On the robustness of learning in games with stochastically perturbed payoff observations

Mario Bravo, Panayotis Mertikopoulos

ON THE ROBUSTNESS OF LEARNING IN GAMES WITH STOCHASTICALLY PERTURBED PAYOFF OBSERVATIONS

MARIO BRAVO AND PANAYOTIS MERTIKOPOULOS

Abstract. We study a general class of game-theoretic learning dynamics in the presence of random payoff disturbances and observation noise, and we provide a unified framework that extends several rationality properties of the (stochastic) replicator dynamics and other game dynamics. In the unilateral case, we show that the stochastic dynamics under study lead to no regret, irrespective of the noise level. In the multi-player case, we find that dominated strategies become extinct (a.s.) and strict Nash equilibria remain stochastically asymptotically stable – again, independently of the perturbations' magnitude. Finally, we establish an averaging principle for 2-player games and we show that the empirical distribution of play converges to Nash equilibrium in zero-sum games under any noise level.

Contents

2010 *Mathematics Subject Classification.* Primary 60H10, 91A26; secondary 60H30, 60J70, 91A22.

Key words and phrases. Learning; dominated strategies; Nash equilibrium; regret minimization; regularized best responses; stochastic replicator dynamics; stability; time averages.

The authors are greatly indebted to Roberto Cominetti for arranging the visit of the second author to the University of Chile and for his many insightful comments and suggestions that greatly improved all aspects of this manuscript. The authors would also like to express their gratitude to Bill Sandholm and Yannick Viossat for many helpful 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" and during the second author's visit to Universidad de Chile.

The first author was partially supported by Fondecyt grant No. 3130732, the Nucleo Milenio Información y Coordinación en Redes ICM/FIC P10-024F and the Complex Engineering Systems Institute (ICM: P-05-004-F, CONICYT: FBO16).

The second author gratefully acknowledges support by the French National Research Agency under grant nos. GAGA–13– JS01–0004–01 and NETLEARN–13–INFR–004, and the French National Center for Scientific Research (CNRS) under grant no. PEPS–GATHERING–2014.

1. Introduction

The objective of learning in games is to reach a rationally acceptable state (such as a Nash equilibrium or a state where no dominated strategies are present) via a simple, dynamic process. To that end, one of the most widely studied learning processes is the exponential weight (EW) algorithm that was originally introduced by Vovk (1990) and Littlestone and Warmuth (1994) in the context of multi-armed bandit problems. In a game-theoretic setting, the algorithm simply prescribes that players score their actions based on their cumulative payoffs and then assign choice probabilities proportionally to the exponential of each action's score. As such, the EW algorithm in continuous time (Sorin, 2009) is described by the (deterministic) dynamics:

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

where $v_{k\alpha}$ denotes the payoff to the α-th action of player k, $y_{k\alpha}$ is its performance score (cumulative payoff) and $x_{k\alpha}$ is the corresponding mixed strategy weight.

A simple differentiation reveals that the evolution of the players' mixed strategies under (EW) is governed by the (multi-population) replicator dynamics of Taylor and Jonker (1978):

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

The replicator equation is one of the most widely studied dynamical systems for population evolution under selection and its rationality properties have attracted significant interest in the literature. Akin (1980), Nachbar (1990) and Samuelson and Zhang (1992) showed that dominated strategies become extinct under (EW)/(RD) while it is well known that *a*) Lyapunov stable states are Nash; *b*) strict Nash equilibria are asymptotically stable; and *c*) time averages of replicator orbits converge to equilibrium in 2-player games provided that no strategy share becomes arbitrarily small (Hofbauer and Sigmund, 1998). More recently, Sorin (2009) showed that the single-player version of (EW)/(RD) is also *universally consistent*, i.e. players have no regret for following (EW) instead of any other fixed strategy (Fudenberg and Levine, 1995).

In this paper, we consider a broad class of reinforcement learning processes obtained by replacing the exponential map in (EW) by an arbitrary *perturbed* (or *regularized*) best response map that reinforces strategies with higher scores (Fudenberg and Levine, 1998; Hofbauer and Sandholm, 2002; Leslie and Collins, 2005; Shalev-Shwartz, 2011). In a deterministic context, the long-term behavior of the resulting dynamics was studied recently by Mertikopoulos and Sandholm (2014) who extended the rationality properties of the replicator dynamics to this much broader setting. However, a key assumption underlying all these deterministic considerations is that the players' payoffs are impervious to exogenous fluctuations and that players possess perfect observations thereof. These requirements are rarely met in practical applications of game theory (e.g. in wireless communications and traffic engineering), so our goal is to investigate the robustness of this general learning scheme in the presence of stochastic payoff perturbations and observation noise.

In a biological setting (where payoffs measure a species' reproductive fitness), Fudenberg and Harris (1992) introduced a stochastic variant of the replicator dynamics where evolution is perturbed by "aggregate shocks" reflecting the impact of

weather-like effects and other fluctuations in the species' habitat – see also Khasminskii and Potsepun (2006) for a Stratonovich-based model and Vlasic (2012) for the case of random semimartingale jumps incurred by catastrophic, earthquake-like events. In this framework, Cabrales (2000), Imhof (2005) and Hofbauer and Imhof (2009) showed that dominated strategies are still eliminated if the variability of the shocks across different genotypes (strategies) is not too high; likewise, Imhof (2005) and Hofbauer and Imhof (2009) showed that strict Nash equilibria are stochastically asymptotically stable under similar "mild noise" requirements.

On the other hand, Mertikopoulos and Moustakas (2010) showed that the replicator dynamics with aggregate shocks do not coincide with the stochastic replicator dynamics induced by (EW) in the presence of random disturbances and measurement noise. Surprisingly, this learning variant of the stochastic replicator dynamics retains the rationality properties of its deterministic counterpart without any caveats on the noise: dominated strategies become extinct and strict Nash equilibria remain stochastically asymptotically stable irrespective of the perturbations' magnitude.

In this paper, we show that the robustness of (EW) in the presence of noise is a special case of a much more general learning principle that also extends the recent deterministic results of Mertikopoulos and Sandholm (2014) to a broader, stochastic setting. In Section 2, we present our stochastically perturbed reinforcement learning model and we derive the system of coupled stochastic differential equations (SDEs) that governs the evolution of the players' mixed strategies. In Section 3, we show that these dynamics lead to no regret in a unilateral setting, whatever the noise level. In Section 4, we investigate dominated strategies: we show that dominated strategies become extinct (a.s.) and we derive an explicit bound for the probability (and corresponding first passage time) that a pure dominated strategy is above a given level. Section 5 focuses on the dynamics' long-term stability and convergence properties: we show that *a*) stochastically (Lyapunov) stable states and states that attract trajectories of play with positive probability are Nash; and *b*) strict Nash equilibria are stochastically asymptotically stable, irrespective of the fluctuations' magnitude. Finally, in Section 6, we provide an averaging principle for 2-player games in the spirit of Hofbauer, Sorin, and Viossat (2009). This principle allows us to show that empirical distributions of play converge to Nash equilibrium in zero-sum games (again, no matter the noise level).

1.1. **Notation and preliminaries.** If V is a vector space, we will write V^* for its dual and $\langle y|x\rangle$ for the pairing between $x \in V$ and $y \in V^*$. The real space spanned by the finite set $\mathcal{S} = \{s_\alpha\}_{\alpha=1}^{d+1}$ will be denoted by $\mathbb{R}^{\mathcal{S}}$ and its canonical basis by $\{e_s\}_{s\in\mathcal{S}}$. In a slight abuse of notation, we will also use α to refer interchangeably to either s_α or e_α and we will 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 $\Delta = \{x \in \mathbb{R}^{\mathcal{S}} : \sum_\alpha x_\alpha = 1 \text{ and } x_\alpha \geq 0\}$ of $\mathbb{R}^{\mathcal{S}}$ and the relative interior of Δ will be denoted by Δ°. For simplicity, if $\{\mathcal{S}_k\}_{k\in\mathcal{N}}$ is a finite family of finite sets, we use the shorthand $(\alpha_k;\alpha_{-k})$ for the tuple $(\dots,\alpha_{k-1},\alpha_k,\alpha_{k+1},\dots)$ and we write $\sum_\alpha^k$ for $\sum_{\alpha\in\mathcal{S}_k}$. Finally, we suppress the dependence of the law of a process $X(t)$ on its initial condition $X(0) = x$ and we write $\mathbb{P}$ instead of $\mathbb{P}_x$.

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 set of all joint action profiles $(\alpha_1, \dots, \alpha_N)$. 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* $x = (x_1, \dots, x_N)$ is called the game's *strategy space*. Unless mentioned otherwise, we will write $V_k \equiv \mathbb{R}^{\mathcal{A}_k}$ and $V \equiv \prod_k V_k \cong \mathbb{R}^{\coprod_k \mathcal{A}_k}$ for the ambient spaces of $\mathcal{X}_k$ and $\mathcal{X}$ respectively.

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

$$u_k(x) = \sum\nolimits_{\alpha_1}^{1} \cdots \sum\nolimits_{\alpha_N}^{N} u_k(\alpha_1, \dots, \alpha_N)\, x_{1,\alpha_1} \cdots x_{N,\alpha_N}, \tag{1.1}$$

where $u_k(\alpha_1, \dots, \alpha_N)$ denotes the payoff of player k in the profile $(\alpha_1, \dots, \alpha_N) \in \mathcal{A}$. Accordingly, the payoff corresponding to $\alpha \in \mathcal{A}_k$ in the mixed profile $x \in \mathcal{X}$ is

$$v_{k\alpha}(x) = \sum\nolimits_{\alpha_1}^{1} \cdots \sum\nolimits_{\alpha_N}^{N} u_k(\alpha_1, \dots, \alpha_N)\, x_{1,\alpha_1} \cdots \delta_{\alpha_k,\alpha} \cdots x_{N,\alpha_N}, \tag{1.2}$$

and we have

$$u_k(x) = \sum\nolimits_{\alpha}^{k} x_{k\alpha} v_{k\alpha}(x) = \langle v_k(x) | x_k \rangle \tag{1.3}$$

where $v_k(x) = (v_{k\alpha}(x))_{\alpha \in \mathcal{A}_k}$ denotes the *payoff vector* of player k at $x \in \mathcal{X}$. In the above, v_k is treated as a dual vector in V_k^* that is paired to the mixed strategy $x_k \in \mathcal{X}_k$; on account of this duality, mixed strategies will be regarded throughout this paper as *primal* variables and payoff vectors as *duals*.

2. Reinforcement Learning with Noisy Observations

In this section, we introduce the class of stochastic game dynamics under study and we discuss some of their main properties.

2.1. The deterministic case.

Consider a general reinforcement learning process where, at each $t \geq 0$, every player of a finite game $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ employs an "approximate" best response to the vector of his cumulative payoffs up to time t. Specifically, this corresponds to the deterministic continuous-time process

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

where:

1. the *score vector* $y_k(t) \in V_k^*$ ranks strategies $\alpha \in \mathcal{A}_k$ based on their cumulative payoffs up to time t.
2. $Q_k \colon V_k^* \to \mathcal{X}_k$ is a *regularized best response* (or *choice*) *map* which reinforces strategies with higher scores (see below for a rigorous definition).
3. $\eta_k(t) > 0$ is a *learning parameter* which can be tuned freely by each player.

A natural choice for the "scores-to-strategies" map Q_k would be the best response correspondence $y_k \mapsto \arg\max_{x_k \in \mathcal{X}_k} \langle y_k | x_k \rangle$, i.e. to greedily assign all weight to the strategy (or strategies) with the highest score. However, since the arg max operator is multi-valued, we will focus on choice maps of the general form

$$Q_k(y_k) = \arg\max_{x_k \in \mathcal{X}_k} \{\langle y_k | x_k \rangle - h_k(x_k)\}, \quad y_k \in V_k^*, \tag{2.1}$$

where the *penalty function* $h_k \colon \mathcal{X}_k \to \mathbb{R}$ satisfies the following properties:

a) h_k is continuous on $\mathcal{X}_k$.
b) h_k is smooth on the relative interior of every face of $\mathcal{X}_k$.

c) h_k is *strongly convex* on $\mathcal{X}_k$: there exists some $K > 0$ such that

$$h_k(tx_k + (1-t)x'_k) \leq th_k(x_k) + (1-t)h_k(x'_k) - \tfrac{1}{2}Kt(1-t)\|x'_k - x_k\|^2, \quad (2.2)$$

for all $x_k, x'_k \in \mathcal{X}_k$ and for all $t \in [0,1]$.

This "softening" of the arg max operator has a long history in game theory, learning and optimization, and the resulting map Q_k is commonly referred to as a *softmax* or *perturbed best response map*; for an in-depth discussion, we refer the reader to van Damme (1987), Fudenberg and Levine (1998), Hofbauer and Sandholm (2002), Leslie and Collins (2005), Nesterov (2009), Shalev-Shwartz (2011) and Mertikopoulos and Sandholm (2014). For our immediate purposes, the key observation is that the (strictly) concave problem (2.1) admits a unique solution for every $y_k \in V_k^*$, so $Q_k(y_k)$ can be seen as a single-valued approximation of the standard best response correspondence $y_k \mapsto \arg\max_{x_k \in \mathcal{X}_k} \langle y_k | x_k \rangle$.

Finally, regarding the learning rate parameter η_k, its role in (RL) is to temper the growth of the cumulative payoff vector $y_k(t)$ so as to allow the player to better explore his strategies (instead of prematurely reinforcing one or another). Accordingly, with $y(t)$ growing as $\mathcal{O}(t)$, we will assume throughout this paper that:

Assumption 1. $\eta_k(t)$ is C^1-smooth, nonincreasing and $\lim_{t\to\infty} t\eta_k(t) = +\infty$.

