Convergence in Models with Bounded Expected Relative Hazard Rates*

Carlos Oyarzun[†] Johannes Ruf[‡]

September 30, 2014

Abstract

We provide a general framework to study stochastic sequences related to individual learning in economics, learning automata in computer sciences, social learning in marketing, and other applications. More precisely, we study the asymptotic properties of a class of stochastic sequences that take values in $[0, 1]$ and satisfy a property called "bounded expected relative hazard rates." Sequences that satisfy this property and feature "small step-size" or "shrinking step-size" converge to 1 with high probability or almost surely, respectively. These convergence results yield conditions for the learning models in Erev and Roth (1998), Schlag (1998), and Börgers et al. (2004) to choose expected payoff maximizing actions with probability one in the long run.

Keywords: Hazard rate, individual learning, social learning, two-armed bandit algorithm, dynamic system, stochastic approximation, submartingale, convergence.

1 Introduction

Stochastic sequences arising in the analysis of several models in economics often exhibit expected hazard rates that are proportional to the sequence's current value. For instance, models of technology adoption often satisfy that the change in the fraction of a population that adopts a new technology is proportional to the product of the current fraction of adopters and the current fraction of non-adopters (see, e.g., Young (2009)). This follows from the assumption that diffusion of technology requires non-adopters to observe adopters in order to learn about the new technology. A similar reasoning applies to models in

*We thank J. Hedlund, S. Lakshmivarahan and M.A.L. Thathachar for helpful correspondence on the subject matter of this paper, and two referees and the Associate Editor for insightful and constructive comments. Oyarzun acknowledges financial support of the Ministerio de Ciencia y Tecnologia, FEDER funds under project SEJ2007-62656, and of the Instituto Valenciano de Investigaciones. Ruf acknowledges financial support of the Visitors Program of the School of Economics at the University of Queensland and of the Oxford-Man Institute of Quantitative Finance at the University of Oxford, where a major part of this work was completed.

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

[‡]Department of Mathematics, University College London. E-Mail: j.ruf@ucl.ac.uk

other disciplines, such as Bass' celebrated model of new product growth (see, e.g., Bass (1969), Jackson and Yariv (2011)) and selection models in biological evolution (see, e.g., Nowak (2006)). As we discuss below, models of individual and social learning provide another class of examples for stochastic sequences with expected hazard rates that are proportional to the sequences' current value. In these models, the sequences represent the probability of choosing optimal actions.

The analysis of such models usually concerns the question whether a new technology or a product gets fully adopted, a certain type takes over in a biological selection process, or an optimal action is played almost surely in the long run. Towards this end, this paper provides general conditions on expected hazard rates of a bounded stochastic sequence that guarantee the convergence to the upper bound. Here, the sequence is interpreted as a fraction of a certain type or the probability of playing an optimal action at any point in time. This paper thus provides conditions that guarantee that, in the long run, a certain type takes over the whole group of types or only optimal actions are chosen, as illustrated in the applications discussed below.

It turns out that constraints on the *relative hazard rates* of a stochastic sequence, i.e., the proportions of the hazard rates to the values of the sequence,[1] provide helpful conditions for the convergence to the upper bound. In contrast to the deterministic case, in a stochastic framework, lower bounds for the relative hazard rates are not sufficient for almost sure convergence. For example, in the case of technology adoption, full adoption might fail as the new technology may be completely abandoned at some point in time by chance, or adoption rates may drop too fast. The analysis below reveals that if the underlying submartingale moves in small or shrinking steps, convergence to the upper bound holds, nevertheless. Thus, in the long run, new technologies are used or optimal actions chosen if adoption or learning occurs in small or shrinking steps.

The first main result of this paper, Theorem 2.1, analyzes the asymptotical properties of a sequence that changes with *small step-size* and satisfies weak bounds on its relative hazard rates. Theorem 2.1 asserts that the probability of *convergence to optimality*, i.e., the event that the stochastic sequence converges to the upper bound, is arbitrarily high for sequences with sufficiently small step-size. This result allows us to obtain novel convergence results in different contexts, including, for instance, the models of individual and social learning that we discuss below. A limitation of Theorem 2.1 is that the question of how small the step-size needs to be in order to achieve any given probability of convergence to 1 is usually directly related to the probability measure of the underlying probability space. In applications, however, this probability measure is assumed to be unknown. This issue is addressed by Theorem 2.2 and Corollary 2.1, which provide sufficient conditions for achieving convergence to optimality almost surely under an extra condition that may be interpreted as requiring an arbitrary *shrinking step-size* over time.

[1] Formally, if the values of the sequence are denoted by $\{P_t\}_{t\in\mathbb{N}_0}$, then the corresponding relative hazard rates are defined as $(P_{t+1} - P_t)/((1 - P_t)P_t)$ for all $t \in \mathbb{N}_0$.

