Online Appendix to: Convergence in Models with Bounded Expected Relative Hazard Rates

Carlos Oyarzun* Johannes Ruf†

September 6, 2014

In this appendix, we apply the convergence results of Section 2 in Oyarzun and Ruf (2014) to a framework that nests several models of individual and social learning.

B.1 General framework for learning applications

We consider a finite set W of individuals called the *population.* These individuals choose actions in some finite set A that yield some payoff in the set $[0,1]$. Formally, if individual $i \in W$ chooses action $a \in A$ at time $t \in \mathbb{N}$, she obtains a payoff, denoted by $x_t^{(i)}(a)$ or simply $x_t^{(i)}$. We denote individual i's chosen action at time t by $a_t^{(i)}$. Individual i may only observe the payoff she obtained or she may observe both the obtained and forgone payoffs, i.e., the whole profile $\{x_t^{(i)}(a)\}_{a\in A}$. In these cases it is said that the individual has *partial information* or *full information*, respectively. We will consider these two possibilities below when we analyze models of individual learning.

Since we consider applications to social learning as well, individuals may also be allowed to observe the chosen action(s) and obtained payoff(s) of other individual(s) in the population. The set of individuals that $i \in W$ observes at time $t \in \mathbb{N}$ (including herself) is called the *sample* and is denoted by $s_t^{(i)}$. The profile of actions chosen by the individuals in sample s at time $t \in \mathbb{N}$ is denoted by $a_t^{(s)}$ for all (non-empty) $s \in \mathcal{P}(W)$, where $\mathcal{P}(W)$ denotes the set of all subsets of W, and the corresponding profile of payoffs is denoted by $x_t^{(s)}(a_t^{(s)})$, or simply by $x_t^{(s)}$.[1]

We now specify the framework:

(a) Probability space. The set of states that may occur at time $t \in \mathbb{N}$, corresponds to the set $\Omega_t = \big([0,1]^{|A|} \times \mathcal{P}(W)\big)^{|W|} \times [0,1]^{|W|}$. The first two components of the state correspond to the obtained

*School of Economics, University of Queensland. E-Mail: c.oyarzun@uq.edu.au

†Oxford-Man Institute of Quantitative Finance, University of Oxford. E-Mail: johannes.ruf@oxford-man.ox.ac.uk

[1] In principle, the analysis here could allow for more possibilities. For instance, individuals could observe the forgone payoffs of other individuals as well. Our choice of the level of generality of the analysis reflects our attempt to capture the main elements that drive the results.

and forgone payoffs $\{x_t^{(i)}(a)\}_{i\in W, a\in A}$ and observed samples $\{s_t^{(i)}\}_{i\in W}$. The third component $[0,1]^{|W|}$ corresponds to all possible realizations of a randomization device that, as described below, determines the actual choice of each individual, given her probabilistic rule of choice. We denote the corresponding Borel sigma algebra by $\mathcal{F}_t$.

Now, rewrite Ω as the product space of two spaces Ω^E and Ω^R (along with the corresponding sigma algebras); more precisely, write

$$\Omega = \Omega^E \times \Omega^R := \prod_{t=1}^{\infty} \Big([0,1]^{|A|} \times \mathcal{P}(W) \Big)^{|W|} \times \prod_{t=1}^{\infty} [0,1]^{|W|} =: \prod_{t=1}^{\infty} \Omega_t^E \times \prod_{t=1}^{\infty} \Omega_t^R =: \prod_{t=1}^{\infty} \Omega_t,$$

with $\Omega_t^E := \big([0,1]^{|A|} \times \mathcal{P}(W) \big)^{|W|}$, $\Omega_t^R := [0,1]^{|W|}$, and $\Omega_t = \Omega_t^E \times \Omega_t^R$ for all $t \in \mathbb{N}$.

It remains to specify the probability distribution $\mathbb{P}$ over $(\Omega, \otimes_{t=1}^{\infty} \mathcal{F}_t)$. We will always assume that $\mathbb{P}$ is the product measure of two probability measures; to wit, $\mathbb{P} = \mathbb{P}^E \times \mathbb{P}^R$, with $\mathbb{P}^E$ and $\mathbb{P}^R$ defined over the respective spaces. In particular, the two components corresponding to each state of the world, are independent.

The first component $\mathbb{P}^E$ of $\mathbb{P}$ specifies the environment in which individuals live, such as the distribution of the payoffs. Individuals are not assumed to know $\mathbb{P}^E$. In particular, they do not know the distribution over the payoff profile. We also assume that the distribution of $x_t^{(i)}(a)$ does not depend on $i \in W$ for all $a \in A$ and and $t \in \mathbb{N}$. We denote by $x_t^{(s)} = x_t^{(s)}(a^{(s)}, \omega)$ the profile of payoffs obtained at time $t \in \mathbb{N}$ by a sample of individuals s if they choose the actions $a^{(s)}$ and the state of the world is $\omega \in \Omega$.

The second component $\mathbb{P}^R$ of $\mathbb{P}$ is a product measure (corresponding to different times) of product measures (corresponding to different individuals) of $|W|$ uniformly distributed random variables over $[0,1]$. The realizations of these random devices determine individuals' choices. Towards this end, we define the configuration space $\mathfrak{S}$ as the set that contains all possible profiles of vectors of probabilities of choosing each action, i.e., $\mathfrak{S} = (\Delta(A))^{|W|}$. That is, the component $\sigma_t^{(i)}(a)$ of $\sigma_t = (\sigma_t^{(1)}, \ldots, \sigma_t^{(|W|)}) \in \mathfrak{S}$ represents the probability that individual $i \in W$ chooses action $a \in A$ at time $t+1$ for all $t \in \mathbb{N}_0$. Formally, each individual i partitions the interval $[0,1]$ in $|A|$ subintervals assigned to each action and whose measures are the elements of $\sigma_t^{(i)}$. The realization of the component of the randomization device corresponding to individual i determines that action a is chosen if that realization is contained in the interval of action a.

(b) Behavioral rule. The initial probabilities of choosing each action, described by the configuration $\sigma_0 \in \mathfrak{S}$, are exogenously given. Upon observing the new information at time $t \in \mathbb{N}$, e.g., obtained payoffs or other individuals' choices and payoffs, the probabilities of choosing each action are updated according to a function called the *behavioral rule*, denoted by $L_{t-1}^{(i)}$. Hence, $\sigma_t^{(i)} = L_{t-1}^{(i)}(\cdot)$ for all $i \in W$ and $t \in \mathbb{N}$, which defines a $\mathfrak{S}$–valued process $\{\sigma_t\}_{t\in\mathbb{N}_0}$. The specification of the behavioral rule $L_{t-1}^{(i)}$ varies across the different learning models that we consider, as the information on actions and payoffs observed by the individual at time t depends on whether we are studying a setup where forgone payoffs are observed

(along with the obtained payoff), and whether we analyze individual or social learning. In all applications, however, the "shape" of $L_{t-1}^{(i)}$ is assumed to be determined by the information available to individual i up to time $t-1$. Thus, at time $t-1$, we already know how each possible realization of the random variables that will be observed at time t will be mapped to updated probabilities of choosing each action.

(c) Aggregator and optimality. Given an optimality criterion O to determine the set of optimal actions $A^{*,O}$, we are interested in the average probability of choosing an optimal action implied by any given configuration $\sigma \in \mathfrak{S}$. Towards this end, we introduce an aggregator $\mathfrak{A}^O$ that maps $\mathfrak{S}$ to $[0,1]$ defined by

$$\mathfrak{A}^O(\sigma) = \frac{1}{|W|} \sum_{i \in W} \sum_{a \in A^{*,O}} \sigma^{(i)}(a) \tag{B.1}$$

for all $\sigma \in \mathfrak{S}$. To avoid trivial cases, we shall assume that both $A^{*,O}$ and $A \setminus A^{*,O}$ are non-empty.