Previous work and examples. The cumulative reinforcement principle behind the dynamics (RL) can be traced back to the seminal work of Vovk (1990), Littlestone and Warmuth (1994) and Sutton and Barto (1998) on multi-armed bandit problems in a discrete-time setting (mostly). Game-theoretic variants of (RL) have also been studied (in both discrete and continuous time) by Fudenberg and Levine (1995), Freund and Schapire (1999), Hopkins (2002), Leslie and Collins (2005), Tuyls, 't Hoen, and Vanschoenwinkel (2006), Cominetti, Melo, and Sorin (2010), Coucheney, Gaujal, and Mertikopoulos (2014) and many others; for a systematic account, see Fudenberg and Levine (1998), Shalev-Shwartz (2011), Mertikopoulos and Sandholm (2014) and references therein.

Below we provide two characteristic examples of the reinforcement learning scheme (RL) based on logit and projected best responses:

Example 2.1. The prototype penalty function on the unit d-dimensional simplex Δ is the Gibbs (negative) entropy $h(x) = \sum_\alpha x_\alpha \log x_\alpha$. By a standard calculation, the associated regularized best response is given by the so-called *logit map*:

$$G_\alpha(y) = \frac{\exp(y_\alpha)}{\sum_\beta \exp(y_\beta)}. \quad (2.3)$$

For constant $\eta = 1$, (2.3) leads to the continuous-time exponential weight algorithm (EW) that was presented in Section 1 (Littlestone and Warmuth, 1994; Sorin, 2009; Vovk, 1990). An easy differentiation then yields

$$\dot{x}_{k\alpha} = \frac{e^{y_{k\alpha}}\dot{y}_{k\alpha}}{\sum_\beta^k e^{y_{k\beta}}} - \frac{e^{y_{k\alpha}}\sum_\beta^k e^{y_{k\beta}}\dot{y}_{k\beta}}{\left(\sum_\beta^k e^{y_{k\beta}}\right)^2} = x_{k\alpha}\left[v_{k\alpha}(x) - \sum\nolimits_\beta^k x_{k\beta}\, v_{k\beta}(x)\right], \quad (2.4)$$

which is simply the (multi-population) replicator equation of Taylor and Jonker (1978) for population evolution under natural selection. For a more thorough treatment of the links between (EW) and (RD), see Rustichini (1999), Hofbauer et al. (2009), Mertikopoulos and Moustakas (2009, 2010) and references therein.

Example 2.2. As another example, consider the quadratic penalty function $h(x) = \frac{1}{2}\sum_\alpha x_\alpha^2$. This penalty function leads to the *projected best response map*

$$\Pi(y) = \arg\min_{x\in\Delta}\left\{\langle y|x\rangle - \tfrac{1}{2}\left\|x\right\|^2\right\} = \arg\min_{x\in\Delta}\left\|y-x\right\|^2, \tag{2.5}$$

and, as was shown by Mertikopoulos and Sandholm (2014), the orbits $x(t) = \Pi(y(t))$ of (RL) with projected best responses satisfy the *projection dynamics*

$$\dot{x}_{k\alpha} = \begin{cases} v_{k\alpha}(x) - |\operatorname{supp}(x_k)|^{-1}\sum_{\beta\in\operatorname{supp}(x_k)} v_{k\beta}(x) & \text{if } x_{k\alpha} > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{PD}$$

over an open dense set of times (in particular, except when the support of $x(t)$ changes). The dynamics (PD) were introduced in game theory by Friedman (1991) as a geometric model of the evolution of play in population games; for a closely related model (but with important differences), see Nagurney and Zhang (1997), Lahkar and Sandholm (2008) and Sandholm, Dokumacı, and Lahkar (2008).

2.2. **Learning in the presence of noise.** A key assumption underlying the reinforcement learning scheme (RL) is that payoffs are impervious to any sort of exogenous random noise and that players have access to perfect payoff observations with which to update their cumulative payoff vectors y_k. However, this assumption is rarely met in practical applications of game-theoretic learning: for instance, in telecommunication networks and traffic engineering, signal strength and latency measurements are constantly subject to stochastic fluctuations which introduce noise to the input of any learning algorithm (Kang, Kelly, Lee, and Williams, 2009; Kelly, Maulloo, and Tan, 1998; Li, Lee, and Guo, 2004). Thus, in the rest of the paper, we will focus on the *stochastically perturbed reinforcement learning* process:

$$\begin{aligned} dY_{k\alpha} &= v_{k\alpha}(X)\,dt + \sigma_{k\alpha}(X)\,dW_{k\alpha}, \\ X_k &= Q_k(\eta_k Y_k), \end{aligned} \tag{SRL}$$

where $W_{k\alpha}$ is a family of independent Wiener processes and the diffusion coefficients $\sigma_{k\alpha}\colon \mathcal{X}\to\mathbb{R}$ (assumed Lipschitz) measure the strength of the players' payoff observation noise.

By Proposition A.1, the regularized best response maps Q_k are Lipschitz, so (SRL) admits a unique (strong) solution $Y(t)$ for every initial condition $Y(0)\in V^*$; standard arguments can then be used to show that these solutions exist for all time (a.s.). With this in mind, our first task will be to derive the stochastic dynamics that govern the evolution of the orbits $X(t) = Q(\eta(t)Y(t))$ of (SRL) in $\mathcal{X}$. For simplicity, following Alvarez, Bolte, and Brahic (2004), we only present here the special case where each player's penalty function is of the decomposable form

$$h_k(x_k) = \sum\nolimits_\alpha^k \theta_k(x_{k\alpha}) \tag{2.6}$$

for some strongly convex *kernel function* $\theta_k\in C^0[0,1]\cap C^3(0,1]$. We have:

Proposition 2.1. *Let $X(t)$ be an orbit of (SRL) in $\mathcal{X}$ and let I be an open interval over which the support of $X(t)$ remains constant. Then, the evolution of $X(t)$ over*

I is governed by the stochastic differential equation

$$dX_{k\alpha} = \frac{\eta_k}{\theta''_{k\alpha}}\left[v_{k\alpha} - \Theta''_k \sum\nolimits_\beta v_{k\beta}/\theta''_{k\beta}\right] dt \tag{2.7a}$$

$$+ \frac{\eta_k}{\theta''_{k\alpha}}\left[\sigma_{k\alpha}\, dW_{k\alpha} - \Theta''_k \sum\nolimits_\beta \sigma_{k\beta}/\theta''_{k\beta}\, dW_{k\beta}\right] \tag{2.7b}$$

$$+ \frac{\dot\eta_k}{\eta_k}\frac{1}{\theta''_{k\alpha}}\left[\theta'_{k\alpha} - \Theta''_k \sum\nolimits_\beta \theta'_{k\beta}/\theta''_{k\beta}\right] dt \tag{2.7c}$$

$$- \frac{1}{2}\frac{1}{\theta''_{k\alpha}}\left[\theta'''_{k\alpha} U^2_{k\alpha} - \Theta''_k \sum\nolimits_\beta \theta'''_{k\beta}/\theta''_{k\beta}\, U^2_{k\beta}\right] dt, \tag{2.7d}$$

where all summations are taken over $\beta \in \operatorname{supp}(X_k)$ and:

a) $\theta'_{k\alpha} = \theta'_k(X_{k\alpha})$, $\theta''_{k\alpha} = \theta''_k(X_{k\alpha})$, *and* $\theta'''_{k\alpha} = \theta'''_k(X_{k\alpha})$.

b) $\Theta''_k = \left(\sum_\beta 1/\theta''_{k\beta}\right)^{-1}$.

c) $U^2_{k\alpha} = \left(\dfrac{\eta_k}{\theta''_{k\alpha}}\right)^2 \left[\sigma^2_{k\alpha}\left(1 - \Theta''_k/\theta''_{k\alpha}\right)^2 + \sum_{\beta\neq\alpha}\left(\Theta''_k/\theta''_{k\beta}\right)^2 \sigma^2_{k\beta}\right]$.

In particular, if $\lim_{z\to 0^+}\theta'_k(z) = -\infty$ for all $k \in \mathcal{N}$, $X(t)$ is an ordinary (strong) solution of (2.7); otherwise, $X(t)$ satisfies (2.7) on an open dense subset of $[0,\infty)$.

Remark 2.1. Even though the dynamics (2.7) appear quite complicated, each of the constituent terms (2.7a)–(2.7d) has a relatively simple interpretation:

a) The term (2.7a) drives the process in the baseline case $\sigma = 0$, $\eta =$ constant; as such, (2.7a) recovers the deterministic reinforcement learning dynamics studied by Mertikopoulos and Sandholm (2014).
b) The diffusion term (2.7b) reflects the direct impact of the noise on (SRL).
c) The term (2.7c) is due to the variability of the players' learning rate η so its impact on (2.7) vanishes if $\eta(t) \to 0$ sufficiently fast.
d) Finally, the term (2.7d) is the Itô correction induced on dX_k through (SRL).

Example 2.3. As we saw in Example 2.1, the replicator dynamics (RD) correspond to the entropic kernel $\theta(x) = x\log x$. In this case, (2.7) leads to the following stochastic variant of the replicator dynamics:

$$\begin{aligned} dX_{k\alpha} &= \eta_k X_{k\alpha}\left[v_{k\alpha} - \sum\nolimits_\beta^k X_{k\beta}\, v_{k\beta}\right] dt \\ &+ \eta_k X_{k\alpha}\left[\sigma_{k\alpha}\, dW_{k\alpha} - \sum\nolimits_\beta^k \sigma_{k\beta} X_{k\beta}\, dW_{k\beta}\right] \\ &+ \frac{\dot\eta_k}{\eta_k} X_{k\alpha}\left[\log X_{k\alpha} - \sum\nolimits_\beta^k X_{k\beta}\log X_{k\beta}\right] dt \\ &+ \frac{1}{2} X_{k\alpha}\left[\sigma^2_{k\alpha}(1 - 2X_{k\alpha}) - \sum\nolimits_\beta^k \sigma^2_{k\beta} X_{k\beta}\,(1 - 2X_{k\beta})\right] dt. \end{aligned} \tag{SRD}$$

When η is constant, (SRD) is simply the stochastic replicator dynamics of exponential learning studied by Mertikopoulos and Moustakas (2010). On the other hand, (SRD) should be contrasted to the evolutionary *replicator dynamics with aggregate*

shocks of Fudenberg and Harris (1992):

$$\begin{aligned} dX_{k\alpha} &= X_{k\alpha}\left[v_{k\alpha} - \sum\nolimits_{\beta}^{k} X_{k\beta} v_{k\beta}\right] dt \\ &\quad + X_{k\alpha}\left[\sigma_{k\alpha} dW_{k\alpha} - \sum\nolimits_{\beta}^{k} \sigma_{k\beta} X_{k\beta}\, dW_{k\beta}\right] \\ &\quad - X_{k\alpha}\left[\sigma_{k\alpha}^2 X_{k\alpha} - \sum\nolimits_{\beta}^{k} \sigma_{k\beta}^2 X_{k\beta}^2\right] dt, \end{aligned} \tag{ASRD}$$

where $X_{k\alpha}$ denotes the population share of the α-th genotype of species k in a multi-species environment, $v_{k\alpha}$ represents its reproductive fitness, and the noise coefficients $\sigma_{k\alpha}$ measure the impact of random weather-like effects on population evolution – for a comprehensive account, see Cabrales (2000), Imhof (2005) and Hofbauer and Imhof (2009); see also Khasminskii and Potsepun (2006) for a Stratonovich-based model and Vlasic (2012) for a model that accounts for random jump discontinuities induced by semimartingale shocks. Besides the absence of the learning rate η, the fundamental difference between (SRD) and (ASRD) is in their Itô correction: this term leads to a drastically different long-term behavior and highlights an important contrast between learning and evolution in the presence of noise.

Example 2.4. In the case of the projected reinforcement learning scheme (PD), substituting $\theta(x) = x^2/2$ in (2.7) yields the *stochastic projection dynamics*:

$$\begin{aligned} dX_{k\alpha} &= \left[v_{k\alpha} - |\operatorname{supp}(X_k)|^{-1} \sum\nolimits_{\beta\in\operatorname{supp}(X_k)} v_{k\beta}\right] dt \\ &\quad + \left[\sigma_{k\alpha}\, dW_{k\alpha} - |\operatorname{supp}(X_k)|^{-1} \sum\nolimits_{\beta\in\operatorname{supp}(X_k)} \sigma_{k\beta}\, dW_{k\beta}\right] \\ &\quad + \frac{\dot\eta_k}{\eta_k}\left[X_{k\alpha} - |\operatorname{supp}(X_k)|^{-1}\right] dt. \end{aligned} \tag{SPD}$$

There are two important qualitative differences between (SRD) and (SPD): first, (SRD) holds for all $t \geq 0$ whereas (SPD) describes the evolution of the solution orbits of (SRL) only on intervals over which the support of X remains constant. Second, the projection mapping Π of (2.5) is piecewise linear, so there is no Itô correction in (SPD).

Proof of Proposition 2.1. For simplicity, we suppress the player index k; also, all summation indices are assumed to run over the (constant) support $\mathcal{A}'$ of $X(t)$.