These results can be applied to the analysis of several models of boundedly rational learning (see, e.g., Erev and Roth (1998), Schlag (1998), Börgers et al. (2004)). In models of individual learning, in every period individuals choose one action out of a finite set and observe a payoff realization yielded by the action they choose (sometimes along with forgone payoffs). In models of social learning, individuals also observe the payoffs from the actions chosen by a sample of other individuals. Learning is assumed to be "adaptive," i.e., in every period, individuals make their choice according to a probability distribution over actions and this distribution is revised as new payoff observations arrive. As discussed in Section 3, our results can be used to provide conditions for learning to yield convergence to choose expected-payoff maximizing actions, either with high probability or almost surely. All details are provided in the online appendix (Oyarzun and Ruf (2014)).

Small and shrinking step-size appear often in applications. Small step-size has been used in both theoretical and experimental work in economics (see, e.g., Börgers and Sarin (1997) and Van Huyck et al. (2007), respectively). Shrinking step-size appears endogenously in the Roth-Erev model (see, e.g., Erev and Roth (1998)).[2] Researchers using the Cross (1973) model in applications often assume shrinking step-size (see, e.g., Sarin and Vahid (2004)), even though the benchmark version of this model has a fixed step-size. The condition of shrinking step-size captures the "power law of practice" in learning (see, e.g., Erev and Roth (1998) and the references therein): initial periods typically exhibit a substantial response of behavior to experience and are followed by gradually decreasing responses, such as those implied by shrinking step-size.

The question then arises when and why the "power law of practice" is relevant. Psychologists have long studied this problem. For instance, Bills (1934) and Newell and Rosenbloom (1981) study the decrease over time of motivation, psychophysical performance, or cognitive gains from experience, as possible explanations of the "power law of practice." These explanations have appeal in the analysis of economic applications, as well. In particular, motivation, psychophysical performance, and cognitive gains play an important role in the analysis of data in experimental economics, where subjects tire and lose concentration. More importantly, in real-world economic problems, the "power law of practice" seems to hold for similar reasons. For instance, Choi et al. (2009) analyze reinforcement learning and saving behavior, and provide evidence supporting the "power law of practice" hypothesis: younger investors are more responsive to their personal return realizations than older investors in terms of their 401(k) saving rates. We believe the "power law of practice" plays a role in individual and social learning in economics.

Related literature. Norman (1968) formally analyzes a two-armed bandit algorithm to study the

[2] Polya-urn-based schemes, akin to the Roth-Erev learning model (and hence yielding decreasing step-size), are useful to study different allocation problems. For instance, Durham et al. (1998) and Laruelle and Pages (2013) apply these models to the study of patient allocation in clinical trials.

asymptotic properties of reinforcement learning models considered by experimental psychologists (see, e.g., Weinstock et al. (1965)) who study learning when success or failure are the only possible outcomes. In this pioneering work, he shows that certain learning models converge with high probability to choose the action that is more likely to yield success, provided that changes in the probability of choosing each action are small. Computer scientists (see, e.g., Shapiro and Narendra (1969), Narendra and Thathachar (1974), Lakshmivarahan and Thathachar (1976), and Torkestani and Meybodi (2009)), provide similar results in the context of learning automata. Oyarzun and Sarin (2013) adapt these techniques to prove convergence of a class of learning models to risk averse choice. The settings in these papers are more restrictive than in this work, and their convergence results are implied by Theorem 2.1 below. None of these papers has a counterpart to the almost-sure convergence results in Theorem 2.2 and Corollary 2.1 below, as the models they analyze fail to satisfy our conditions on shrinking step-size over time.

The paper closest to our analysis is that of Lamberton et al. (2004), who thoroughly analyze the asymptotical properties of the two-armed bandit algorithm. This analysis is of particular interest because the algorithm may have a positive probability of converging to a non-optimal state, i.e., a "trap," despite of the probability of choosing an optimal action being a submartingale. Lamberton et al. (2004) take an approach similar to ours based on shrinking step-size to provide conditions that yield convergence to optimality almost surely. Their analysis is tailored to the specific characteristics of the two-armed bandit algorithm, whereas this paper's framework allows us to apply its results in more general settings such as the models in economics that we study in the applications.

2 Convergence for updating rules with small and shrinking step-size

2.1 Framework

In this subsection, we provide the analytical framework and introduce the condition of bounded expected relative hazard rates.