Given $\omega \in \Omega$, the components of ω in Ω_t^R in conjunction with $\sigma_{t-1}(\omega) \in \mathfrak{S}$ determine the profile of choices $\{a_t^{(i)}\}_{i \in W} = \{a_t^{(i)}(\omega, \sigma_{t-1}^{(i)}(\omega))\}_{i \in W}$ at each time $t \in \mathbb{N}$. Convergence to optimality of a learning model, with respect to the criterion O, is defined as the event that all individuals in the population converge to choose an optimal action with probability one, i.e., the event $\{\lim_{t \uparrow \infty} \mathfrak{A}^O(\sigma_t) = 1\}$.

While many criteria could be considered, in the applications below we focus on expected value ($O = E$) and first-order stochastic dominance ($O = S$). Accordingly, we define

$$A^{*,E} := \left\{ a \in A : \mathbb{E}_t \left[x_{t+1}^{(1)}(a) \right] \geq \mathbb{E}_t \left[x_{t+1}^{(1)}(b) \right] \text{ almost surely for all } b \in A \text{ and } t \in \mathbb{N}_0 \right\}$$

and

$$A^{*,S} := \left\{ a \in A : \mathbb{E}_t \left[u \left(x_{t+1}^{(1)}(a) \right) \right] \geq \mathbb{E}_t \left[u \left(x_{t+1}^{(1)}(b) \right) \right] \text{ almost surely for all } b \in A, t \in \mathbb{N}_0 \text{ and } u \in \mathcal{U} \right\},$$

where $\mathcal{U}$ denotes the set of all bounded, non-decreasing functions $u : [0,1] \to \mathbb{R}$. Note that both sets $A^{*,E}$ and $A^{*,S}$ are determined by the *environment*, i.e., the underlying probability measure $\mathbb{P}$.

B.2 Individual learning

In this section, we focus on individual learning, thus, $|W| = 1$. At each time $t \in \mathbb{N}$, the individual observes the payoff she would have obtained from any action if she had chosen such an action or some part of this information. The behavioral rule, denoted by L_{t-1},[2] is a function mapping the current probabilities of choosing each action at time $t \in \mathbb{N}$ and the observed part of the vector of actions and corresponding payoffs to the probability of choosing each action at time $t+1$.

[2]For the analysis of individual learning we omit the superscript (i).

First we consider the case of *partial information* (P), i.e., the individual only observes the payoff of the action she chose. We focus on the class of *monotone behavioral rules* $L^P = \{L_t^P\}_{t\in\mathbb{N}_0}$, where

$$L_t^P : A \times [0,1] \times \Delta(A) \to \Delta(A)$$

is given by

$$L_t^P(a,x,\sigma)_a = \sigma(a) + (1-\sigma(a))\,(C_{t,a,a} + D_{t,a,a}x) \quad \text{and} \tag{B.2}$$

$$L_t^P(b,x,\sigma)_a = \sigma(a) - \sigma(a)(C_{t,b,a} + D_{t,b,a}x) \tag{B.3}$$

for all $t \in \mathbb{N}_0$, $a \in A$, $b \in A \setminus \{a\}$, $x \in [0,1]$, and $\sigma \in \Delta(A)$. Here, $\{C_{t,b,a}\}_{t\in\mathbb{N}_0,a,b\in A}$ and $\{D_{t,b,a}\}_{t\in\mathbb{N}_0,a,b\in A}$ are two sequences of deterministic matrixes satisfying

$$C_{t,a,a} = \sum_{c\in A}\sigma(c)C_{t,c,a}, \quad D_{t,a,a} = \sum_{c\in A}\sigma(c)D_{t,c,a}, \quad C_{t,b,a} \geq 0, \quad D_{t,b,a} > 0, \quad \text{and} \quad C_{t,b,a} + D_{t,b,a} \leq 1$$

for all $a, b \in A$ and $t \in \mathbb{N}_0$. Börgers et al. (2004) prove, in a one-period setup, that these behavioral rules yield, in expectation, an increase in the probability of choosing an expected payoff maximizing action.[3]

The analysis also covers the case of *full information* (F), i.e., the individual observes the payoff that she would have obtained with each action. We focus on *pairwise symmetric-switch behavioral rules* $L^F = \{L_t^F\}_{t\in\mathbb{N}_0}$, defined as

$$L_t^F : [0,1]^{|A|} \times \Delta(A) \to \Delta(A),$$

with

$$L_t^F\left((x(c))_{c\in A}, \sigma\right)_a = \sigma(a) + \sigma(a)\sum_{b\in A}\sigma(b)g_{t,b,a}(x(b),x(a)) \tag{B.4}$$

for all $t \in \mathbb{N}_0$, $a \in A$, $x \in [0,1]^{|A|}$, and $\sigma \in \Delta(A)$; here, $\{g_{t,b,a} : [0,1]^2 \to [-1,1]\}_{t\in\mathbb{N}_0,b,a\in A}$ is a family of *symmetric-switch* functions:

Definition B.1. *We say that a function $g : [0,1]^2 \to [-1,1]$ is* symmetric-switch *if*

1. *$g(x_1,x_2) = -g(x_2,x_1)$ for all $x_1, x_2 \in [0,1]$,*
2. *$g(x_1,\cdot)$ is non-decreasing for all $x_1 \in [0,1]$, and*
3. *$g(x_1,x_2) \geq 0$ for all $x_1, x_2 \in [0,1]$ with $x_2 > x_1$.*

[3] Corollaries B.1 and B.2 below would also hold in a more general setup. First, the strict positivity of $\{D_{t,b,a}\}_{t\in\mathbb{N}_0,b,a\in A}$ is more than required (see, e.g., Börgers et al. (2004)). Second, $\{C_{t,b,a}\}_{t\in\mathbb{N}_0,b,a\in A}$ and $\{D_{t,b,a}\}_{t\in\mathbb{N}_0,b,a\in A}$ might depend on the configuration $\sigma \in \mathfrak{S}$; in particular, the lower bound on $\{C_{t,b,a}\}_{t\in\mathbb{N}_0,b,a\in A}$ could be weakened, so that L^P might have a penalizing component, as well. Third, $\{C_{t,b,a}\}_{t\in\mathbb{N}_0,b,a\in A}$ and $\{D_{t,b,a}\}_{t\in\mathbb{N}_0,b,a\in A}$ do not need to be deterministic, but might be $\{\mathcal{F}_{[0,t]}\}_{t\in\mathbb{N}_0}$–adapted.

Two simple examples of symmetric-switch functions are the functions $g_1(x_1, x_2) = \mathbf{1}_{x_2>x_1} - \mathbf{1}_{x_1>x_2}$ and $g_2(x_1, x_2) = x_2^p - x_1^p$ for some $p \geq 0$.

We say that g is *linear in the difference* if $g(x_1, x_2) = c(x_2 - x_1)$ for some scalar $c \geq 0$ for all $x_1, x_2 \in [0, 1]$. We say that L^F is *linear in the differences* if $g_{t,b,a}$ is linear in the difference for some scalar $c_{t,b,a} \geq 0$, for all $t \in \mathbb{N}_0$, $a \in A$, and $b \in A \setminus \{a\}$.

Pairwise symmetric-switch behavioral rules can be interpreted as if the individual was making pairwise comparisons between all possible pairs of actions and moving probability from one action to the other. The use of symmetric-switch functions guarantees that no probability is swapped when the individual is comparing two actions that yielded the same payoff. Furthermore, when the payoff of one action is greater than the payoff of another action, probability is moved toward the action that yielded the higher payoff in the corresponding pairwise comparison.