By the Karush–Kuhn–Tucker (KKT) conditions for the convex problem (2.1), we readily obtain $Y_\alpha - \eta^{-1}\theta'(X_\alpha) = \zeta$ where ζ is the Lagrange multiplier corresponding to the probability constraint $\sum_{\alpha\in\operatorname{supp}(X_k)} X_{k\alpha} = 1$. Itô's formula then gives

$$d\zeta = dY_\alpha + \frac{\dot\eta}{\eta^2}\theta'_\alpha\, dt - \eta^{-1}\theta''_\alpha\, dX_\alpha - \frac{1}{2}\eta^{-1}\theta'''_\alpha (dX_\alpha)^2 \tag{2.8}$$

and hence:

$$dX_\alpha = \frac{\eta}{\theta''_\alpha}\left[dY_\alpha + \frac{\dot\eta}{\eta^2}\theta'_\alpha\, dt - \frac{1}{2\eta}\theta'''_\alpha (dX_\alpha)^2 - d\zeta\right]. \tag{2.9}$$

Now, write dX_α in the general form:

$$dX_\alpha = b_\alpha\, dt + \sum\nolimits_{\beta} c_{\alpha\beta}\, dW_\beta, \tag{2.10}$$

where the drift and diffusion coefficients b_α and $c_{\alpha\beta}$ are to be determined. To do so, recall that the Wiener processes W_α are independent (in the sense that $dW_\alpha \cdot dW_\beta = \delta_{\alpha\beta}$), so we get $(dX_\alpha)^2 = \sum_\beta c_{\alpha\beta}^2\, dt$; therefore, summing (2.9) over $\alpha \in \mathcal{A}'$ and solving for $d\zeta$ yields:

$$\frac{1}{\Theta''}\, d\zeta = \sum\nolimits_\alpha \frac{1}{\theta''_\alpha}\left[dY_\alpha + \frac{\dot\eta}{\eta^2}\theta'_\alpha\, dt - \frac{1}{2\eta}\theta'''_\alpha \sum\nolimits_\beta c_{\alpha\beta}^2\, dt\right]. \tag{2.11}$$

Substituting this last expression in (2.9) leads to (2.7) with $U_\alpha^2 = \sum_\beta c_{\alpha\beta}^2$, so we are left to show that U_α^2 has the prescribed form. To that end, by comparing the diffusion terms of (2.7) and (2.10), we get

$$c_{\alpha\beta} = \frac{\eta}{\theta''_\alpha}\left[\sigma_\alpha \delta_{\alpha\beta} - \frac{\Theta''}{\theta''_\beta}\sigma_\beta\right], \tag{2.12}$$

and hence:

$$\begin{aligned} U_\alpha^2 = \sum\nolimits_\beta c_{\alpha\beta}^2 &= \left(\frac{\eta}{\theta''_\alpha}\right)^2 \sum\nolimits_\beta \left(\sigma_\alpha\delta_{\alpha\beta} - \Theta''\sigma_\beta/\theta''_\beta\right)^2 \\ &= \left(\frac{\eta}{\theta''_\alpha}\right)^2 \left[\sigma_\alpha^2\left(1 - \Theta''/\theta''_\alpha\right)^2 + \sum\nolimits_{\beta\neq\alpha}\left(\Theta''/\theta''_\beta\right)^2\sigma_\beta^2\right], \end{aligned} \tag{2.13}$$

which concludes our derivation of (2.7).

To show that the support of $X(t)$ is piecewise constant on a dense open subset of $[0,\infty)$, fix some $\alpha \in \mathcal{A}$ and let $A = \{t : X_\alpha(t) > 0\}$ so that $A^c = X_\alpha^{-1}(0)$. Then, A is open because $X(t)$ is continuous (a.s.), so it suffices to show that $A \cup \operatorname{int}(A^c)$ is dense in $[0,\infty)$; however, this is trivially true because of the identity $\operatorname{cl}(A) \cup \operatorname{int}(A^c) = [0,\infty)$. Finally, if $\lim_{x\to 0^+}\theta'(x) = -\infty$, standard convex analysis arguments show that the (necessarily unique) solution of (2.1) lies in the relative interior of $\mathcal{X}_k$ (Rockafellar, 1970, Chap. 26), so we conclude that $X(t) \in \mathcal{X}^\circ$ for all $t \geq 0$ by the well-posedness of (SRL). □

3. Consistency and Regret Minimization

We begin our rationality analysis with the unilateral case where there is a single player whose payoffs are determined by the state of his environment – which, in turn, may evolve *arbitrarily* over time (including adversarially if there are other players involved).

Specifically, following Sorin (2009), consider a decision process where, at each $t \geq 0$, the player chooses an action from a finite set $\mathcal{A}$ according to some mixed strategy $x(t) \in \mathcal{X} \equiv \Delta(\mathcal{A})$ and obtains a reward based on the (a priori unknown) payoff vector $v(t) = (v_\alpha(t))_{\alpha\in\mathcal{A}}$ of stage t. In this context, the performance of the strategy $x(t)$ is measured by comparing the player's (expected) cumulative payoff to the payoff that he could have obtained if the state of nature were known in advance and the player had best-responded to it. More precisely, given a (locally integrable) stream of payoffs $v(t)$, the player's *cumulative regret* at time t is defined as

$$\operatorname{Reg}(t) = \max_{\alpha\in\mathcal{A}} \int_0^t v_\alpha(s)\, ds - \int_0^t \langle v(s)|x(s)\rangle\, ds, \tag{3.1}$$

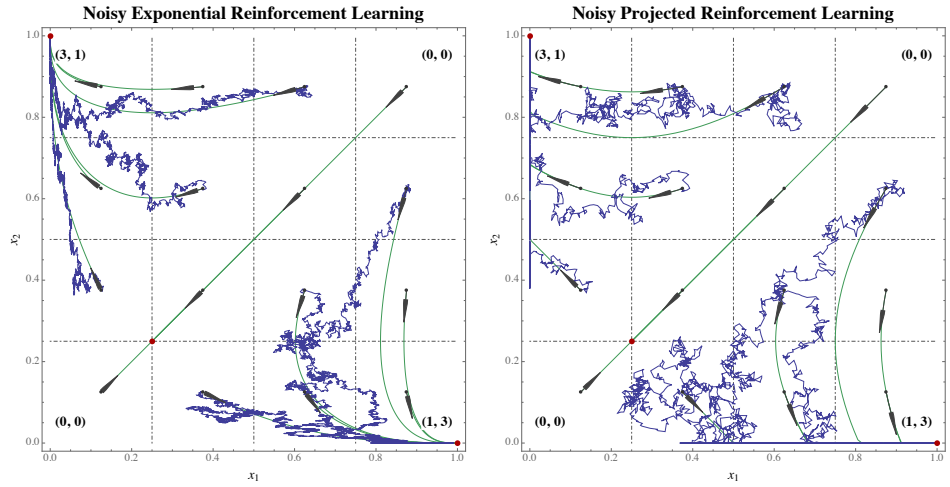


FIGURE 1. Evolution of play under (SRL) with logit (left) and projected best responses (right) for a 2×2 congestion game (Nash equilibria are depicted in red; for the game's payoffs, see the vertex labels). For comparison purposes, we took $\sigma_{k\alpha} = 1$ for all $\alpha \in \mathcal{A}_k$, $k = 1, 2$, and we used the same Wiener process realization in both cases: the orbits of projected reinforcement learning attain the boundary of $\mathcal{X}$ and converge to Nash equilibrium much faster than in the case of exponential learning.

and we say that a strategy $x(t)$ is *consistent* if it leads to *no* (*average*) *regret*, i.e.

$$\limsup_{t\to\infty} \frac{1}{t} \operatorname{Reg}(t) \leq 0, \tag{3.2}$$

or, equivalently:

$$\operatorname{Reg}(t) = o(t) \quad \text{as } t \to \infty. \tag{3.3}$$

Remark 3.1. The notion of consistency presented above is commonly referred to as *external* or *universal* consistency and was originally introduced in a discrete-time context: the agent receives a payoff vector $v_n \in \mathbb{R}^{\mathcal{A}}$ at each stage $n \in \mathbb{N}$ and, observing the past realizations $v_1, \dots, v_{n-1}$, he chooses an action $\alpha_n \in \mathcal{A}$ with law $x_n \in \Delta(\mathcal{A})$. The induced process x_n is then said to be externally consistent if, on average, it earns more than any action (or expert suggestion) $\alpha \in \mathcal{A}$. For an overview of the rich literature surrounding the topic, see Fudenberg and Levine (1998), Hart and Mas-Colell (2000), Cesa-Bianchi and Lugosi (2006), Sorin (2009), Shalev-Shwartz (2011) and references therein.

The main question that we seek to address here is whether the perturbed reinforcement learning process (SRL) is consistent. Formally, with notation as before, we will focus on the *unilateral* process:

$$\begin{aligned} dY_\alpha(t) &= v_\alpha(t)\,dt + \sigma_\alpha(t)\,dW_\alpha(t), \\ X(t) &= Q(\eta(t)Y(t)), \end{aligned} \tag{SRL-U}$$

where $v(t)$ is a locally integrable stream of payoffs, $W(t)$ is a Wiener process in $\mathbb{R}^{\mathcal{A}}$ and the noise coefficients σ_α (assumed continuous and bounded) represent the error in the player's payoff observations. In the deterministic case ($\sigma = 0$), Sorin (2009) proved that the unilateral variant of (EW) is consistent, a result which was recently extended by Kwon and Mertikopoulos (2014) to (RL) run with arbitrary regularized best responses. Below, we show that (SRL-U) is consistent even when the player's payoff observations are subject to arbitrarily high measurement errors:

Theorem 3.1. *Assume that* (SRL-U) *is run with learning parameter* $\eta(t)$ *satisfying* $\lim_{t\to\infty} \eta(t) = 0$ *(in addition to Assumption 1). Then,* (SRL-U) *is consistent.*

The basic idea of our proof will be to study the generating process $Y(t)$ of (SRL-U) as it evolves in the dual space V^* and to examine how far the resulting induced trajectory $X(t) = Q(\eta(t)Y(t))$ can stray from a benchmark strategy $p \in \mathcal{X}$. To carry out this "primal-dual" comparison, we will consider the so-called *Fenchel coupling* $F\colon \mathcal{X} \times V^* \to \mathbb{R}$ defined as:

$$F(x, y) = h(x) + h^*(y) - \langle y|x\rangle\,, \tag{3.4}$$

where

$$h^*(y) = \max_{x\in\mathcal{X}}\{\langle y|x\rangle - h(x)\} \tag{3.5}$$

denotes the *convex conjugate* of h (Rockafellar, 1970). The terminology "Fenchel coupling" is due to Mertikopoulos and Sandholm (2014) and reflects the fact that (3.4) collects all the terms of Fenchel's inequality. As a result, $F(x, y)$ is non-negative and (strictly) convex in both arguments, so it provides a "congruity" measure between x and y which can be seen as a primal-dual analogue of the well-known (primal-primal) Bregman divergence of h (Bregman, 1967).

Our goal will be to express the player's regret in terms of this Fenchel coupling and show that the latter grows sublinearly in t. To that end, we will also require the following easy consequence of the law of the iterated logarithm:

Lemma 3.2. *Let* $W(t) = (W_1(t), \dots, W_{d+1}(t))$, $t \geq 0$, *be a Wiener processes in* $\mathbb{R}^{d+1}$ *and let* $Z(t)$ *be a bounded, continuous process in* $\mathbb{R}^{d+1}$. *Then:*

$$f(t) + \int_0^t Z(s)\cdot dW(s) \sim f(t) \quad \text{as } t\to\infty \text{ (a.s.)}, \tag{3.6}$$

whenever $\lim_{t\to\infty} (t\log\log t)^{-1/2} f(t) = +\infty$.