In our applications to models of individual and social learning, the realization of the state of the world in each period determines the action chosen by each individual, the obtained and forgone payoffs, and the information revealed to each individual. After observing this information, individuals adjust their behavior, i.e., the probability of choosing each action according to their behavioral rule. We now provide a formal model that encompass these models. First, we introduce the probability space and the states of the world. Then we introduce the *updating rule*, which is a function that maps the observable part of the state of the world to the current value of a stochastic process. This process represents performance, i.e., the probability of choosing optimal actions. The applications discussed in Section 3 illustrate that (a slight generalization of) this setup is broad enough to accommodate an array of models of learning.

(a) Probability space. The possible states of the world are represented by the measurable product space $(\Omega, \mathcal{F}) = (\prod_{t=1}^{\infty} \Omega_t, \otimes_{t=1}^{\infty} \mathcal{F}_t)$, where Ω_t stands for the set of states that may occur at time $t \in \mathbb{N}$ and is equipped with a sigma algebra $\mathcal{F}_t$, describing the set of events. Furthermore, let $\Omega_{[0,t]} := \prod_{\tau=1}^{t} \Omega_\tau$ denote the set of all histories up to time t and let $\Omega_{[0,0]}$ be an arbitrary singleton. Analogously, let $\mathcal{F}_{[0,t]} := \otimes_{\tau=1}^{t} \mathcal{F}_\tau$ denote the set of all events that may occur up to time $t \in \mathbb{N}$ and set $\mathcal{F}_{[0,0]} := \{\emptyset, \Omega\}$. Let $\mathbb{P}$ be a probability measure on $(\Omega, \mathcal{F})$, and $\mathbb{P}_t$ its conditional version given $\mathcal{F}_{[0,t]}$, along with its conditional expectation $\mathbb{E}_t[\cdot]$.

(b) Updating rule. An updating rule is a sequence $\Pi = \{\Pi_t\}_{t\in\mathbb{N}}$ of functions $\Pi_t : \Omega \times [0,1] \to [0,1]$ such that $\Pi_t(\cdot, p)$ is $\mathcal{F}_{[0,t]}$–measurable for all $t \in \mathbb{N}$ and $p \in [0,1]$. We shall usually omit the first argument of Π_t, for sake of notation. For a given $P_0 \in [0,1]$, the sequence $P = \{P_t\}_{t\in\mathbb{N}_0}$, defined via the iteration $P_t = \Pi_t(P_{t-1})$ for all $t \in \mathbb{N}$, is called the *performance measure* (corresponding to the initial value P_0 and the updating rule Π).

The performance measure changes in a probabilistic way over time, according to the information that becomes available and to the updating rule. Since the updating rule is allowed to depend on all information revealed up to that period, the performance measure may depend on the whole sequence of realized states up to that date.

In order to get an intuitive idea of the class of sequences studied in this paper, let us recall the interpretation borrowed from the literature on technology adoption mentioned in the introduction. Let $P_t \in [0,1]$ denote the fraction of a continuum population that has adopted a new technology by time $t \in \mathbb{N}_0$. The change in the fraction of the population that adopts the new technology is then assumed to be proportional to the rate at which a non-adopter observes an adopter. More precisely, if each individual observes only one other individual, uniformly picked from the population, then the fraction of time t non-adopters who observe time t adopters is $P_t\,(1 - P_t)$. Only a fraction of these non-adopters are assumed to adopt the new technology at time $t + 1$. In the spirit of survival theory, Young (2009) calls the ratio $(P_{t+1} - P_t) \,/\, (P_t\,(1 - P_t))$ the *relative hazard rate* of P at time $t \in \mathbb{N}_0$.

The relative hazard rates of the models studied below are allowed to be stochastic. This paper provides conditions on the expected values of the relative hazard rates that guarantee that P converges to 1 (with a high probability or almost surely). Towards this end, we next introduce an important property for expected relative hazard rates:

Definition 2.1. *An updating rule* Π *satisfies the* weakly bounded expected relative hazard rates property (WBERHR) *with lower bound sequence* $\delta := \{\delta_t\}_{t\in\mathbb{N}_0}$*, where* $\delta_t \geq 0$ *is* $\mathcal{F}_{[0,t]}$*–measurable, if*

$$\mathbb{E}_t[\Pi_{t+1}\,(p)] - p \geq \delta_t \cdot p\,(1 - p) \tag{2.1}$$

for all $p \in [0,1]$ *and* $t \in \mathbb{N}_0$. *The updating rule* Π *satisfies the* bounded expected relative hazard rates property (BERHR) *if it satisfies WBERHR with lower bound sequence* $\delta := \{\delta_t\}_{t\in\mathbb{N}_0}$ *such that* $\inf_{t\in\mathbb{N}_0}\{\delta_t\} > 0$.[3]

The nature of phenomena that yield sequences satisfying BERHR or WBERHR is diverse. These conditions often hold in learning models, such as those discussed in Section 3. They also arise in biological evolution models (see, e.g., Nowak (2006)), where the variable of interest may be the relative size of the population of a certain type of cells with respect to the whole population.[4]

The two-armed bandit algorithm in Lamberton et al. (2004) satisfies WBERHR. The analysis in this paper, however, requires a more general setup than theirs in order to accommodate the applications below. For instance, if the two-armed bandit algorithm is modified to allow that the outcomes of both arms are observed (while perhaps some past outcomes are forgotten), then the results in Lamberton et al. (2004) do not directly apply. In contrast, in the example of learning with full information that we analyze in Section B.2 of the online appendix, we illustrate how to use the results provided here in that setup.

