997. Cycling in a stochastic learning algorithm for normal form games. *Journal of Evolutionary Economics* **7** 193–207.

[47] Robinson, Clark. 1995. *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*. CRC Press, Boca Raton, FL.

[48] Rockafellar, Ralph Tyrrell. 1970. *Convex Analysis*. Princeton University Press, Princeton, NJ.

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

[50] Samuelson, Larry, Jianbo Zhang. 1992. Evolutionary stability in asymmetric games. *Journal of Economic Theory* **57** 363–391.

[51] Sandholm, William H., Emin Dokumacı, Ratul Lahkar. 2008. The projection dynamic and the replicator dynamic. *Games and Economic Behavior* **64** 666–683.

[52] Shalev-Shwartz, Shai. 2011. Online learning and online convex optimization. *Foundations and Trends in Machine Learning* **4**(2) 107–194.

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

[54] Sutton, R. S., A. G. Barto. 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

[55] Taylor, Peter D., Leo B. Jonker. 1978. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences* **40**(1-2) 145–156.

[56] Tsallis, Constantino. 1988. Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics* **52** 479–487.

[57] Tuyls, Karl, Pieter Jan 't Hoen, Bram Vanschoenwinkel. 2006. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems* **12** 115–153.

[58] Vovk, Volodimir G. 1990. Aggregating strategies. *COLT '90: Proceedings of the 3rd Workshop on Computational Learning Theory*. 371–383.

[59] Weibull, Jörgen W. 1995. *Evolutionary Game Theory*. MIT Press, Cambridge, MA.

(Panayotis Mertikopoulos) CNRS (French National Center for Scientific Research), LIG, F-38000 Grenoble, France, and Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France

E-mail address: panayotis.mertikopoulos@imag.fr

URL: http://mescal.imag.fr/membres/panayotis.mertikopoulos

(William H. Sandholm) Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison WI 53706, USA

E-mail address: whs@ssc.wisc.edu

URL: http://www.ssc.wisc.edu/~whs/