Proof. Let $\xi(t) = \int_0^t Z(s)\cdot dW(s) = \sum_{\alpha=1}^{d+1}\int_0^t Z_\alpha(s)\,dW_\alpha(s)$. Then, the quadratic variation $\rho = [\xi, \xi]$ of ξ satisfies:

$$d[\xi,\xi] = d\xi\cdot d\xi = \sum\nolimits_{\alpha=1}^{d+1} Z_\alpha Z_\beta \delta_{\alpha\beta}\,dt \leq M\,dt, \tag{3.7}$$

where $M = \sup_{t\geq 0}\|Z(t)\|^2 < +\infty$ (recall that $Z(t)$ is bounded by assumption). On the other hand, by the time-change theorem for martingales (Øksendal, 2007, Cor. 8.5.4), there exists a Wiener process $\widetilde{W}(t)$ such that $\xi(t) = \widetilde{W}(\rho(t))$, and hence:

$$\frac{f(t) + \xi(t)}{f(t)} = 1 + \frac{\widetilde{W}(\rho(t))}{f(t)}. \tag{3.8}$$

Obviously, if $\lim_{t\to\infty}\rho(t) \equiv \rho(\infty) < +\infty$, $\widetilde{W}(\rho(\infty))$ will be normally distributed so $\widetilde{W}(\rho(t))/f(t) \to 0$ and there is nothing to show. Otherwise, if $\lim_{t\to\infty}\rho(t) =$

$+\infty$, the quadratic variation bound (3.7) and the law of the iterated logarithm yield:

$$\frac{\left|\widetilde{W}(\rho(t))\right|}{f(t)} \leq \frac{\left|\widetilde{W}(\rho(t))\right|}{\sqrt{2\rho(t)\log\log\rho(t)}} \times \frac{\sqrt{2Mt\log\log Mt}}{f(t)} \to 0 \quad \text{as } t \to \infty, \tag{3.9}$$

and our claim follows. □

Proof of Theorem 3.1. Our proof hinges on the rate-adjusted Fenchel coupling

$$H_p \equiv \frac{1}{\eta} F(p, \eta Y) = \frac{1}{\eta} \cdot [h(p) + h^*(\eta Y) - \langle \eta Y | p \rangle] \tag{3.10}$$

between a benchmark strategy $p \in \mathcal{X}$ and the generating process ηY. To begin with, the Itô formula of Lemma A.3 gives:

$$\begin{aligned} dH_p &= -\frac{\dot\eta}{\eta} H_p \, dt + \frac{1}{\eta} \langle d(\eta Y) | X - p \rangle + \frac{1}{2\eta} \sum\nolimits_\beta \frac{\partial^2 h^*}{\partial y_\beta^2} \eta^2 \sigma_\beta^2 \, dt \\ &= -\frac{\dot\eta}{\eta} H_p \, dt + \frac{\dot\eta}{\eta} \langle Y | X - p \rangle \, dt + \langle dY | X - p \rangle + \frac{\eta}{2} \sum\nolimits_\beta \frac{\partial^2 h^*}{\partial y_\beta^2} \sigma_\beta^2 \, dt, \end{aligned} \tag{3.11}$$

so, by combining the definition of H_p with (SRL-U), we get:

$$\begin{aligned} dH_p = &-\frac{\dot\eta}{\eta^2} [h(p) - h(X)] \, dt + \langle v | X - p \rangle \, dt \\ &+ \sum\nolimits_\beta (X_\beta - p_\beta) \, \sigma_\beta \, dW_\beta + \frac{\eta}{2} \sum\nolimits_\beta \frac{\partial^2 h^*}{\partial y_\beta^2} \sigma_\beta^2 \, dt, \end{aligned} \tag{3.12}$$

where we used the fact that $h^*(\eta Y) = \langle \eta Y | X \rangle - h(X)$ in the first line. Therefore, the player's cumulative regret for not playing p up to time t will be:

$$\int_0^t \langle v(s) | p - X(s) \rangle \, ds = H_p(0) - H_p(t) \tag{3.13a}$$

$$- \int_0^t \frac{\dot\eta(s)}{\eta^2(s)} [h(p) - h(X(s))] \, ds \tag{3.13b}$$

$$+ \sum\nolimits_\beta \int_0^t (X_\beta(s) - p) \, \sigma_\beta(s) \, dW_\beta(s) \tag{3.13c}$$

$$+ \frac{1}{2} \sum\nolimits_\beta \int_0^t \eta(s) \frac{\partial^2 h^*}{\partial^2 y_\beta^2} \sigma_\beta^2(s) \, ds. \tag{3.13d}$$

We now proceed to bound each term of (3.13) by a sublinear function:

a) Since $H_p \geq 0$, the term (3.13a) is bounded from above by $H_p(0)$, a constant.

b) For the second term, let $R = \sqrt{2 \max_{x,x' \in \mathcal{X}} \{h(x) - h(x')\}}$ denote the so-called h-radius of $\mathcal{X}$ (Nesterov, 2009). Then, $h(p) - h(X(s)) \leq R^2/2$ and hence:

$$(3.13b) \leq -\frac{R^2}{2} \int_0^t \frac{\dot\eta(s)}{\eta^2(s)} \, ds = \frac{R^2}{2} \left(\frac{1}{\eta(t)} - \frac{1}{\eta(0)} \right) = o(t) \tag{3.14}$$

because $\lim_{t\to\infty} t\eta(t) = \infty$ by assumption (recall also that $\dot\eta \leq 0$).

c) For (3.13c), sublinearity follows directly from Lemma 3.2.

d) Finally, for (3.13d), recall that the Hessian of $h^*(y)$ is equal to the inverse Hessian of $h(x)$, suitably restricted to the affine hull of the face of the simplex

spanned by $x = Q(y)$ – see e.g. Chap. 26 in Rockafellar (1970). Thus, given that h is K-strongly convex, (2.2) readily yields $\partial^2 h^* / \partial y_\beta^2 \leq K^{-1}$, and hence:

$$(3.13\text{d}) \leq \frac{|\mathcal{A}|}{2K} \sigma_{\max}^2 \int_0^t \eta(s)\, ds, \tag{3.15}$$

where $\sigma_{\max}^2 = \sup_{t \geq 0} \max_\beta \sigma_\beta^2(t)$. L'Hôpital's rule then yields $t^{-1} \int_0^t \eta(s)\, ds \sim \eta(t) \to 0$, so (3.13d) is also sublinear in t.

Combining all of the above, we conclude that

$$\operatorname{Reg}(t) = \max_{p \in \mathcal{X}} \int_0^t \langle v(s) | p - X(s) \rangle \, ds = o(t), \tag{3.16}$$

i.e. (SRL-U) is consistent. □

Theorem 3.1 posits that the player's learning parameter decreases to 0 at a controlled rate so that $t\eta(t) \to \infty$. A standard choice satisfying these desiderata is to pick $\eta(t) = t^{-\gamma}$ for some $\gamma \in (0, 1)$; in this case, we obtain the following regret minimization rate:

Proposition 3.3. *Assume that* (SRL-U) *is run with learning rate* $\eta(t) \sim t^{-\gamma}$ *for some* $\gamma \in (0, 1)$. *Then:*

$$\operatorname{Reg}(t) = \begin{cases} \mathcal{O}\left(t^{1-\gamma}\right) & \text{if } 0 < \gamma < \frac{1}{2}, \\ \mathcal{O}\left(\sqrt{t \log\log t}\right) & \text{if } \gamma = \frac{1}{2}, \\ \mathcal{O}\left(t^{\gamma}\right) & \text{if } \frac{1}{2} < \gamma < 1. \end{cases} \tag{3.17}$$

Proof. By substituting $\eta(t) \propto t^{-\gamma}$ in the bound (3.13), we get an asymptotic behavior of the form $\mathcal{O}(1)$ for (3.13a), $\mathcal{O}(t^\gamma)$ for (3.13b), $\mathcal{O}(\sqrt{t \log\log t})$ for (3.13c) (by the law of the iterated logarithm), and $\mathcal{O}(t^{1-\gamma})$ for (3.13d). The bound (3.17) then follows by identifying the dominant term in each case. □

We close this section by discussing the links of (SRL-U) with the process of vanishingly smooth fictitious play that was recently introduced by Benaïm and Faure (2013) in the discrete-time context described in Remark 3.1 Specifically, by interpreting $t^{-1} \int_0^t v(s)\, ds$ as the payoff that a player obtains in a 2-player game against his opponent's empirical frequency of play (cf. Section 6), the strategy

$$x(t) = Q\left(\eta(t) \int_0^t v(s)\, ds\right) = Q\left(t\eta(t) \cdot t^{-1} \int_0^t v(s)\, ds\right) \tag{3.18}$$

can be interpreted itself as a "vanishingly smooth" best response to the empirical frequency of play of one's opponent: it is "smooth" because the player is employing a regularized best response map instead of the hard $\arg\max$ correspondence, and it is "vanishingly smooth" because the factor $t\eta(t)$ hardens to ∞ as $t \to \infty$ (for a more detailed discussion, see Benaïm and Faure, 2013). In this way, (SRL-U) can be seen as a stochastically perturbed variant of vanishingly smooth fictitious play in continuous time and Proposition 3.3 provides the analogue of Theorem 1.8 of Benaïm and Faure (2013) in a continuous-time, stochastic setting.

4. Extinction of Dominated Strategies

A fundamental rationality requirement for any game-theoretic learning process is the elimination of suboptimal, dominated strategies. Formally, given a finite game $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$, we say that $p_k \in \mathcal{X}_k$ *is dominated by* $p'_k \in \mathcal{X}_k$ (and we write $p_k \prec p'_k$) if

$$\langle v_k(x)|p_k\rangle < \langle v_k(x)|p'_k\rangle \quad \text{for all } x \in \mathcal{X}. \tag{4.1}$$

Strategies that are *iteratively dominated*, *undominated*, or *iteratively undominated* are defined similarly; also, if the strategies in question are pure, we obviously have $\alpha \prec \beta$ if and only if

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

With this in mind, given a trajectory of play $x(t) \in \mathcal{X}$, $t \geq 0$, we say that a 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 (1992), we will say that the mixed strategy $p_k \in \mathcal{X}_k$ becomes extinct along $x(t)$ if $\min\{x_{k\alpha}(t) : \alpha \in \operatorname{supp}(p_k)\} \to 0$; otherwise, we say that p_k *survives*.

In a deterministic context, Akin (1980), Nachbar (1990) and Samuelson and Zhang (1992) showed that dominated strategies become extinct under the replicator dynamics (RD); this result was then extended by Mertikopoulos and Sandholm (2014) to the more general case of the deterministic reinforcement learning dynamics (RL) with arbitrary regularized best response maps. In the stochastic case, Cabrales (2000), Imhof (2005) and Hofbauer and Imhof (2009) showed that dominated strategies are eliminated under the replicator dynamics with aggregate shocks (ASRD), provided that the variance of the noise across different strategies is small enough. Surprisingly however, no "small noise" condition is required for the stochastic replicator dynamics of exponential learning: Mertikopoulos and Moustakas (2010) showed that the constant η variant of (SRD) eliminates dominated strategies, irrespective of the noise level.

Our main result in this section is that this is a special case of a much more general elimination principle:

Theorem 4.1. *Let $X(t)$ be a solution orbit of (SRL). If $p_k \in \mathcal{X}_k$ is dominated (even iteratively), then it becomes extinct along $X(t)$ almost surely.*

The basic idea of our proof will be to show that the Itô process $X(t)$ has a dominant drift coefficient that pushes it away from any dominated strategy p_k. However, given the complicated form of the (non-autonomous) dynamics (2.7), we will do so indirectly, by studying the Fenchel coupling between $p_k \in \mathcal{X}_k$ and the generating process $Y_k(t) \in V_k^*$.

Proof of Theorem 4.1. Suppose $p_k \prec p'_k$ so that $\langle v_k(x)|p'_k - p_k\rangle \geq m_k$ for some $m_k > 0$ and for all $x \in \mathcal{X}$. Then, if $Y(t)$ is a solution orbit of (SRL), we get:

$$\begin{aligned} d\langle Y_k|p'_k - p_k\rangle = \langle dY_k|p'_k - p_k\rangle &= \langle v_k|p'_k - p_k\rangle\, dt + \sum\nolimits_\beta^k (p'_{k\beta} - p_{k\beta})\, \sigma_{k\beta}\, dW_{k\beta} \\ &\geq m_k\, dt + \sum\nolimits_\beta^k (p'_{k\beta} - p_{k\beta})\, \sigma_{k\beta}\, dW_{k\beta}, \end{aligned} \tag{4.3}$$

or, equivalently:

$$\langle Y_k(t)|p_k - p'_k\rangle \geq c_k + m_k t + \xi_k(t), \tag{4.4}$$

where $c_k = \langle Y_k(0)|p_k - p'_k\rangle$ and

$$\xi_k(t) = \sum\nolimits_{\beta}^{k}(p_{k\beta} - p'_{k\beta}) \int_0^t \sigma_{k\beta}(X(s))\, dW_{k\beta}(s). \tag{4.5}$$

Consider now the rate-adjusted "cross-coupling"

$$\begin{aligned} V_k(y_k) &= \eta_k^{-1}\left[F_k(p_k, \eta_k y_k) - F_k(p'_k, \eta_k y_k)\right] \\ &= \eta_k^{-1}\left[h_k(p_k) - h_k(p'_k)\right] - \langle y_k|p_k - p'_k\rangle\,, \end{aligned} \tag{4.6}$$

with $F_k\colon \mathcal{X}_k \times V_k^* \to \mathbb{R}$ is defined as in (3.4). Then, by substituting (4.4) in (4.6) and recalling that $F_k(p'_k, y_k) \geq 0$, we obtain:

$$F_k(p_k, \eta_k Y_k) \geq h_k(p_k) - h_k(p'_k) + \eta_k \cdot [c_k + m_k t + \xi_k(t)]\,. \tag{4.7}$$

However, by Lemma 3.2, we also have $m_k t + \xi_k(t) \sim m_k t$, so the RHS of (4.7) tends to infinity as $t \to \infty$ on account of the fact that $t\eta_k(t) \to +\infty$ (cf. Assumption 1). In turn, this gives $F_k(p_k, \eta_k(t)Y_k(t)) \to +\infty$, so p_k becomes extinct along $X(t) = Q(\eta(t)Y(t))$ by virtue of Proposition A.1. Finally, the result for iteratively dominated strategies follows by a standard induction argument on the rounds of elimination of dominated strategies – see e.g. Cabrales (2000, Proposition 1A). □

Theorem 4.1 shows that dominated strategies become extinct under (SRL) but it does not provide any information on how fast they vanish. Below we provide a "large deviations" bound for the probability of observing a dominated strategy above a given level at time $t \geq 0$; for simplicity, we present our result in the case where the players' regularized best responses are derived from decomposable penalty functions of the general form (2.6):

Proposition 4.2. *Let $\alpha \in \mathcal{A}_k$ be dominated by $\beta \in \mathcal{A}_k$ and assume that the regularized best response map of player k is generated by a steep penalty function of the decomposable form (2.6) with $\lim_{x\to 0^+} \theta'_k(x) = -\infty$. Then, for all $\delta > 0$ and for all large enough $t \geq 0$, we have:*

$$\mathbb{P}\left(X_{k\alpha}(t) > \delta\right) \leq \frac{1}{2}\operatorname{erfc}\left[\frac{1}{2\sigma_{\alpha\beta}}\left(m_k\sqrt{t} - \frac{C_k - \theta'_k(\delta)}{\eta_k(t)\sqrt{t}}\right)\right], \tag{4.8}$$

where $\operatorname{erfc}(z) = \frac{2}{\sqrt{\pi}}\int_z^\infty e^{-t^2}dt$ is the complementary error function and:

a) $m_k = \min_{x\in\mathcal{X}}\{v_{k\beta}(x) - v_{k\alpha}(x)\} > 0$ is the minimum payoff difference between α and β.
b) C_k is a constant that depends only on the initial conditions of (SRL).
c) $\sigma^2_{\alpha\beta} = \frac{1}{2}\max_{x\in\mathcal{X}}\left\{\sigma^2_{k\alpha}(x) + \sigma^2_{k\beta}(x)\right\} > 0$.