As shown below, for any optimality criterion O, the expected relative hazard rates of the probability of choosing an optimal action, in the partial and full information setup, are bounded from below by

$$\delta_t^{P,O} := \min_{a \in A^{*,O}, b \in A \setminus A^{*,O}} \{D_{t,b,a}(\mathbb{E}_t[x_{t+1}(a)] - \mathbb{E}_t[x_{t+1}(b)])\} \quad \text{and} \tag{B.5}$$

$$\delta_t^{F,O} := \min_{a \in A^{*,O}, b \in A \setminus A^{*,O}} \{\mathbb{E}_t[g_{t,b,a}(x_{t+1}(b), x_{t+1}(a))]\}, \tag{B.6}$$

respectively, for all $t \in \mathbb{N}_0$.

Regardless the optimality criterion O is expected value or first-order stochastic dominance, we have $\delta_t^{P,O} \geq 0$ for all $t \in \mathbb{N}_0$. Similarly, we have $\delta_t^{F,S} \geq 0$; to see this, fix $a \in A^{*,S}$, $b \in A \setminus \{a\}$, $t \in \mathbb{N}_0$, and the symmetric-switch function $g(\cdot, \cdot) \equiv g_{t,b,a}(\cdot, \cdot)$. By the rule of iterated expectations,

$$\begin{aligned}\mathbb{E}_t[g(x_{t+1}(b), x_{t+1}(a))] &= \mathbb{E}_t[\mathbb{E}[g(x_{t+1}(b), x_{t+1}(a))|\mathcal{F}_{[0,t]} \vee \boldsymbol{\sigma}(x_{t+1}(b))]] \\ &\geq \mathbb{E}_t[\mathbb{E}[g(x_{t+1}(b), \widetilde{x}_{t+1}(b))|\mathcal{F}_{[0,t]} \vee \boldsymbol{\sigma}(x_{t+1}(b))]] = 0,\end{aligned}$$

where $\boldsymbol{\sigma}(x_{t+1}(b))$ is the sigma algebra generated by $x_{t+1}(b)$ and $\widetilde{x}_{t+1}(b)$ is an independent copy of $x_{t+1}(b)$. The inequality follows from the fact that g is non-decreasing in the second component and that $x_{t+1}(a)$ first-order stochastically dominates $\widetilde{x}_{t+1}(b)$. The last equality follows from the anti-symmetry property of g. This yields $\delta_t^{F,S} \geq 0$.[4] Finally, the relative hazard rates in the full information setup with the expected value criterion $\delta_t^{F,E}$ are non-negative if L^F is linear in the differences. Thus, we have proved the following lemma:

Lemma B.1. *Fix an optimality criterion $O \in \{E, S\}$ and an information setup $I \in \{P, F\}$. If $I = F$ and $O = E$, then additionally assume that L^F is linear in the differences. Then, $\delta_t^{I,O} \geq 0$ for all $t \in \mathbb{N}_0$.*

[4] Lemma 3 in Oyarzun and Ruf (2009) states that, for fixed $a, b \in A$ and $t \in \mathbb{N}_0$, if g is symmetric-switch with $g(x_1, x_2) > 0$ for all $x_1, x_2 \in [0, 1]$ such that $x_2 > x_1$, and $x_{t+1}(a)$ strictly first-order stochastically dominates $x_{t+1}(b)$, then the strict inequality $\mathbb{E}_t[g(x_{t+1}(b), x_{t+1}(a))] > 0$ holds. This result can be helpful to show that $\delta_t^{F,S} > 0$ in specific applications.

Now we are ready to provide statements on convergence to optimality of individual learning:

Corollary B.1. *Fix an optimality criterion $O \in \{E, S\}$ and an information setup $I \in \{P, F\}$. If $O = E$ and $I = F$, then additionally assume that L^F is linear in the differences. Suppose that*

$$\sum_{t=0}^{\infty} (\delta_t^{I,O})^2 = \infty \quad \textit{and} \quad \mathfrak{A}^O(\sigma_0) > 0.$$

For an arbitrary sequence $\theta = \{\theta_t\}_{t \in \mathbb{N}_0}$ with $\theta_t \in [0, 1]$ consider the behavioral rule $L^{P,\theta}$ defined as in (B.2) *and* (B.3) *with*

$$C^{\theta}_{t,b,a} := \theta_t C_{t,b,a} \quad \textit{and} \quad D^{\theta}_{t,b,a} := \theta_t D_{t,b,a}$$

and $L^{F,\theta}$ defined as in (B.4) *with*

$$g^{\theta}_{t,b,a} := \theta_t g_{t,b,a}$$

for all $t \in \mathbb{N}_0$, $a \in A$, and $b \in A \setminus \{a\}$. Then, for all $\varepsilon > 0$, there exists a sequence θ such that convergence to optimality holds with probability at least $1 - \varepsilon$ for the behavioral rule $L^{P,\theta}$ or $L^{F,\theta}$, respectively.

The condition on the non square-summability of the relative hazard rates in Corollary B.1 is not stated explicitly in terms of the primitives of the learning model. It is transparent, however, that this condition holds in many cases. For instance, it holds if we assume that the payoffs are identical and independently distributed over time, and the sequences $\{D_{t,b,a}\}_{t \in \mathbb{N}_0, b, a \in A}$ and $\{g_{t,b,a}\}_{t \in \mathbb{N}_0, b, a \in A}$ are time-invariant, with $g_{t,b,a}$ strictly increasing in its second argument.

The behavioral rule L^P with parameter constellation $C_{t,b,a} \equiv 0$ and $D_{t,b,a} \equiv 1$ for all $t \in \mathbb{N}_0$ and $a, b \in A$ corresponds to the standard Cross (1973) model. Börgers and Sarin (1997) show that in the Cross (1973) model with a constant but small step-size, the probabilities of choosing each action converge to those in Taylor (1979) replicator dynamics as the step-size goes to zero. Van Huyck et al. (2007) study this model with small step-size (e.g., $\theta_t = 0.05$ or $\theta_t = 0.01$ for all $t \in \mathbb{N}_0$) to replicate experimental data from a coordination game. In this model, the small step-size may be interpreted as "cautious" learning in the sense that a single observation of a payoff cannot lead to dramatic changes in the probability of choosing each action.

The following result provides conditions for achieving convergence to optimality almost surely in models of individual learning.

Corollary B.2. *Fix an optimality criterion $O \in \{E, S\}$ and an information setup $I \in \{P, F\}$. Suppose payoffs are independent and identically distributed over time, and $\mathfrak{A}^O(\sigma_0) > 0$. For any $c \in (0, 1]$, consider the behavioral rules $\widetilde{L}^P$, defined as in* (B.2) *and* (B.3) *with*

$$\widetilde{C}_{t,b,a} := \frac{1 - c}{t + 2} \quad \textit{and} \quad \widetilde{D}_{t,b,a} := \frac{c}{t + 2},$$

and $\widetilde{L}^F$*, defined as in* (B.4) *with*

$$\widetilde{g}_{t,b,a}(x_1, x_2) := \frac{c(x_2^p - x_1^p)}{t+2},$$

such that $p = 1$ *if* $O = E$ *and* $p > 0$ *otherwise, for all* $t \in \mathbb{N}_0$*,* $a, b \in A$*, and* $x_1, x_2 \in [0,1]$*. Then, convergence to optimality holds almost surely for the behavioral rules* $\widetilde{L}^P$ *and* $\widetilde{L}^F$*.*[5]

The model with partial information in Corollary B.2 is a specific version of our multi-period extension of Börgers et al. (2004). In this model, c is a parameter that determines the marginal effect of the obtained payoff on the updated probability of choosing each action. This behavioral rule does not only induce that the probability of choosing an optimal action is a submartingale, as implied by the results in Börgers et al. (2004), but also yields convergence to choosing optimal actions with probability one. The individual does not need to know the payoff distributions, but can be certain to choose an expected payoff maximizing action in the long-run, as long as her behavior is determined by a behavioral rule with step-size that shrinks arbitrarily over time. The shrinking step-size behavioral rule $\widetilde{L}^P$ is considered, for instance, by Sarin and Vahid (2004) in the analysis of experimental subjects' behavior in games. The decreasing step-size allows this learning model to exhibit the "power law of practice" (see, e.g., Erev and Roth (1998)), and Corollary B.2 proves convergence to optimality for this model.