Proof. Let $X(t)$ be a solution of (SRL) with $X(0) = x \in \mathcal{X}$. Then, suppressing the player index k for simplicity, we obtain:

$$\begin{aligned} dY_\alpha - dY_\beta &= (v_\alpha(X) - v_\beta(X))\, dt + \sigma_\alpha(X)\, dW_\alpha - \sigma_\beta(X)\, dW_\beta \\ &\leq -m\, dt - d\xi, \end{aligned} \tag{4.9}$$

where $\xi(t) = \int_0^t \sigma_\beta(X(s))\, dW_\beta(s) - \int_0^t \sigma_\alpha(X(s))\, dW_\alpha(s)$ and m is defined as above. We thus get

$$Y_\alpha(t) - Y_\beta(t) \leq Y_\alpha(0) - Y_\beta(0) - mt - \xi(t), \tag{4.10}$$

so the KKT conditions for the problem (2.1) give:

$$\theta'(X_\alpha(t)) - \theta'(X_\beta(t)) \leq \eta(t) \cdot [Y_\alpha(0) - Y_\beta(0) - mt - \xi(t)] \tag{4.11}$$

(recall that θ is steep so the solutions of (2.1) are interior). Thus, with $\eta(t)$ decreasing, we finally get:

$$\theta'(X_\alpha(t)) \leq C - \eta(t) \cdot [mt + \xi(t)], \tag{4.12}$$

where $C = \theta'(1) - \eta(0) \cdot |Y_\alpha(0) - Y_\beta(0)|$ depends only on the initial conditions of (SRL).

Now, if $\rho = [\xi, \xi]$ denotes the quadratic variation of ξ (cf. the proof of Lemma 3.2), the independence of W_α and W_β yields $d\rho = d\xi \cdot d\xi = \left(\sigma_\alpha^2 + \sigma_\beta^2\right) dt$ and hence:

$$\rho(t) = \int_0^t \left[\sigma_\alpha^2(X(s)) + \sigma_\beta^2(X(s))\right] ds \leq 2\sigma_{\alpha\beta}^2\, t. \tag{4.13}$$

Consequently, invoking the time-change theorem for martingales, let $\widetilde{W}$ be a Wiener process such that $\xi(t) = \widetilde{W}(\rho(t))$; then, combining (4.12) and (4.13), we obtain:

$$\begin{aligned}
\mathbb{P}\left(X_\alpha(t) > \delta\right) &= \mathbb{P}\left(\theta'(X_\alpha(t)) > \theta'(\delta)\right) \\
&\leq \mathbb{P}\left(\widetilde{W}(\rho(t)) < -mt + \frac{C - \theta'(\delta)}{\eta(t)}\right) \\
&= \frac{1}{2} \operatorname{erfc}\left[\frac{1}{\sqrt{2\rho(t)}}\left(mt - \frac{C - \theta'(\delta)}{\eta(t)}\right)\right].
\end{aligned} \tag{4.14}$$

Since $m > 0$, Assumption 1 for the rate of decay of $\eta(t)$ guarantees that $mt\eta(t) > C - \theta'(\delta)$ for large t; (4.8) then follows by substituting (4.13) in (4.14). □

Proposition 4.2 shows that $\mathbb{P}(X_{k\alpha}(t) < \delta)$ vanishes as $t \to \infty$; in particular, by expanding the complementary error function around $t = \infty$, we get:

$$\mathbb{P}(X_\alpha(t) > \delta) = \mathcal{O}\left(t^{-1/2} \exp\left(-A_k^2 t\right)\right), \tag{4.15}$$

with $A_k = m_k/(2\sigma_{\alpha\beta})$. The following proposition establishes a bound for the expected value of the corresponding hitting time:

Proposition 4.3. *With notation as in Proposition 4.2, assume that* (SRL) *is run with constant learning rates η_k and noisy observations with constant variance. If $\tau_\delta = \inf\{t > 0 : X_{k\alpha}(t) \leq \delta\}$, then:*

$$\mathbb{E}[\tau_\delta] \leq \frac{\left[C_k - \theta_k'(\delta)\right]_+}{\eta_k m_k}. \tag{4.16}$$

Remark 4.1. Note that the bound (4.16) is independent of the variance of the noise in (SRL); in other words, the diffusion coefficients σ affect the probability of observing a dominated strategy at a high share, but not its mean elimination rate.

Proof. With notation as in the proof of Proposition 4.2, let $Z_{\alpha\beta}$ denote the RHS of (4.9), viz.:

$$dZ_{\alpha\beta} = m_k\, dt + \sigma_{k\beta}\, dW_{k\beta} - \sigma_{k\alpha}\, dW_{k\alpha}. \tag{4.17}$$

Then, by (4.12), we readily obtain

$$\tau_\delta \leq \tilde{\tau}_a \equiv \inf\left\{t > 0 : Z_{\alpha\beta}(t) \leq a\right\}, \tag{4.18}$$

with $a = \eta_k^{-1}[C_k - \theta_k'(\delta)]_+$ (obviously, $\tau_\delta = 0$ if $C_k - \theta_k'(\delta) < 0$). Since $Z_{\alpha\beta}$ is simply a Wiener process with drift m_k, a standard argument based on Dynkin's formula yields $\mathbb{E}[\tilde{\tau}_a] = a/m_k$ and (4.16) follows. □

5. Long-Term Stability and Convergence Analysis

In this section, we focus on the long-term stability and convergence properties of the stochastic reinforcement learning dynamics (SRL) with respect to equilibrium play. To that end, recall first that a strategy profile $x^* \in \mathcal{X}$ is a *Nash equilibrium* of $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ if it is unilaterally stable for all players, i.e.

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

or, equivalently:

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

Strict equilibria are defined by requiring that (5.1) hold as a strict inequality for all $x_k \neq x^*_k$; obviously, such equilibria are also pure in the sense that they correspond to pure strategy profiles in $\mathcal{A} = \prod_k \mathcal{A}_k$ (i.e. vertices of $\mathcal{X}$).

In the noiseless case ($\sigma = 0$) with constant learning rates ($\dot{\eta} = 0$), Mertikopoulos and Sandholm (2014) recently showed that the deterministic dynamics (RL) exhibit the following properties with respect to Nash equilibria of $\mathfrak{G}$:

a) If a solution orbit of (RL) converges to x^*, then x^* is Nash.
b) If $x^* \in \mathcal{X}$ is (Lyapunov) stable, then it is also Nash.
c) Strict Nash equilibria are asymptotically stable in (RL).

In turn, these properties are generalizations of the long-term stability and convergence properties of the (multi-population) replicator dynamics which are sometimes referred to as the "folk theorem" of evolutionary game theory (Hofbauer and Sigmund, 1998, 2003). That being said, the situation is quite different in the presence of noise: for instance, interior Nash equilibria are not even traps (almost sure rest points) of the stochastic reinforcement learning dynamics (SRL), so the ordinary (deterministic) definitions of stability and convergence no longer apply. In the context of stochastic differential equations (for an in-depth treatment, see Khasminskii, 2012), Lyapunov and asymptotic stability are defined as follows:

Definition 5.1. Let $x^* \in \mathcal{X}$. We will say that:

(1) x^* is *stochastically* (*Lyapunov*) *stable* under (SRL) if, for every $\varepsilon > 0$ and for every neighborhood U_0 of x^* in $\mathcal{X}$, there exists a neighborhood $U \subseteq U_0$ of x^* such that

$$\mathbb{P}(X(t) \in U_0 \text{ for all } t \geq 0) \geq 1 - \varepsilon, \tag{5.3}$$

whenever $X(0) \in U$.

(2) x^* is *stochastically asymptotically stable* under (SRL) if it is stochastically stable and attracting: for every $\varepsilon > 0$ and for every neighborhood U_0 of x^* in $\mathcal{X}$, there exists a neighborhood $U \subseteq U_0$ of x^* such that

$$\mathbb{P}\left(X(t) \in U_0 \text{ for all } t \geq 0 \text{ and } \lim_{t\to\infty} X(t) = x^*\right) \geq 1 - \varepsilon, \tag{5.4}$$

whenever $X(0) \in U$.

In the evolutionary setting of (ASRD), Imhof (2005) and Hofbauer and Imhof (2009) showed that strict Nash equilibria are stochastically asymptotically stable provided that the variability of the shocks across different strategies is small enough. More recently, in a learning context, Mertikopoulos and Moustakas (2010) showed that the same holds for the stochastic replicator dynamics (SRD) of exponential learning (with constant η), irrespective of the variance of the observation noise. However, this last result relies heavily on the properties of the logit map (2.3) and

the specific form of the infinitesimal generator of (SRD). In our case, the convoluted (and non-autonomous) form of the stochastic dynamics (2.7) complicates things considerably, so such an approach does not seem possible. Nonetheless, by working directly on the dual space V^*, we obtain the following general result:

Theorem 5.2. *Let $X(t)$ be a solution orbit of (SRL) and let $x^* \in \mathcal{X}$. Then:*

1. *If $\mathbb{P}(\lim_{t\to\infty} X(t) = x^*) > 0$, x^* is a Nash equilibrium of $\mathfrak{G}$.*
2. *If x^* is stochastically (Lyapunov) stable, it is also Nash.*
3. *If x^* is a strict Nash equilibrium of $\mathfrak{G}$, it is stochastically asymptotically stable under (SRL).*

Contrary to the approach of Imhof (2005), Hofbauer and Imhof (2009) and Mertikopoulos and Moustakas (2010), our proof does not rely on the stochastic Lyapunov method (Khasminskii, 2012) and the infinetesimal generator of (2.7). Instead, we take a more direct approach that relies on two auxiliary results, one regarding the structure of (SRL) near quasi-stable points, and an estimate of the probability that a Brownian motion intersects the line $a + bt$ in finite time (a textbook application of Girsanov's theorem).

Proposition 5.3. *With notation as in Theorem 5.2, assume that every neighborhood U of x^* in $\mathcal{X}$ admits with positive probability a solution orbit $X(t)$ of (SRL) such that $X(t) \in U$ for all $t \geq 0$. Then, x^* is a Nash equilibrium.*

Proof. If x^* is not Nash, we must have $v_{k\alpha}(x^*) < v_{k\beta}(x^*)$ for some player $k \in \mathcal{N}$ and for some $\alpha \in \operatorname{supp}(x_k^*)$, $\beta \in \mathcal{A}_k$. On that account, let U be a sufficiently small neighborhood of x^* in $\mathcal{X}$ such that $v_{k\beta}(x) - v_{k\alpha}(x) \geq m_k$ for some $m_k > 0$ and for all $x \in U$. Then, conditioning on the positive probability event that there exists an orbit $X(t) = Q(\eta(t)Y(t))$ of (SRL) that is contained in U for all $t \geq 0$, we have:

$$\begin{aligned} dY_{k\alpha} - dY_{k\beta} &= (v_{k\alpha}(X) - v_{k\beta}(X))\,dt + \sigma_{k\alpha}\,dW_{k\alpha} - \sigma_{k\beta}\,dW_{k\beta} \\ &\leq -m_k\,dt - d\xi_k, \end{aligned} \tag{5.5}$$

where $\xi_k(t)$ is defined as in the proof of Proposition 4.2. By Lemma 3.2, it follows that $\xi_k(t) + m_k t \sim m_k t$ (a.s.), so

$$\eta_k(t) \cdot \big[Y_{k\alpha}(t) - Y_{k\beta}(t)\big] \leq -\frac{m_k t + \xi_k(t)}{t} \cdot \eta_k(t) t \to -\infty, \tag{5.6}$$

on account of Assumption 1. Proposition A.1 then shows that $Q_{k\alpha}(\eta_k Y_k(t)) \to 0$ as $t \to \infty$, contradicting the assumption that $X(t) \in U$ for all t (recall that $\alpha \in \operatorname{supp}(x_k^*)$). We thus conclude that x^* is a Nash equilibrium of $\mathfrak{G}$, as claimed. □

Lemma 5.4. *Let $W(t)$ be a standard one-dimensional Wiener process and consider the hitting time $\tau_{a,b} = \inf\{t > 0 : W(t) = a + bt\}$, $a, b \in \mathbb{R}$. Then:*

$$\mathbb{P}(\tau_{a,b} < \infty) = \exp(-ab - |ab|). \tag{5.7}$$

Lemma 5.4 is a textbook exercise in Girsanov's theorem and its proof is fairly standard; we only provide a short proof for the sake of completeness:

Proof. Let $\overline{W}(t) = W(t) - bt$ so that $\tau_{a,b} = \inf\{t > 0 : \overline{W}(t) = a\}$. By Girsanov's theorem (see e.g. Øksendal, 2007, Chap. 8), there exists a probability measure $\mathbb{Q}$

such that *a*) $\overline{W}$ is a Brownian motion with respect to $\mathbb{Q}$; and *b*) the Radon–Nikodym derivative of $\mathbb{Q}$ with respect to $\mathbb{P}$ satisfies

$$\left.\frac{d\mathbb{Q}}{d\mathbb{P}}\right|_{\mathcal{F}_t} = \exp\left(-b^2t/2 + bW(t)\right) = \exp\left(b^2t/2 + b\overline{W}(t)\right), \tag{5.8}$$

where $\mathcal{F}_t$ denotes the natural filtration of $W(t)$. We then get

$$\begin{aligned}
\mathbb{P}\left(\tau_{a,b} < t\right) &= \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}(\tau_{a,b} < t)\right] \\
&= \mathbb{E}_{\mathbb{Q}}\left[\mathbb{1}(\tau_{a,b} < t) \cdot \exp(-b^2t/2 - b\overline{W}(t))\right] \\
&= \mathbb{E}_{\mathbb{Q}}\left[\mathbb{1}(\tau_{a,b} < t) \cdot \exp(-b^2\tau_{a,b}/2 - b\overline{W}(\tau_{a,b}))\right] \\
&= \exp(-ab)\mathbb{E}_{\mathbb{Q}}\left[\mathbb{1}(\tau_{a,b} < t) \cdot \exp(-b^2\tau_{a,b}/2)\right],
\end{aligned} \tag{5.9}$$

and hence:

$$\begin{aligned}
\mathbb{P}\left(\tau_{a,b} < \infty\right) &= \lim_{t\to\infty} \mathbb{P}\left(\tau_{a,b} < t\right) \\
&= \lim_{t\to\infty} \exp(-ab)\, \mathbb{E}_{\mathbb{Q}}\left[\mathbb{1}(\tau_{a,b} < t) \cdot \exp(-b^2\tau_{a,b}/2)\right] \\
&= \exp(-ab)\, \mathbb{E}_{\mathbb{Q}}\left[\exp(-b^2\tau_{a,b}/2)\right] \\
&= \exp(-ab - |ab|),
\end{aligned} \tag{5.10}$$

where, in the last step above, we used the well-known expression $\mathbb{E}[\exp(-\lambda\tau_a)] = \exp(-a\sqrt{2\lambda})$ for the Laplace transform of the Brownian hitting time $\tau_a = \inf\{t > 0 : W(t) = a\}$ (see e.g. Karatzas and Shreve, 1998). □

Proof of Theorem 5.2. Parts 1 and 2 of Theorem 5.2 follow readily from Proposition 5.3 – simply note that the hypothesis of Proposition 5.3 is satisfied in both cases. As such, we only need to show that strict Nash equilibria are stochastically asymptotically stable under (SRL).

To that end, let $x^* = (\alpha_1^*, \ldots, \alpha_N^*)$ be a strict equilibrium of $\mathfrak{G}$ and let $\mathcal{A}_k^* \equiv \mathcal{A}_k \setminus \{\alpha_k^*\}$. Moreover, for all $\alpha \in \mathcal{A}_k^*$, set

$$Z_{k\alpha} = \eta_k \left(Y_{k\alpha} - Y_{k\alpha_k^*}\right), \tag{5.11}$$

so that $X(t) \to x^*$ if and only if $Z_{k\alpha}(t) \to -\infty$ for all $\alpha \in \mathcal{A}_k^*$, $k \in \mathcal{N}$ (cf. Proposition A.1).

Now, fix some tolerance $\varepsilon > 0$ and a neighborhood U_0 of x^* in $\mathcal{X}$. Since x^* is a strict equilibrium of $\mathfrak{G}$, there exist $m_k > 0$ and a neighborhood $U \subseteq U_0$ of x^* such that

$$v_{k\alpha_k^*}(x) - v_{k\alpha}(x) \geq m_k \quad \text{for all } x \in U \text{ and for all } \alpha \in \mathcal{A}_k^*,\ k \in \mathcal{N}. \tag{5.12}$$

With this in mind, let $M > 0$ be sufficiently large so that $X(t) \in U$ if $Z_{k\alpha}(t) \leq -M$ for all $\alpha \in \mathcal{A}_k^*$, $k \in \mathcal{N}$ (that such an M exists is again a consequence of Proposition A.1); furthermore, with a fair degree of hindsight, assume also that

$$M > m_k^{-1}\eta_k(0)\sigma_{k,\max}^2 \log(N/\varepsilon) \quad \text{for all } k = 1, \ldots, N, \tag{5.13}$$

where $\sigma_{k,\max}^2 = \max_{\alpha,x} \sigma_{k\alpha}^2(x)$. We will show that if $Z_{k\alpha}(0) \leq -2M$, then $X(t) \in U$ for all $t \geq 0$ and $Z_{k\alpha}(t) \to -\infty$ with probability at least $1 - \varepsilon$.

Indeed, assume that $Z_{k\alpha}(0) \leq -2M$ in (5.11) and define the escape time:

$$\tau_U = \inf\{t > 0 : X(t) \notin U\}. \tag{5.14}$$

The stochastic dynamics (SRL) then give:

$$d\left(Y_{k\alpha} - Y_{k\alpha_k^*}\right) = \left[v_{k\alpha} - v_{k\alpha_k^*}\right] dt + \sigma_{k\alpha}\, dW_{k\alpha} - \sigma_{k\alpha_k^*}\, dW_{k\alpha_k^*}, \tag{5.15}$$

so, for all $t \leq \tau_U$, we will have:

$$\begin{aligned} Z_{k\alpha}(t) &= Z_{k\alpha}(0) + \eta_k(t) \int_0^t \left[v_{k\alpha}(X(s)) - v_{k\alpha_k^*}(X(s))\right] ds + \eta_k(t)\xi_k(t) \\ &\leq -2M - \eta_k(t)\left[mt - \xi_k(t)\right], \end{aligned} \tag{5.16}$$

where we have set $\xi_k(t) = \int_0^t \sigma_{k\alpha}(X(s))\, dW_{k\alpha}(s) - \int_0^t \sigma_{k\alpha_k^*}(X(s))\, dW_{k\alpha_k^*}(s)$.

We will first show that $\mathbb{P}(\tau_U < \infty) \leq \varepsilon$. To that end, note that the time-change theorem for martingales (Øksendal, 2007, Cor. 8.5.4) provides a standard Wiener process $\widetilde{W}_k(t)$ such that $\xi_k(t) = \widetilde{W}_k(\rho_k(t))$ where $\rho_k = [\xi_k, \xi_k]$ is the quadratic variation of ξ_k. Then, from the fact that η is nonincreasing, we conclude that $Z_{k\alpha}(t) \leq -M$ whenever $m_k t - \widetilde{W}_k(\rho_k(t)) \geq -M/\eta_k(0)$.

Accordingly, with $\rho_k(t) \leq 2\sigma_{k,\max}^2 t$ (cf. the proof of Proposition 4.2), it suffices to show that the hitting time

$$\tau_0 = \inf\left\{t > 0 : \widetilde{W}_k(t) = \frac{m_k t}{2\sigma_{k,\max}^2} + \frac{M}{\eta_k(0)} \text{ for some } k \in \mathcal{N}\right\} \tag{5.17}$$

is finite with probability not exceeding ε. However, if a trajectory of $\widetilde{W}_k(t)$ has $\widetilde{W}_k(t) \leq m_k t/(2\sigma_{k,\max}^2) + M/\eta_k(0)$ for all $t \geq 0$, we will also have

$$\widetilde{W}_k(\rho_k(t)) \leq \frac{m_k \rho_k(t)}{2\sigma_{k,\max}^2} + \frac{M}{\eta_k(0)} \leq m_k t + \frac{M}{\eta_k(0)}, \tag{5.18}$$

so τ_U must be infinite for every trajectory of $\widetilde{W} = (\widetilde{W}_1, \dots, \widetilde{W}_N)$ with infinite τ_0, i.e. $\mathbb{P}(\tau_U < +\infty) \leq \mathbb{P}(\tau_0 < +\infty)$. Thus, if we write E_k for the event that $\widetilde{W}_k(t) \geq m_k t/(2\sigma_{k,\max}^2) + M/\eta_k(0)$ for some finite $t \geq 0$, Lemma 5.4 yields $\mathbb{P}(E_k) = e^{-\lambda_k M}$ with $\lambda_k = m_k/(\eta_k(0)\sigma_{k,\max}^2)$, and we obtain:

$$\mathbb{P}(\tau_0 < +\infty) = \mathbb{P}\left(\bigcup\nolimits_k E_k\right) \leq \sum\nolimits_k \mathbb{P}(E_k) = \sum\nolimits_k e^{-\lambda_k M} \leq \varepsilon, \tag{5.19}$$

by construction of M. In view of the above, by conditioning on the event $\tau_U = +\infty$ and invoking Assumption 1 and Lemma 3.2, Eq. (5.16) finally yields

$$Z_{k\alpha}(t) \leq -2M - \eta_k(t) \cdot \left[mt - \xi_k(t)\right] \sim -\eta_k(t) \cdot mt \to -\infty \quad \text{(a.s.)}. \tag{5.20}$$

We conclude that $Z_{k\alpha}(t) \to -\infty$ for all $\alpha \in \mathcal{A}_k^*$, $k \in \mathcal{N}$, so $\lim_{t\to\infty} X(t) = x^*$ (conditionally a.s.) and our proof is complete. □

Remark 5.1. By considering the modified game $\tilde{\mathfrak{G}}$ with noise-adjusted payoff functions $\tilde{v}_{k\alpha}(x) = v_{k\alpha}(x) - \frac{1}{2}\sigma_{k\alpha}^2$, the above reasoning yields an alternative proof of the stability and convergence results of Hofbauer and Imhof (2009) for the replicator dynamics with aggregate shocks (ASRD).

6. An Averaging Principle for 2-Player Games

In this section, we analyze the asymptotic behavior of the players' empirical distribution of play

$$\bar{X}(t) = \frac{1}{t}\int_0^t X(s)\, ds, \tag{6.1}$$

and we establish an averaging principle for (SRL) in 2-player games. Our analysis is motivated by the original deterministic results of Hofbauer and Sigmund (1998) who showed that $\bar{X}(t)$ converges to Nash equilibrium under the replicator dynamics whenever $\liminf_{t\to\infty} X_{k\alpha}(t) > 0$ for all $\alpha \in \mathcal{A}_k$, $k = 1, 2$ (see also Hofbauer et al., 2009). More recently, Mertikopoulos and Sandholm (2014) proved a version of this result for arbitrary regularized best response maps (always in a deterministic setting), while Hofbauer and Imhof (2009) showed that the time-averages of the aggregate-shocks replicator dynamics (ASRD) converge to the Nash set of a modified game (cf. Remark 5.1 above).

Our main result here is that the averaging principle of Hofbauer and Sigmund (1998) extends to the perturbed reinforcement learning scheme (SRL), even in the presence of arbitrarily large payoff observation errors:

Theorem 6.1. *Let $\mathfrak{G}$ be a 2-player game and let $X(t)$ be a solution orbit of the stochastic dynamics (SRL). If the players' score differences $Y_{k\alpha}(t) - Y_{k\beta}(t)$ grow sublinearly with t for all $\alpha, \beta \in \mathcal{A}_k$, $k = 1, 2$, then the empirical distribution of play converges almost surely to the set of Nash equilibria of $\mathfrak{G}$.*

Remark 6.1. In the case of (EW)/(RD), the sublinear growth requirement for $Y_{k\alpha} - Y_{k\beta}$ boils down to the permanency condition $\liminf_{t\to\infty} X_{k\alpha}(t) > 0$, so we recover the original result of Hofbauer and Sigmund (1998).

Proof of Theorem 6.1. Pick $\alpha, \beta \in \mathcal{A}_k$. Then, by the definition of the dynamics (SRL), we get:

$$\begin{aligned} Y_{k\alpha}(t) - Y_{k\beta}(t) &= Y_{k\alpha}(0) - Y_{k\beta}(0) + \int_0^t [v_{k\alpha}(X(s)) - v_{k\beta}(X(s))]\, ds \\ &\quad + \int_0^t \sigma_{k\alpha}(X(s))\, dW_{k\alpha}(s) - \int_0^t \sigma_{k\beta}(X(s))\, dW_{k\beta}(s) \\ &= c_k + \xi_k(t) + t\left[v_{k\alpha}(\bar{X}(t)) - v_{k\beta}(\bar{X}(t))\right], \end{aligned} \tag{6.2}$$

where $c_k = Y_{k\alpha}(0) - Y_{k\beta}(0)$, $\xi_k(t)$ denotes the martingale part of (6.2) and we have used the fact that $v_k(x)$ is linear in x (and not only multilinear). Dividing by t and using Lemma 3.2 then yields

$$\lim_{t\to\infty} \left[v_{k\alpha}(\bar{X}(t)) - v_{k\beta}(\bar{X}(t))\right] = 0 \quad \text{(a.s.)}, \tag{6.3}$$

so $v_{k\alpha}(x^*) = v_{k\beta}(x^*)$ whenever x^* is an ω-limit of $\bar{X}(t)$. This shows that any ω-limit of $\bar{X}(t)$ is Nash, and since the ω-set of $\bar{X}(t)$ is nonempty (by compactness of $\mathcal{X}$), our claim follows. □

Of course, the applicability of Theorem 6.1 is limited by the growth requirement for $Y_{k\alpha}(t) - Y_{k\beta}(t)$. The following proposition shows that this condition always holds in 2-player zero-sum games under the additional assumption $\lim_{t\to\infty} \eta_k(t) = 0$:

Proposition 6.2. *Let $\mathfrak{G}$ be a 2-player zero-sum game and assume that (SRL) is run with $\eta_k(t)$ satisfying $\lim_{t\to\infty} \eta_k(t) = 0$ (in addition to Assumption 1). Then, the empirical distribution of play under (SRL) converges almost surely to the set of Nash equilibria of $\mathfrak{G}$.*

Proof. Let $p = (p_1, p_2)$ be an interior Nash equilibrium of $\mathfrak{G}$ and let $F_p(t) = \sum_{k=1,2} F_k(p_k, \eta_k(t)Y_k(t))$. Using the Itô formula of Lemma A.3, we get:

$$\begin{aligned} dF_p &= \sum_{k,\beta} (X_{k\beta} - p_{k\beta})\, d(\eta_k Y_k) + \frac{1}{2}\sum_{k,\beta} \frac{\partial^2 h_k^*}{\partial y_{k\beta}^2} \eta_k^2 \sigma_{k\beta}^2\, dt \\ &= \sum_{k=1,2} \dot\eta_k \langle Y_k | X_k - p\rangle\, dt && \text{(6.4a)} \\ &\quad + \sum_{k=1,2} \eta_k \langle v_k | X_k - p_k\rangle\, dt && \text{(6.4b)} \\ &\quad + \sum_{k,\beta} \eta_k (X_{k\beta} - p_{k\beta}) \sigma_{k\beta}\, dW_{k\beta} && \text{(6.4c)} \\ &\quad + \frac{1}{2}\sum_{k,\beta} \frac{\partial^2 h_k^*}{\partial y_{k\beta}^2} \eta_k^2 \sigma_{k\beta}^2\, dt, && \text{(6.4d)} \end{aligned}$$

where we substituted dY from (SRL) to obtain (6.4b) and (6.4c).