The proof of Corollaries B.1 and B.2 requires defining the updating rules of the general framework corresponding to the behavioral rules of the learning models:

$$\Pi_t^P(\omega, \sigma) = L_{t-1}^P\left(a_t(\omega, \sigma), x_t(a_t(\omega, \sigma), \omega), \sigma\right); \tag{B.7}$$

$$\Pi_t^F(\omega, \sigma) = L_{t-1}^F\left(\left(x_t(a, \omega)\right)_{a \in A}, \sigma\right) \tag{B.8}$$

for all $t \in \mathbb{N}$, $\omega \in \Omega$, and $\sigma \in \mathfrak{S}$. We recall the generalized framework of updating rules in Remark 2.1 in Subsection 2.4. Theorem 2.1 and Corollary 2.1, in conjunction with the following lemma, then directly yield Corollaries B.1 and B.2.

Lemma B.2. *Fix an optimality criterion* $O \in \{E, S\}$ *and an information setup* $I \in \{P, F\}$*. If* $O = E$ *and* $I = F$*, then additionally assume that* L^F *is linear in the differences. The system* $(\Pi^I, \mathfrak{A}^O)$ *satisfies WBERHR with lower bound sequences given in* (B.5) *for* $I = P$ *and in* (B.6) *for* $I = F$*.*

Proof. First consider the case $I = P$. As Börgers et al. (2004) show,

$$\mathbb{E}_t[\sigma_{t+1}(a)] - \sigma_t(a) = \sigma_t(a) \sum_{b \in A} \sigma_t(b) D_{t,b,a} \left(\mathbb{E}_t\left[x_{t+1}(a)\right] - \mathbb{E}_t\left[x_{t+1}(b)\right]\right) \tag{B.9}$$

[5]Stationarity or independence of the actions' payoffs over time are not essential for the convergence result. If we drop this assumption and instead impose any condition that yields

$$\inf_{t \in \mathbb{N}_0} \min_{a \in A^{*,O}, b \in A \setminus A^{*,O}} \{\mathbb{E}_t[x_{t+1}(a)^p - x_{t+1}(b)^p]\} > 0$$

with the corresponding p, the result holds as well.

for all $a \in A$ and $t \in \mathbb{N}_0$. An application of Lemma B.1 and (B.9) yield the statement for $I = P$.

Next consider the case $I = F$ and observe that $\mathbb{E}_t[g_{t,b,a}(x_{t+1}(a), x_{t+1}(b), \sigma)] = 0$ if $a, b \in A^{*,O}$ since then $x_{t+1}(a)$ and $x_{t+1}(b)$ have either the same expectation (if $O = E$) or have the same distribution (if $O = S$). Note that

$$\mathbb{E}_t\left[\mathfrak{A}^O\left(\Pi^F_{t+1}(\sigma)\right)\right] - \mathfrak{A}^O(\sigma) = \sum_{a \in A^{*,O}} \sigma(a) \sum_{b \in A \setminus A^{*,O}} \sigma(b) \mathbb{E}_t\left[g_{t,b,a}(x_{t+1}(b), x_{t+1}(a))\right] \tag{B.10}$$

for all $\sigma \in \Delta(A)$ and $t \in \mathbb{N}_0$. As above, an application of Lemma B.1 and (B.10) yield the statement for $I = F$. □

B.3 Social learning

In the model analyzed in this section, individuals can observe other individuals' choices and obtained payoffs, and learn by imitation. Learning is described by a behavioral rule $\{L^{(i)}_t\}_{t \in \mathbb{N}_0}$ that has two components. The first, called the *imitation component*, describes how the observed choices and payoffs affect individuals' behavior. This component represents individuals' drive to imitate what others do. The second, called the *inertial component*, does not depend on the current observations and corresponds to the probabilities of choosing each action in the previous period.

At time $t = 1$, each individual $i \in W$ chooses an action according to some exogenously given probabilities $\sigma^{(i)}_0 \in \Delta(A)$. In the following periods individuals can imitate other individuals. At time $t + 1$, individuals make choices through imitation with probability $\lambda_{t-1} \in [0, 1]$ and make choices with the same probability as in the last period with probability $1 - \lambda_{t-1}$ for all $t \in \mathbb{N}$. We call $\lambda := \{\lambda_t\}_{t \in \mathbb{N}_0}$ the *imitation rate* and we assume λ_t is $\mathcal{F}_{[0,t]}$–measurable and the same for all individuals.

We first specify the imitation component $\widehat{L}^{(i)}$ of individual i's behavioral rule, a function defined on the *observation space*

$$\mathcal{O}^{(i)} := \bigcup_{k=1}^{|W|} \left({}^kW_i \times [0, 1]^k\right)$$

for all $i \in W$, where ${}^kW_i$ is the subset of $\mathcal{P}(W)$ whose elements contain individual i and $k - 1$ other individuals, for $k \in \{1, ..., |W|\}$. Therefore, $\mathcal{O}^{(i)}$ represents the possible new information individual i may receive at any time.[6] The imitation component maps what is observed to the probability of imitating any of the observed individuals, i.e., $\widehat{L}^{(i)} : \mathcal{O}^{(i)} \to \Delta(W)$. Thus, the probability that $i \in W$ imitates at time $t + 1$ what $j \in W$ did at time $t \in \mathbb{N}$ is $\lambda_{t-1}\widehat{L}^{(i)}(\cdot, \cdot)_j$. We say that $\widehat{L} := \{\widehat{L}^{(i)}\}_{i \in W}$ satisfies the "must–see" condition if $\sum_{j \in s} \widehat{L}^{(i)}(s, \cdot)_j = 1$ for all $i \in W$ and $s \in \bigcup_{k=1}^{|W|} {}^kW_i$. This condition is standard

[6] The imitation component $\widehat{L}^{(i)}$ and the observation space $\mathcal{O}^{(i)}$ could be generalized to depend on, for instance, past observations, the observed actions, time, or the current probabilities of choosing each action.

in the literature (see, e.g., Cubitt and Sugden (1998)); it states that individuals may only imitate the individuals they observe.

The behavioral rule $L_{t-1}^{(i)}$ of individual $i \in W$, representing her probability of choosing each action at time $t+1$, is defined in a two-step procedure. First, the range $\Delta(W)$ of the imitation component $\widehat{L}^{(i)}$ is mapped to $\Delta(A)$ by assigning, to each action $a \in A$, the sum of the probabilities of imitating each individual who chose a at time t. Second, the inertial behavior is accounted for. Thus, for each individual $i \in W$, the behavioral rule $L_t^{(i)} : \bigcup_{k=1}^{|W|} \left({}^kW_i \times [0,1]^k \times A^k\right) \times \Delta(A) \to \Delta(A)$ is given by

$$L_t^{(i)}\left(s, x^{(s)}, a^{(s)}, \sigma^{(i)}\right)_a = \lambda_t \sum_{j \in s : a^{(j)} = a} \widehat{L}^{(i)}(s, x^{(s)})_j + (1-\lambda_t)\sigma^{(i)}(a) \tag{B.11}$$

for all $t \in \mathbb{N}_0$, $a \in A$, $(s, x^{(s)}, a^{(s)}) \in \bigcup_{k=1}^{|W|} \left({}^kW_i \times [0,1]^k \times A^k\right)$, and $\sigma^{(i)} \in \Delta(A)$.

The analysis allows for several possibilities regarding sampling. Let $\rho^{(i)}(s)$ be a constant corresponding to the probability that individual i observes sample s for all $i \in W$ and $s \in \mathcal{P}(W)$, and such that $\rho^{(i)}(s) = 0$ for all $s \subseteq W \setminus \{i\}$. For all $i \in W$ and $j \in W \setminus \{i\}$, define $\bar{S}(i,j) := \{s \in \mathcal{P}(W) : i, j \in s\}$, i.e., the set of all samples that contain i and j. There are two conditions on sampling that seem hard to dispense with: (i) We say that a sampling process is *symmetric* if $\rho^{(i)}(s) = \rho^{(j)}(s)$ for all $s \in \bar{S}(i,j)$, $i \in W$, and $j \in W \setminus \{i\}$. Thus, symmetric sampling imposes that the probabilities of i and j observing any given sample s with $i, j \in s$ are equal. (ii) We say that a sampling process satisfies *observability* (with lower bound ξ) if the probability that any individual observes any other individual is positive, to wit, if there exists a constant $\xi > 0$ such that $\sum_{s \in \bar{S}(i,j)} \rho^{(i)}(s) > \xi$ for all $i \in W$ and $j \in W \setminus \{i\}$. In the sequel, we shall assume that sampling is symmetric and observable (with lower bound $\xi > 0$) and $\widehat{L}$ satisfies the must-see condition.[7]

Let

$$g_{j,i,s}(x^{(s)}) := \widehat{L}^{(j)}\left(s, x^{(s)}\right)_i - \widehat{L}^{(i)}\left(s, x^{(s)}\right)_j$$

for all $i \in W$, $j \in W \setminus \{i\}$, $s \in \bar{S}(i,j)$, and $x^{(s)} \in [0,1]^{|s|}$. We also define the sequence $\delta^O = \{\delta_t^O\}_{t \in \mathbb{N}_0}$ by

$$\delta_t^O = \lambda_t(|W|-1)\xi \min_{i \in W, j \in W \setminus \{i\}} \min_{s \in \bar{S}(i,j)} \min_{a^{(s)} \in A^{|s|} : a^{(i)} \in A^{*,O}, a^{(j)} \in A \setminus A^{*,O}} \left\{\mathbb{E}_t\left[g_{j,i,s}(x_{t+1}^{(s)}(a^{(s)}))\right]\right\} \tag{B.12}$$

for all $t \in \mathbb{N}_0$ and $O \in \{E, S\}$.

These definitions enable us to provide a first convergence result:

Corollary B.3. *Fix an optimality criterion $O \in \{E, S\}$ and suppose that $s_{t+1}^{(j)}$ is conditionally independent, given the information up to time t, of $\{x_{t+1}^{(i)}\}_{i \in W}$ for all $j \in W$ and $t \in \mathbb{N}_0$. Assume, moreover, that $g_{j,i,s}(x^{(s \setminus \{i,j\})}, \cdot, \cdot)$ is symmetric-switch (non-decreasing in $x^{(i)}$) for all $i \in W$, $j \in W \setminus \{i\}$, $s \in \bar{S}(i,j)$, and $x^{(s)} \in [0,1]^{|s|}$.*

[7] The sampling process could also be assumed to be random; for instance all our results below will hold if we assume that the probabilty that i observes s at time $t+1$ is $\mathcal{F}_{[0,t]}-$measurable for all $i \in W$, $s \in \bigcup_{k=1}^{|W|} {}^kW_i$, and $t \in \mathbb{N}_0$.

(1) If $O = E$, assume that $g_{j,i,s}(x^{(s\setminus\{i,j\})}, \cdot, \cdot)$ is linear in the difference for all $i \in W$, $j \in W \setminus \{i\}$, $s \in \bar{S}(i,j)$, and $x^{(s)} \in [0,1]^{|s|}$.

(2) If $O = S$, assume that, given the information up to time t, $\{x_{t+1}^{(i)}\}_{i\in W}$ are independent for all $t \in \mathbb{N}_0$. Alternatively, suppose that $g_{j,i,s}$ can be written as a sum of functions that only depend on pairs of payoffs $\left\{(x^{(i)}, x^{(j)})\right\}_{i,j\in s}$ for all $i \in W$, $j \in W \setminus \{i\}$, and $s \in \bar{S}(i,j)$, in which case only pairwise independence of $\{x_{t+1}^{(i)}\}_{i\in W}$ is required.

If

$$\sum_{t=0}^{\infty} (\delta_t^O)^2 = \infty \quad \textit{and} \quad \mathfrak{A}^O(\sigma_0) > 0,$$

then, for all $\varepsilon > 0$, there exists a sequence $\theta = \{\theta_t\}_{t\in\mathbb{N}_0}$ with $\theta_t \in [0,1]$ such that convergence to optimality holds with probability at least $1-\varepsilon$ for the behavioral rule in (B.11) *with λ_t replaced by $\theta_t\lambda_t$ for all $t \in \mathbb{N}_0$.*

An example of a behavioral rule that satisfies either Condition (1) or (2) in Corollary B.3 (along with the standing assumptions of this section) is given by

$$\widehat{L}^{(i)}\left(s, x^{(s)}\right)_j = \frac{x^{(j)}}{|s|} \text{ and } \widehat{L}^{(i)}\left(s, x^{(s)}\right)_i = 1 - \frac{\sum_{k\in s\setminus\{i\}} x^{(k)}}{|s|} \tag{B.13}$$

for all $i \in W$, $j \in W \setminus \{i\}$, $s \in \bar{S}(i,j)$, and $x^{(s)} \in [0,1]^{|s|}$. Another example that satisfies the conditions in Corollary B.3 but does not impose linear dependence on the observed payoffs is given by

$$\widehat{L}^{(i)}\left(s, x^{(s)}\right)_j = \frac{f\left(x^{(j)}\right)}{\sum_{k\in s} f\left(x^{(k)}\right)} \text{ with } \frac{0}{0} := \frac{1}{|s|} \tag{B.14}$$

for all $i, j \in W$, $s \in \bar{S}(i,j)$, and $x^{(s)} \in [0,1]^{|s|}$, where $f : [0,1] \to [0,\infty)$ is any non-negative and non-decreasing function; e.g., $f(x) = x$. The specification of such an imitation component resembles that of the Roth-Erev model of individual learning. In that model, the probability of choosing each action is proportional to the cumulative reinforcement, determined by the payoffs the individual has received over time with each action. Here, the probability of imitating each other sampled individual is proportional to the payoff that such an individual received and hence, the probability of choosing the corresponding action through imitation is proportional to the sum of payoffs it provided to the sampled individuals who have chosen this action.

These examples allow us to provide a second convergence result:

Corollary B.4. *Make the same assumptions as in Corollary B.3. Furthermore, suppose that $\mathfrak{A}^O(\sigma_0) > 0$ and*

$$\inf_{t\in\mathbb{N}_0} \min_{a\in A^{*,O}, b\in A\setminus A^{*,O}} \{\mathbb{E}_t\left[f(x_{t+1}(a)) - f(x_{t+1}(b))\right]\} > 0 \tag{B.15}$$

for either $f(x) = x$, in the case $O = E$, or any nonnegative, non-decreasing and bounded function f on $[0,1]$, in the case $O = S$. Then, with $\lambda_t = 1/(t+2)$ for all $t \in \mathbb{N}_0$, convergence to optimality holds almost

surely for the behavioral rule defined by (B.11) *in combination with* (B.13) *if* $O = E$ *or with* (B.14) *if* $O = S$.