We now claim that $F_p(t)$ grows sublinearly in t; indeed:

a) Lemma 3.2 shows that $Y(t) = \mathcal{O}(t)$, so there exist $M_k > 0$, $k = 1, 2$, such that

$$\sum_{k=1,2} \int_0^t \dot\eta_k(s) \langle X_k(s) - p | Y_k(s)\rangle\, ds \leq \sum_{k=1,2} M_k \int_0^t |\dot\eta_k(s)|\, s\, ds = o(t), \tag{6.5}$$

where the sublinearity estimate follows from Assumption 1 and the fact that $\eta_k(t) \to 0$:

$$\int_0^t \dot\eta_k(s) s\, ds = \eta_k(t) t - \int_0^t \eta_k(s)\, ds = o(t). \tag{6.6}$$

b) The term (6.4b) is identically zero because $\mathfrak{G}$ is zero-sum and p is an interior equilibrium of $\mathfrak{G}$:

$$\begin{aligned} &\langle v_1(X) | X_1 - p_1\rangle + \langle v_2(X) | X_2 - p_2\rangle \\ &\qquad = u_1(X_1, X_2) - u_1(p_1, X_2) + u_2(X_1, X_2) - u_2(X_1, p_2) = 0. \end{aligned} \tag{6.7}$$

c) The term (6.4c) is sublinear (a.s.) on account of (the proof of) Lemma 3.2.

d) Finally, for (6.4d), the same reasoning as in the proof of Theorem 3.1 yields:

$$\frac{1}{2}\sum_{k,\beta} \int_0^t \frac{\partial^2 h_k^*}{\partial y_{k\beta}^2} \eta_k^2(s) \sigma_{k\beta}^2(X(s))\, ds \leq \frac{1}{2K} \sum\nolimits_{k,\beta} \sigma_{k,\max}^2 \int_0^t \eta_k^2(s)\, ds, \tag{6.8}$$

with $\sigma_{k,\max}^2 = \max_{\alpha\in\mathcal{A}_k} \max_{x\in\mathcal{X}} \sigma_{k\alpha}^2(x)$ defined as in the proof of Theorem 5.2. Since $\eta_k(t) \to 0$, this last integral is also sublinear in t, as claimed.

Assume now that $\limsup_{t\to\infty} F_p(t) \geq m$ for some $m > 0$ (otherwise we would have the stronger result $X(t) \to p$). Then, Lemma A.2 gives $|Y_{k\alpha}(t) - Y_{k\beta}(t)| = \mathcal{O}(F_p(t)) = o(t)$, so our claim follows from Theorem 6.1. □

Remark 6.2. We should note here that if players employ a constant learning parameter η_k, the term (6.4a) vanishes identically and only the Itô correction of (6.4) can lead to a linear growth for $Y_{k\alpha} - Y_{k\beta}$. This explains why the deterministic results of Hofbauer et al. (2009) and Mertikopoulos and Sandholm (2014) do not require a vanishing learning parameter; for an alternate approach covering the constant η case, see Theorem 6.3 below.

We close this section by linking the long-term behavior of time-averaged orbits of (SRL) to the deterministic best response dynamics of Gilboa and Matsui (1991):

$$\dot{x}_k \in \mathtt{br}_k(x) - x_k, \tag{BRD}$$

where $\mathtt{br}_k(x) \equiv \arg\max_{x'_k \in \mathcal{X}_k} \langle v_k(x) | x'_k \rangle$ denotes the standard (non-regularized) best response correspondence of player k. Hofbauer et al. (2009) then showed that the ω-limit set Ω of time-averaged solutions of (RD) is *internally chain transitive* (ICT) under (BRD), i.e. any two points in Ω may be joined by a piecewise continuous "chain" of arbitrarily long orbit segments of (BRD) in Ω broken by arbitrarily small jump discontinuities (in particular, ICT sets are invariant, connected and have no proper attractors; for a full development, see Benaïm, Hofbauer, and Sorin, 2005).

Mertikopoulos and Sandholm (2014) subsequently established a more general averaging principle linking the deterministic reinforcement learning scheme (RL) to (BRD). Importantly, as we show below, the *stochastic* dynamics (SRL) share the same connection to the *deterministic* best response dynamics (BRD), despite all the noise:

Theorem 6.3. *Let $X(t)$ be a solution orbit of (SRL) for a 2-player game $\mathfrak{G}$. Then, the ω-limit set of the empirical distribution of play $\bar{X}(t)$ is internally chain transitive under the deterministic best response dynamics (BRD).*

Proof. Our proof follows the approach of Hofbauer et al. (2009). Specifically, from (SRL), we have:

$$\begin{aligned} Y_k(t) &= Y_k(0) + \int_0^t v_k(X(s))\,ds + \int_0^t \sigma_k(X(s))\,dW_k(s) \\ &= t v_k(\bar{X}(t)) + Y_k(0) + \xi_k(t), \end{aligned} \tag{6.9}$$

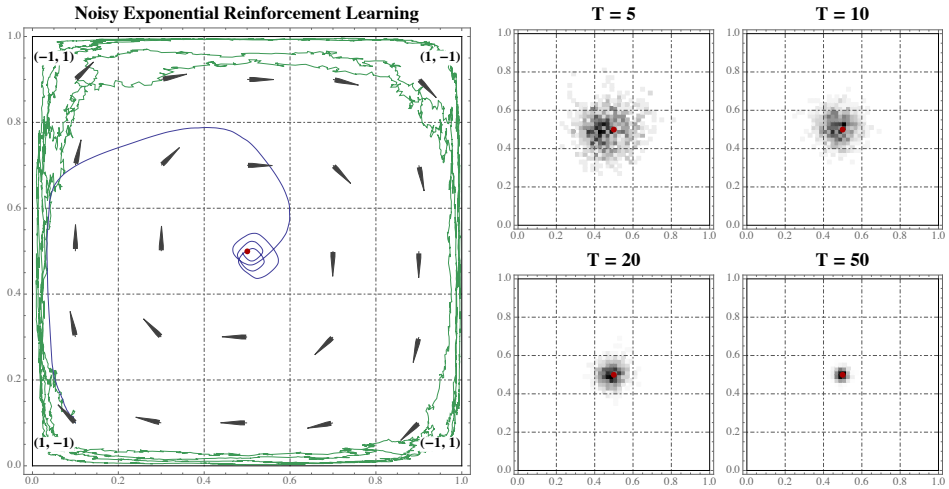
where we have set $\xi_{k\alpha}(t) = \int_0^t \sigma_{k\alpha}(X(s))\,dW_{k\alpha}(s)$ and we have used the fact that $\mathfrak{G}$ is a 2-player game in order to carry the integral inside the argument of v_k. Consequently, by the definition (2.1) of the players' regularized best response maps, $X_k(t) = Q_k(\eta_k(t) Y_k(t))$ solves the (strictly) concave maximization problem:

$$\begin{aligned} &\text{maximize} && \langle v_k(\bar{X}(t)) | x_k \rangle + \frac{1}{t} \langle Y_k(0) + \xi_k(t) | x_k \rangle - \frac{1}{t\eta_k(t)} h_k(x_k), \\ &\text{subject to} && x_k \in \mathcal{X}_k. \end{aligned} \tag{6.10}$$

By Lemma 3.2 and Assumption 1, the last two terms of (6.10) vanish as $t \to 0$. Hence, by the maximum theorem of Berge (1997, p. 116), it follows that $X_k(t)$ lies within a vanishing distance of $\arg\max_{x_k \in \mathcal{X}_k} \langle v_k(\bar{X}(t)) | x_k \rangle \equiv \mathtt{br}_k(\bar{X}(t))$. On the other hand, differentiating $\bar{X}(t)$ yields:

$$\frac{d}{dt}\bar{X}(t) = t^{-1} X(t) - t^{-2} \int_0^t X(s)\,ds = t^{-1} \left[X(t) - \bar{X}(t) \right], \tag{6.11}$$

and, after changing time to $\tau = \log t$, the expression above becomes $\frac{d}{dt}\bar{X}(t) = X - \bar{X}$. Combining all of the above, we conclude that $\bar{X}(t)$ tracks a perturbed version of the best reply dynamics (BRD) in the sense of Benaïm et al. (2005, Def. III), and our assertion follows from Theorem 3.6 in the same paper. □



(a) A sample trajectory and its time average. (b) Distribution of time averages at time T.

FIGURE 2. Time averages of (SRL) with logit best responses in a game of Matching Pennies (as in Fig. 1, Nash equilibria are depicted in red and the game's payoff's are displayed inline; for benchmarking purposes, we also took $\sigma_{k\alpha} = 1$ for all $\alpha \in \mathcal{A}_k$, $k = 1, 2$). Fig. 2(a) shows the evolution of a sample trajectory and its time average; in Fig. 2(b), we show a density plot of the distribution of 10^4 time-averaged trajectories for different values of the integration horizon T. In tune with Proposition 6.2, we see that time averages converge to the game's Nash equilibrium.

Thanks to Theorem 6.3, several conclusions of Hofbauer et al. (2009) for 2-player games can be readily generalized to the full stochastic setting of (SRL) simply by exploiting the properties of the deterministic dynamics (BRD):

(1) If the empirical distribution of play converges under (SRL), its limit is Nash.
(2) Any global attractor of (BRD) also attracts the time averages of (SRL), independently of the noise level. In particular, since the set of Nash equilibria is globally attracting under (BRD) in zero-sum games, this observation extends Proposition 6.2 to the constant η case.
(3) The only ICT sets of (BRD) in potential games consist of (isolated) components of Nash equilibria. Thus, combining Theorems 5.2 and 6.3, it follows that (SRL) converges to the set of strict Nash equilibria in 2-player potential games.

APPENDIX A. PROPERTIES OF THE FENCHEL COUPLING

In this appendix, we collect some basic properties of the Fenchel coupling and regularized best response maps. To state them, let Δ denote the d-dimensional simplex of $V \equiv \mathbb{R}^{d+1}$ and let $h\colon \Delta \to \mathbb{R}$ be a penalty function on Δ (i.e. h is continuous and strongly convex on Δ and smooth on the relative interior of any

face of Δ). Also, recall that the *regularized best response map* $Q\colon V^* \to \Delta$ induced by h is given by

$$Q(y) = \arg\max_{x\in\Delta}\{\langle y|x\rangle - h(x)\}, \quad y \in V^*, \tag{A.1}$$

while the *Fenchel coupling* between $p \in \Delta$ and $y \in V^*$ is defined as:

$$F(p, y) = h(p) + h^*(y) - \langle y|p\rangle\,, \tag{A.2}$$

where

$$h^*(y) = \max_{x\in\Delta}\{\langle y|x\rangle - h(x)\} \tag{A.3}$$

denotes the convex conjugate of h. We then have:

Proposition A.1. *Let $h\colon \Delta \to \mathbb{R}$ be a penalty function on Δ with strong convexity constant K. Then:*

1. *Q is $1/K$-Lipschitz and $Q(y) = dh^*(y)$ for all $y \in V^*$.*
2. *If $y_\alpha - y_\beta \to -\infty$ for some $\beta \neq \alpha$, then $Q_\alpha(y) \to 0$.*
3. *$F(p, y) \geq \frac{1}{2}\|Q(y) - p\|^2$ for all $p \in \Delta$ and for all $y \in V^*$; in particular, $F(p, y) \geq 0$ with equality if and only if $p = Q(y)$.*
4. *If $F(p, y_n) \to +\infty$ for some sequence $y_n \in V^*$, the sequence $x_n = Q(y_n)$ converges to the union of faces of Δ that do not contain p; in particular, $\liminf_{n\to+\infty}\{x_{n,\alpha} : \alpha \in \operatorname{supp}(p)\} = 0$.*

Proof. The first part of Proposition A.1 is well known (see e.g. Nesterov, 2009; Shalev-Shwartz, 2011); the rest of our claims are due to Mertikopoulos and Sandholm (2014), but we provide a proof for the sake of completeness.