In contrast with previous convergence results in the literature that rely on populations that are a continuum, Corollary B.4 reveals that behavioral rules with an imitation component whose magnitude decreases over time, as described in this corollary, yield convergence to optimality almost surely even for finite populations. The condition in (B.15) is satisfied if the payoffs are independent and identically distributed over time, either trivially if $f(x) = x$, or due to Lemma 3 in Oyarzun and Ruf (2009) if f is additionally assumed to be strictly increasing (see Footnote 4).

When we impose that each individual only observes one other individual, i.e., $\sum_{j \in W \setminus \{i\}} \rho^{(i)}(\{i,j\}) = 1$ for all $i \in W$, the imitation component of the behavioral rules in this section correspond to the first-order monotone behavioral rules in Oyarzun and Ruf (2009). When we further assume that the symmetric-switch functions are linear, the class of imitation components of the behavioral rules contains Schlag (1998) *improving* behavioral rules. When we impose that each individual observes two other individuals, i.e., $\sum_{j \in W \setminus \{i\}, k \in W \setminus \{i,j\}} \rho^{(i)}(\{i,j,k\}) = 1$ for all $i \in W$, and the symmetric-switching functions are linear, the imitation component satisfies the characterization of *strictly improving rules* in Schlag (1999). We are not aware of any paper in the literature that provides implications for the relative hazard rates for the other possibilities of sampling that we allow.

Models exhibiting small step-size as those in Corollary B.3 can be interpreted as exhibiting "cautious" social learning, analogously to the models of individual learning analyzed in Corollary B.1. Similarly, the decreasing imitation rates in Corollary B.4 allow these models to exhibit the "power law of practice," analogously to the models in Corollary B.2.

Analogously to the arguments in individual learning, we now associate an updating rule to the behavioral rule in order to prove Corollaries B.3 and B.4. Formally, we set

$$\Pi_t(\omega, \sigma) = \left\{ L_{t-1}^{(i)} \left(s_t^{(i)}(\omega), x_t^{(s_t^{(i)}(\omega))} \left(a_t^{(s_t^{(i)}(\omega))}(\omega, \sigma), \omega \right), a_t^{(s_t^{(i)}(\omega))}(\omega, \sigma), \sigma^{(i)} \right) \right\}_{i \in W} \tag{B.16}$$

for all $t \in \mathbb{N}, \omega \in \Omega$, and $\sigma \in \mathfrak{S}$. Now we can apply Theorem 2.1 and Corollary 2.1, along with Remark 2.1. Towards this end, fix $O \in \{E, S\}$ and recall the sequence δ^O defined in (B.12). It is sufficient to show that the conditions in Corollaries B.3 and B.4 yield (a) $\delta_t^O \geq 0$ for all $t \in \mathbb{N}_0$, and (b) the system $(\Pi, \mathfrak{A}^O)$ satisfies (W)BEHR with sequence δ^O. The argument that proves (a) is analogous to the one in Section B.2 establishing that $\delta_t^{F,O} \geq 0$ for all $t \in \mathbb{N}_0$ and is omitted. Lemma B.3 below takes care of (b), which completes the proof of Corollaries B.3 and B.4.

Lemma B.3. *Fix an optimality criterion* $O \in \{E, S\}$ *and suppose that* $s_{t+1}^{(j)}$ *is conditionally independent, given the information up to time* t*, of* $\{x_{t+1}^{(i)}\}_{i \in W}$ *for all* $j \in W$ *and* $t \in \mathbb{N}_0$*. Provided that* $\delta_t^O \geq 0$ *for all*

$t \in \mathbb{N}_0$, *the system* $(\Pi, \mathfrak{A}^O)$, *defined by* (B.1), (B.11), *and* (B.16), *satisfies WBERHR with lower bound sequence* δ^O.

The proof of this lemma is provided in Section B.5.

Thus, symmetric and observable sampling yield positive relative hazard rates. Symmetric sampling imposes that for every pair of individuals, they are equally visible to each other in each sample that contains both of them. Symmetric switch yields that, in expected value, an individual i observing other individual j doing better to be more likely to switch to j's action than j to switch to i's action. As a result, in expected value, the average probability of choosing optimal actions in the population increases over time.

B.4 Roth-Erev learning model

In the previous sections we have used Theorem 2.1 and Corollary 2.1 to provide sufficient conditions for achieving convergence to optimality either with high probability or almost surely. In this section, we show that the almost-sure convergence to optimality of Roth and Erev's behavioral rule can be derived from the argument developed in Theorem 2.2. This learning model is widely used to describe individual learning in experimental economics (see, e.g., Roth and Erev (1995)) and its convergence properties are studied in several places in the literature (see, e.g., Beggs (2005) and Hopkins and Posch (2005), and the references therein).

Roth and Erev's behavioral rule can be interpreted in terms of a vector of current "attractions" corresponding to each action. The probability of choosing each of them is proportional to its attraction that formally is defined by $V_t(a) = V_{t-1}(a) + \mathbf{1}_{\{a_t=a\}} x_t(a)$ for all $t \in \mathbb{N}$ and $V_0(a) > 0$ exogenously given for all $a \in A$. In other words, only the attraction of the chosen action is updated, and it is increased in an amount equal to the obtained payoff.

Let $V_t := \sum_{a \in A} V_t(a)$ for all $t \in \mathbb{N}_0$ be the sum of attractions at time $t \in \mathbb{N}_0$. The behavioral rule of Roth and Erev $L_t : A \times [0,1] \to \Delta(A)$ is then given by

$$L_t(a,x)_a = \frac{V_t(a) + x}{V_t + x} \quad \text{and} \quad L_t(a,x)_b = \frac{V_t(b)}{V_t + x}$$

for all $a \in A$, $b \in A \setminus \{a\}$, $x \in [0,1]$, and $t \in \mathbb{N}_0$.

The set of optimal actions that we consider here is the set of expected payoff maximizing actions. Instead of deriving the system associated to this behavioral rule, here we derive directly the process $P = \{P_t\}_{t \in \mathbb{N}_0}$ defining $P_t = \mathfrak{A}^E(L_{t-1}(a_t, x_t))$ for all $t \in \mathbb{N}$ and $P_0 = \mathfrak{A}^E((V_0(a))_{a \in A}/V_0)$.

Beggs (2005) provides a thorough analysis of the Roth and Erev (1995) model of individual learning with partial information. Here we recover and slightly generalize his convergence result[8] using a different

[8] E.g., Beggs (2005) shows his result under the additional assumption that payoffs are bounded away from zero.

argument and hence, provide a different interpretation of the convergence properties of this learning model. The proof we provide here follows from Theorem 2.2 and can be found in Section B.5. Hence, our argument is based on the analysis of the properties of the expected relative hazard rates of this learning model.

Corollary B.5. *Suppose that there exists an almost surely strictly positive random variable ε such that*

$$\inf_{t\in\mathbb{N}_0} \quad \min_{a\in A^{*,E}, b\in A\setminus A^{*,E}} \quad \{\mathbb{E}_t[x_{t+1}(a) - x_{t+1}(b)]\} \geq \varepsilon. \tag{B.17}$$

If an individual makes choices according to Roth-Erev's behavioral rule, then $P_\infty = 1$ almost surely; that is, the individual will choose, in the limit, almost surely an optimal action.

Apart from (B.17), we have made no assumptions on either stationarity or independence of the actions' payoffs. Finally, we remark that our proof of convergence to optimality is based on Theorem 2.2, so the proof of Corollary B.5 is based only on the analysis of the performance measure, not the underlying system.

B.5 Proofs of the online appendix

We first provide the proof of Lemma B.3:

Proof of Lemma B.3. Let $P_t := \mathfrak{A}^O(\sigma_t)$ and $P_t^{(i)} := \sum_{a\in A^{*,O}} \sigma_t{}^{(i)}$ for all $i \in W$ and $t \in \mathbb{N}_0$. Fix $t \in \mathbb{N}_0$ and observe that

$$\begin{aligned}
\mathbb{E}_t[P_{t+1}] - P_t &= \frac{1}{|W|}\sum_{j\in W}\left(\mathbb{E}_t\left[P_{t+1}^{(j)}\right] - P_t^{(j)}\right) \\
&= \frac{\lambda_t}{|W|}\sum_{i,j\in W}\mathbb{E}_t\left[\widehat{L}^{(j)}\left(s_{t+1}^{(j)}, x_{t+1}^{(s_{t+1}^{(j)})}\right)_i\left(\mathbf{1}_{\{a_{t+1}^{(i)}\in A^{*,O}\}} - \mathbf{1}_{\{a_{t+1}^{(j)}\in A^{*,O}\}}\right)\right] \\
&= \frac{\lambda_t}{|W|}\sum_{i,j\in W}\sum_{s\in\bar{S}(i,j)}\rho^{(j)}(s)\mathbb{E}_t\left[\widehat{L}^{(j)}\left(s, x_{t+1}^{(s)}\right)_i\left(\mathbf{1}_{\{a_{t+1}^{(i)}\in A^{*,O}\}} - \mathbf{1}_{\{a_{t+1}^{(j)}\in A^{*,O}\}}\right)\right]
\end{aligned}$$

since $\sum_{i\in W}\widehat{L}^{(j)}(\cdot,\cdot)_i = 1$, $s_{t+1}^{(j)}$ is assumed to be conditionally independent of $\{a_{t+1}^{(i)}\}_{i\in W}$ and $\{x_{t+1}^{(i)}\}_{i\in W}$ for all $j \in W$, and $\widehat{L}^{(j)}(s,\cdot)_i = 0$ for all $s \notin \bar{S}(i,j)$ and all $i,j \in W$ by assumption. Observe now, by the assumption (on the randomization device) of conditional independence of the choice of actions $\{a_{t+1}^{(i)}\}_{i\in W}$ and the payoffs $\{x_{t+1}^{(i)}\}_{i\in W}$, that

$$\begin{aligned}
&\sum_{i,j\in W}\sum_{s\in\bar{S}(i,j)}\rho^{(j)}(s)\mathbb{E}_t\left[\widehat{L}^{(j)}\left(s, x_{t+1}^{(s)},\right)_i\left(\mathbf{1}_{\{a_{t+1}^{(i)}\in A^{*,O}\}} - \mathbf{1}_{\{a_{t+1}^{(j)}\in A^{*,O}\}}\right)\right] \\
&\quad= \sum_{i,j\in W}\sum_{s\in\bar{S}(i,j)}\rho^{(j)}(s)\sum_{a^{(s)}\in A^{|s|}: a^{(i)}\in A^{*,O}, a^{(j)}\in A\setminus A^{*,O}}\mathbb{P}_t\left(a_{t+1}^{(s)} = a^{(s)}\right)\mathbb{E}_t\left[\widehat{L}^{(j)}\left(s, x_{t+1}^{(s)}\right)_i\right]
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{i,j\in W} \sum_{s\in\bar{S}(i,j)} \rho^{(j)}(s) \sum_{a^{(s)}\in A^{|s|}:a^{(i)}\in A\setminus A^{*,O},a^{(j)}\in A^{*,O}} \mathbb{P}_t\left(a_{t+1}^{(s)} = a^{(s)}\right) \mathbb{E}_t\left[\widehat{L}^{(j)}\left(s, x_{t+1}^{(s)}\right)_i\right] \\
&\geq \widetilde{\delta}_t^O \sum_{i,j\in W} \sum_{s\in\bar{S}(i,j)} \rho^{(j)}(s) \sum_{a^{(s)}\in A^{|s|}:a^{(i)}\in A^{*,O},a^{(j)}\in A\setminus A^{*,O}} \mathbb{P}_t\left(a_{t+1}^{(s)} = a^{(s)}\right),
\end{aligned}
$$

with

$$
\widetilde{\delta}_t^O := \min_{i\in W, j\in W\setminus\{i\}} \min_{s\in\bar{S}(i,j)} \min_{a^{(s)}\in A^{|s|}:a^{(i)}\in A^{*,O},a^{(j)}\in A\setminus A^{*,O}} \left\{\mathbb{E}_t\left[g_{j,i,s}(x_{t+1}^{(s)}(a^{(s)}))\right]\right\} \tag{B.18}
$$

for all $t \in \mathbb{N}_0$ and $O \in \{E, S\}$. In the last step we first exchanged i and j and used the assumption that $\rho^{(j)}(s) = \rho^{(i)}(s)$ for all $s \in \bar{S}(i,j)$. Due to the conditional independence of $\{a_{t+1}^{(i)}\}_{i\in W}$ imposed by the randomization device we have

$$
\begin{aligned}
\sum_{a^{(s)}\in A^{|s|}:a^{(i)}\in A^{*,O},a^{(j)}\in A\setminus A^{*,O}} \mathbb{P}_t\left(a_{t+1}^{(s)} = a^{(s)}\right) &= \sum_{a^{(i)}\in A^{*,O},a^{(j)}\in A\setminus A^{*,O}} \mathbb{P}_t\left(a_{t+1}^{(i)} = a^{(i)}\right) \cdot \mathbb{P}_t\left(a_{t+1}^{(j)} = a^{(j)}\right) \\
&= P_t^{(i)}\left(1 - P_t^{(j)}\right)
\end{aligned}
$$

for all $i \in W$ and $j \in W \setminus \{i\}$. This yields

$$
\begin{aligned}
\sum_{i,j\in W} \sum_{s\in\bar{S}(i,j)} \rho^{(j)}(s) \sum_{a^{(s)}\in A^{|s|}:a^{(i)}\in A^{*,O},a^{(j)}\in A\setminus A^{*,O}} \mathbb{P}_t\left(a_{t+1}^{(s)} = a^{(s)}\right) &= \sum_{i\in W, j\in W\setminus\{i\}} P_t^{(i)}\left(1 - P_t^{(j)}\right) \sum_{s\in\bar{S}(i,j)} \rho^{(j)}(s) \\
&\geq \xi \sum_{i\in W, j\in W\setminus\{i\}} P_t^{(i)}\left(1 - P_t^{(j)}\right) \\
&= \xi \sum_{i\in W} P_t^{(i)}\left(|W| - 1 - \left(|W|P_t - P_t^{(i)}\right)\right) \\
&\geq \xi|W|(|W| - 1)P_t\left(1 - P_t\right),
\end{aligned}
$$

where the first inequality follows from the assumed observability and the second inequality from the fact that $\sum_{i\in W}(P_t^{(i)})^2 \geq |W|P_t^2$, which is implied by Jensen's inequality. The statement then follows directly. □

The following simple observation, which is closely related to Lemma 2 in Beggs (2005), will be useful in the proof of Corollary B.5:

Lemma B.4. *In the setup of Section B.4, if for some $a \in A$, we have $\mathbb{E}_t[x_{t+1}(a)] > \varepsilon$ almost surely for all $t \in \mathbb{N}_0$, for some almost surely strictly positive random variable ε, then $\lim_{t\uparrow\infty} V_t(a) = \infty$.*