Part 2. 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 have $\langle y_n|x_n\rangle - h(x_n) \geq \langle y_n|x'\rangle - h(x')$ for all $x' \in \Delta$, so if we set $x'_n = x_n + \varepsilon(e_\beta - e_\alpha)$, we readily obtain

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

This contradicts our original assumption that $y_{n,\alpha} - y_{n,\beta} \to -\infty$; since Δ is compact, we get $x^*_\alpha = 0$ for any limit point x^* of x_n, i.e. $Q_\alpha(y_n) \to 0$.

Part 3. Let $x = Q(y)$ so that $h^*(y) = \langle y|x\rangle - h(x)$ and $F(p, y) = h(p) - h(x) - \langle y|p - x\rangle$. Since $x = dh^*(y)$ by Part 1, we will also have

$$\langle y|x' - x\rangle \leq h(x') - h(x) \quad \text{for all } x' \in \Delta, \tag{A.5}$$

by standard convex analysis arguments (Rockafellar, 1970, Chap. 26). On the other hand, letting $z = p - x$, the definition (2.2) of strong convexity yields:

$$h(x + tz) \leq th(p) + (1 - t)h(x) - \tfrac{1}{2}Kt(1 - t)\|z\|^2 \quad \text{for all } t \in (0, 1). \tag{A.6}$$

Hence, combining (A.6) with (A.5) for $x' = x + tz$, we get

$$h(x) + \langle y|tz\rangle \leq th(p) + (1 - t)h(x) - \tfrac{1}{2}Kt(1 - t)\|z\|^2\,, \tag{A.7}$$

and, after rearranging and dividing by t, we obtain

$$F(p, y) = h(p) - h(x) - \langle y|z\rangle \geq \tfrac{1}{2}K(1 - t)\|z\|^2\,. \tag{A.8}$$

Our claim then follows by letting $t \to 0^+$ in (A.8).

Part 4. Let Λ_p denote the union of faces of Δ that do not contain p and set $\Delta_p \equiv \Delta \setminus \Lambda_p$. Then, if $x = Q(y) \in \Delta_p$ for some $y \in V^*$, the function $h(x + t(p - x))$ will be finite and smooth for all t in a neighborhood of 0 (simply note that $x + t(p - x)$ lies in the relative interior of the same face of Δ for small t). This implies that h admits a two-sided derivative at x along $p - x$, so we will also have $\langle y|p - x\rangle = h'(x; p - x) < +\infty$ by Theorem 23.5 in Rockafellar (1970); in particular, this shows that $F(p, y) = h(p) - h(x) - h'(x; p - x) < +\infty$ whenever $x = Q(y) \in \Delta_p$.

Assume now ad absurdum that $x_n = Q(y_n)$ has a limit point in Δ_p, so there exists a compact neighborhood K of p in Δ_p which is visited infinitely often by x_n. By continuity and compactness, this implies that there exists some $M > 0$ such that $F(p, y_n) = h(p) - h(x_n) - h'(x_n; p - x_n) < M$ infinitely often, a contradiction to our original assumption that $F(p, y_n) \to +\infty$. We conclude that every limit point of x_n lies in Δ_p, as claimed. □

For the proof of Proposition 6.2, we also require the following lemma:

Lemma A.2. *Let $p \in \Delta^\circ$ and let $y_n \in V^*$ have $F(p, y_n) \geq m$ for some $m > 0$. Then, for all α, β, there exists some $K > 0$ such that $|y_{n,\alpha} - y_{n,\beta}| \leq K F(p, y_n)$.*

Proof. By descending to a subsequence if necessary, assume ad absurdum that $y_{n,\alpha} - y_{n,\beta} \geq K_n F(p, y_n)$ for some increasing sequence $K_n \to +\infty$. Then, by relabeling indices (and passing to a finer subsequence if needed), we may assume that *a)* $y_{n,\alpha} \geq y_{n,\gamma} \geq y_{n,\beta}$ for all $\gamma = 1, \dots, d+1$; and *b)* the index set $\{1, \dots, d+1\}$ can be decomposed into two nonempty sets, $\mathcal{A}_0$ and $\mathcal{A}_\infty$, such that $y_{n,\alpha} - y_{n,\gamma} \leq K' F(p, y_n)$ for some $K' > 0$ and for all $\gamma \in \mathcal{A}_0$, while $y_{n,\alpha} - y_{n,\gamma} > K'_n F(p, y_n)$ for some $K'_n \to +\infty$ and for all $\gamma \in \mathcal{A}_\infty$. Thus, letting $x_n = Q(y_n)$, we get:

$$\begin{aligned}\langle y_n|p - x_n\rangle &= \sum\nolimits_{\gamma=1}^{d+1} y_{n,\gamma}(p_\gamma - x_{n,\gamma}) = \sum\nolimits_{\gamma=1}^{d+1} (y_{n,\gamma} - y_{n,\alpha})(p_\gamma - x_{n,\gamma}) \\ &= \sum_{\gamma\in\mathcal{A}_0} (y_{n,\gamma} - y_{n,\alpha})(p_\gamma - x_{n,\gamma}) + \sum_{\gamma\in\mathcal{A}_\infty} (y_{n,\gamma} - y_{n,\alpha})(p_\gamma - x_{n,\gamma}).\end{aligned} \tag{A.9}$$

By construction, the first sum is bounded in absolute value by $2dK'F(p, y_n)$. As for the second sum, since $F(p, y_n) \geq M > 0$, we will also have $y_{n,\alpha} - y_{n,\gamma} \to +\infty$, so $x_{n,\gamma} \to 0$ by Proposition A.1; with $p \in \Delta^\circ$, we then get $\liminf_n (p_\gamma - x_{n,\gamma}) \geq \delta$ for some $\delta > 0$, so every $\mathcal{A}_\infty$ summand in (A.9) is eventually bounded from above by $-\delta K'_n F(p, y_n)$. This shows that $\langle y_n|p - x_n\rangle \leq -K''_n F(p, y_n)$ for some positive sequence $K''_n \to +\infty$; however, $\langle y_n|p - x_n\rangle = h(p) - h(x_n) - F(p, y_n)$, a contradiction. □

From a stochastic analysis standpoint, the importance of the Fenchel coupling is captured by the following calculation:

Lemma A.3. *Let $h\colon \Delta \to \mathbb{R}$ be a penalty function on Δ and let $Q\colon V^* \to \Delta$ and $F\colon \Delta \times V^* \to \mathbb{R}$ be defined as above. Moreover, let $dY_\alpha = \mu_\alpha\, dt + \sum_\beta \sigma_{\alpha\beta}\, dW_\beta$ be an Itô process with values in V^* and set $X(t) = Q(Y(t))$. Then, for all $p \in \Delta$, we have:*

$$dF(p, Y) = \sum\nolimits_\beta (X_\beta - p_\beta)\, dY_\beta + \frac{1}{2} \sum\nolimits_\beta \frac{\partial^2 h^*}{\partial y_\beta^2} \sigma_\beta^2\, dt. \tag{A.10}$$

Proof. Let $H(t) = F(p, Y(t))$, $t \geq 0$. Then, Itô's formula yields:

$$\begin{aligned} dH &= \sum\nolimits_\beta \frac{\partial F}{\partial y_\beta} dY_\beta + \frac{1}{2} \sum\nolimits_{\beta,\gamma} \frac{\partial^2 F}{\partial y_\beta \partial y_\gamma} \, dY_\beta \cdot dY_\gamma \\ &= \sum\nolimits_\beta \left(\frac{\partial h^*}{\partial y_\beta} - x_\beta \right) dY_\beta + \frac{1}{2} \sum\nolimits_{\beta,\gamma} \frac{\partial^2 h^*}{\partial y_\beta^2} \sigma_\beta \sigma_\gamma \delta_{\beta\gamma} \, dt \\ &= \sum\nolimits_\beta (X_\beta - p_\beta) \, dY_\beta + \frac{1}{2} \sum\nolimits_\beta \frac{\partial^2 h^*}{\partial y_\beta^2} \sigma_\beta^2 \, dt, \end{aligned} \tag{A.11}$$

where we have used the definition of F and Proposition A.1 in the second and third lines. □

References

Akin, E., 1980: Domination or equilibrium. *Mathematical Biosciences*, **50 (3-4)**, 239–250.

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

Benaïm, M. and M. Faure, 2013: Consistency of vanishingly smooth fictitious play. *Mathematics of Operations Research*, **38 (3)**, 437–450.

Benaïm, M., J. Hofbauer, and S. Sorin, 2005: Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, **44 (1)**, 328–348.

Berge, C., 1997: *Topological Spaces*. Dover, New York.

Bregman, L. 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.

Cabrales, A., 2000: Stochastic replicator dynamics. *International Economic Review*, **41 (2)**, 451–81.

Cesa-Bianchi, N. and G. Lugosi, 2006: *Prediction, Learning, and Games*. Cambridge University Press.

Cominetti, R., E. Melo, and S. Sorin, 2010: A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior*, **70 (1)**, 71–83.

Coucheney, P., B. Gaujal, and P. Mertikopoulos, 2014: Penalty-regulated dynamics and robust learning procedures in games. *Mathematics of Operations Research*.

Freund, Y. and R. E. Schapire, 1999: Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, **29**, 79–103.

Friedman, D., 1991: Evolutionary games in economics. *Econometrica*, **59 (3)**, 637–666.

Fudenberg, D. and C. Harris, 1992: Evolutionary dynamics with aggregate shocks. *Journal of Economic Theory*, **57 (2)**, 420–441.

Fudenberg, D. and D. K. Levine, 1995: Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, **19 (5-7)**, 1065–1089.

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

Gilboa, I. and A. Matsui, 1991: Social stability and equilibrium. *Econometrica*, **59 (3)**, 859–867.

Hart, S. and A. Mas-Colell, 2000: A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, **68 (5)**, 1127–1150.

Hofbauer, J. and L. A. Imhof, 2009: Time averages, recurrence and transience in the stochastic replicator dynamics. *The Annals of Applied Probability*, **19 (4)**, 1347–1368.

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

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

Hofbauer, J. and K. Sigmund, 2003: Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, **40 (4)**, 479–519.

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

Hopkins, E., 2002: Two competing models of how people learn in games. *Econometrica*, **70 (6)**, 2141–2166.

Imhof, L. A., 2005: The long-run behavior of the stochastic replicator dynamics. *The Annals of Applied Probability*, **15 (1B)**, 1019–1045, doi:http://dx.doi.org/10.1214/105051604000000837.

Kang, W. N., F. P. Kelly, N. H. Lee, and R. J. Williams, 2009: State space collapse and diffusion approximation for a network operating under a fair bandwidth-sharing policy. *Annals of Applied Probability*, **19**, 1719–1780.

Karatzas, I. and S. E. Shreve, 1998: *Brownian Motion and Stochastic Calculus*. Springer-Verlag, Berlin.

Kelly, F. P., A. K. Maulloo, and D. K. H. Tan, 1998: Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, **49**, 237–252.

Khasminskii, R. Z., 2012: *Stochastic Stability of Differential Equations*. 2d ed., No. 66 in Stochastic Modelling and Applied Probability, Springer-Verlag, Berlin.

Khasminskii, R. Z. and N. Potsepun, 2006: On the replicator dynamics behavior under Stratonovich type random perturbations. *Stochastic Dynamics*, **6**, 197–211.

Kwon, J. and P. Mertikopoulos, 2014: A continuous-time approach to online optimization, http://arxiv.org/abs/1401.6956.

Lahkar, R. and W. H. Sandholm, 2008: The projection dynamic and the geometry of population games. *Games and Economic Behavior*, **64**, 565–590.

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

Li, J.-S., C.-T. Lee, and M.-W. Guo, 2004: Analysis, simulation and implementation of wireless TCP flows with forward error correction. *Computer Communications*, **27 (3)**, 222–229.

Littlestone, N. and M. K. Warmuth, 1994: The weighted majority algorithm. *Information and Computation*, **108 (2)**, 212–261.

Mertikopoulos, P. and A. L. Moustakas, 2009: Learning in the presence of noise. *GameNets '09: Proceedings of the 1st International Conference on Game Theory for Networks*.

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

Mertikopoulos, P. and W. H. Sandholm, 2014: Regularized best responses and reinforcement learning in games. http://arxiv.org/abs/1407.6267.

Nachbar, J. H., 1990: Evolutionary selection dynamics in games. *International Journal of Game Theory*, **19**, 59–89.

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

Nesterov, Y., 2009: Primal-dual subgradient methods for convex problems. *Mathematical Programming*, **120 (1)**, 221–259.

Øksendal, B., 2007: *Stochastic Differential Equations*. 6th ed., Springer-Verlag, Berlin.

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

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

Samuelson, L. and J. Zhang, 1992: Evolutionary stability in asymmetric games. *Journal of Economic Theory*, **57**, 363–391.

Sandholm, W. H., E. Dokumacı, and R. Lahkar, 2008: The projection dynamic and the replicator dynamic. *Games and Economic Behavior*, **64**, 666–683.

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

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

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

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

Tuyls, K., P. J. 't Hoen, and B. Vanschoenwinkel, 2006: An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems*, **12 (1)**, 115–153.

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

Vlasic, A., 2012: Long-run analysis of the stochastic replicator dynamics in the presence of random jumps. http://arxiv.org/abs/1206.0344.

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

(M. Bravo) Universidad de Chile, Departamento de Ingeniería Industrial, República 701, Santiago Centro, Chile
E-mail address: mbravo@dii.uchile.cl

(P. 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