Proof. We observe that the probability of choosing action a at time t is bounded from below by $V_0(a)/(V_0 + t - 1)$. Thus, an application of the Borel-Cantelli lemma yields that action a is chosen infinitely often, say at times $\tau_1 < \tau_2 < \ldots$. Set $\tau_0 := 0$ and define the martingale $M = \{M_n\}_{n\in\mathbb{N}_0}$ by $M_n = V_{\tau_n}(a) - \sum_{i=1}^n \mathbb{E}_{\tau_{i-1}}[x_{\tau_i}(a)]$ for all $n \in \mathbb{N}_0$. Then,

$$
\frac{V_{\tau_n}(a)}{n} = \frac{M_n}{n} + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\tau_{i-1}}[x_{\tau_i}(a)] \geq \frac{M_n}{n} + \frac{1}{n}\sum_{i=1}^n \varepsilon \to 0 + \varepsilon = \varepsilon > 0
$$

as $n \uparrow \infty$ due to the Strong Law of Large Numbers for martingales in Chow (1967) applied to the first term. This yields $\lim_{t\uparrow\infty} V_t(a) = \infty$. □

We now provide the proof of Corollary B.5:

Proof of Corollary B.5. We assume, without loss of generality by Lemma B.4, that

$$V_0 > \frac{2}{\varepsilon} \vee 1. \tag{B.19}$$

Let $\sigma_t(a) := V_t(a)/\sum_{c\in A} V_t(c)$ for all $a \in A$ and $t \in \mathbb{N}_0$. Observe that

$$\begin{aligned}
L_t(a,x)_a &= \sigma_t(a) + (1-\sigma_t(a))\,\frac{x}{\sum_{c\in A} V_t(c) + x} = \sigma_t(a) + \sum_{b\in A\setminus\{a\}} \sigma_t(b)\cdot\frac{x}{\sum_{c\in A} V_t(c) + x},\\
L_t(b,x)_a &= \sigma_t(a) - \sigma_t(a)\frac{x}{\sum_{c\in A} V_t(c) + x}
\end{aligned}$$

for all $a \in A$, $b \in A \setminus \{a\}$, $x \in [0,1]$, and $t \in \mathbb{N}_0$. Therefore,

$$\begin{aligned}
\mathbb{E}_t[P_{t+1}] - P_t &= \mathbb{E}_t\left[\sum_{a\in A^{*,E}}\sum_{b\in A} \sigma_t(b) L_t(b, x_{t+1}(b))_a\right] - \sum_{a\in A^{*,E}} \sigma_t(a)\\
&= \sum_{a\in A^{*,E}} \mathbb{E}_t\left[\sigma_t(a)(1-\sigma_t(a))\frac{x_{t+1}(a)}{V_t + x_{t+1}(a)} - \sum_{b\in A\setminus\{a\}} \sigma_t(b)\sigma_t(a)\frac{x_{t+1}(b)}{V_t + x_{t+1}(b)}\right]\\
&= \sum_{a\in A^{*,E}}\sum_{b\in A\setminus A^{*,E}} \sigma_t(a)\sigma_t(b)\cdot\mathbb{E}_t\left[\frac{x_{t+1}(a)}{V_t + x_{t+1}(a)} - \frac{x_{t+1}(b)}{V_t + x_{t+1}(b)}\right] \geq P_t(1-P_t)\delta_t,
\end{aligned}$$

where δ_t is defined as

$$\delta_t := \min_{a\in A^{*,E}, b\in A\setminus A^{*,E}} \{\delta_t(a,b)\} := \min_{a\in A^{*,E}, b\in A\setminus A^{*,E}} \left\{\mathbb{E}_t\left[\frac{x_{t+1}(a)}{V_t+1} - \frac{x_{t+1}(b)}{V_t}\right]\right\}$$

for all $t \in \mathbb{N}_0$. We notice that

$$\begin{aligned}
\delta_t(a,b) &= \frac{1}{V_t+1}\left(\mathbb{E}_t[x_{t+1}(a) - x_{t+1}(b)] - \frac{1}{V_t}\mathbb{E}_t[x_{t+1}(b)]\right)\\
&\geq \frac{1}{V_t+1}\left(\varepsilon - \frac{1}{V_0}\right) \geq \frac{\varepsilon}{2(V_0+t+1)} > 0
\end{aligned}$$

for all $t \in \mathbb{N}_0$, $a \in A^{*,E}$, and $b \in A \setminus A^{*,E}$ due to (B.19). This inequality also yields $\sum_{t=0}^{\infty} \delta_t = \infty$.

We set $\theta_t := 1/V_t \in (0,1)$ for all $t \in \mathbb{N}_0$ and observe that

$$\frac{\delta_t}{\theta_t} \geq \frac{V_t}{V_t+1}\cdot\left(\varepsilon - \frac{1}{V_0}\right) \geq \frac{1}{1+\frac{1}{V_t}}\cdot\frac{\varepsilon}{2} > \tilde{\delta}$$

for some strictly positive random variable $\tilde{\delta}$. Furthermore,

$$-P_t\frac{1}{V_t} \leq \frac{P_tV_t}{V_t+1} - P_t \leq P_{t+1} - P_t \leq \frac{P_tV_t + x_{t+1}(a_{t+1})}{V_t + x_{t+1}(a_{t+1})} - P_t = \frac{x_{t+1}(a_{t+1})(1-P_t)}{V_t + x_{t+1}(a_{t+1})} \leq \frac{1}{V_t}(1-P_t)$$

and thus, (2.5) holds for all $t \in \mathbb{N}_0$. Therefore, Theorem 2.2, in conjunction with Lemma B.4, yields the result. □

References

Beggs, A. (2005). On the convergence of reinforcement learning. *Journal of Economic Theory*, 77:383–405.

Börgers, T., Morales, A., and Sarin, R. (2004). Expedient and monotone learning rules. *Econometrica*, 72:383–405.

Börgers, T. and Sarin, R. (1997). Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77:383–405.

Chow, Y. S. (1967). On a strong law of large numbers for martingales. *Annals of Mathematical Statistics*, 38:610.

Cross, J. (1973). A stochastic learning model of economic behavior. *Quarterly Journal of Economics*, 87:239–266.

Cubitt, R. and Sugden, R. (1998). The selection of preferences by imitation. *Review of Economic Studies*, 65:761–771.

Erev, I. and Roth, A. (1998). Predicting how people play games: Reinforcement learning in experimental games with a unique mixed strategy equilibria. *American Economic Review*, 88:848–881.

Hopkins, E. and Posch, M. (2005). Attainability of boundary points under reinforcement learning. *Games and Economic Behavior*, 53:110–125.

Oyarzun, C. and Ruf, J. (2009). Monotone imitation. *Economic Theory*, 41:411–441.

Oyarzun, C. and Ruf, J. (2014). Convergence in models with bounded expected relative hazard rates. *Journal of Economic Theory*, 154:229–244.

Roth, A. and Erev, I. (1995). Learning in extensive form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8:164–212.

Sarin, R. and Vahid, F. (2004). Strategy similarity and coordination. *Economic Journal*, 114:506–527.

Schlag, K. (1998). Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. *Journal of Economic Theory*, 78:130–156.

Schlag, K. (1999). Which one should I imitate? *Journal of Mathematical Economics*, 31:493–522.

Taylor, P. (1979). Evolutionary stable strategies with two types of players. *Journal of Applied Probability*, 16:76–83.

Van Huyck, J. B., Battalio, R. C., and Rankin, F. W. (2007). Selection dynamics and adaptive behavior without much information. *Economic Theory*, 33:53–65.