(c) Convergence to optimality. If the updating rule Π satisfies BERHR or WBERHR, then the corresponding performance measure P is a bounded submartingale and hence, there exists (almost surely) the random variable $P_\infty = \lim_{t\uparrow\infty} P_t$. We call the event $\{P_\infty = 1\}$ convergence to optimality. The notion of almost sure convergence to optimality also appears as "infallibility" in the literature (see, e.g., Lamberton et al. (2004)).

We conclude this subsection with a relatively standard observation (see, e.g., Norman (1968)):

Lemma 2.1. *If the updating rule* Π *satisfies WBERHR with lower bound sequence* $\delta = \{\delta_t\}_{t\in\mathbb{N}_0}$ *and if* $\sum_{t=0}^{\infty}\delta_t = \infty$, *then* $P_\infty \in \{0,1\}$.

Proof. Assume that $\{P_t\}_{t\in\mathbb{N}_0}$ does not almost surely converge to either 0 or 1. Then, there exists an $\varepsilon > 0$ such that $\mathbb{P}(\lim_{t\uparrow\infty} P_t(1-P_t) > 2\varepsilon) > 2\varepsilon$. Thus, there exists a $t_0 \in \mathbb{N}$ such that the event $B := \{P_t(1-P_t) > \varepsilon \text{ for all } t \geq t_0\}$ satisfies $\mathbb{P}(B) > \varepsilon$.

By the hypothesis,

$$\mathbb{E}_t[P_{t+1}] - P_t \geq \delta_t P_t(1-P_t)$$

for all $t \in \mathbb{N}_0$; thus,

$$1 \geq \mathbb{E}[P_t] = P_0 + \sum_{\tau=0}^{t-1}\mathbb{E}\left[P_{\tau+1} - P_\tau\right] \geq \sum_{\tau=0}^{t-1}\mathbb{E}\left[\delta_\tau P_\tau(1-P_\tau)\right] \geq \sum_{\tau=t_0}^{t-1}\mathbb{E}\left[\mathbf{1}_B\delta_\tau P_\tau(1-P_\tau)\right] \uparrow \infty,$$

[3]We emphasize that the BERHR property does not require that $\inf_{t\in\mathbb{N}_0}\{\delta_t\}$ is uniformly (in $\omega \in \Omega$) bounded away from zero.

[4]Consider the populations of two types of cells, whose sizes at time $t \in \mathbb{N}_0$ are X_t and Y_t and whose exogenously given growth rates are $a \geq -1$ and $b \geq -1$. Thus $X_{t+1} - X_t = aX_t$ and $Y_{t+1} - Y_t = bY_t$. If $a > b$, then the fraction of the first type of cells in the population $P_t := X_t/(X_t + Y_t)$ satisfies BERHR: $(P_{t+1} - P_t)/(P_t(1-P_t)) \geq (a-b)/(1+a)$. In this case, since P_t is deterministic, standard arguments yield that $\lim_{t\uparrow\infty} P_t = 1$.

as t tends to infinity, leading to a contradiction. $\square$

2.2 Example for the lack of almost sure convergence to optimality

The following example, adapted from Viswanathan and Narendra (1972), illustrates that even updating rules that satisfy a strong version of BERHR, so that the lower bound sequence is uniformly (in $\omega \in \Omega$) bounded away from zero, and that have a performance sequence P that never gets absorbed, i.e., $P_t \in (0, 1)$ for all $t \in \mathbb{N}_0$, may not achieve convergence to optimality almost surely. Consider a two-armed bandit algorithm that, at each time t, chooses one out of two arms and observes the realization of a failure or success. Arm 1 and arm 2 succeed with probability μ_1 and μ_2, respectively, where $0 < \mu_1 < \mu_2 < 1$. The probability of choosing arm 1 at time t is $1 - P_t$ and the probability of choosing arm 2 is P_t. If at time t the observed realization is a failure, then $P_{t+1} = P_t$. Observed successes, however, increase the probability of choosing the same arm in the next period: if arm 2 is chosen at time t and a success is observed, then $P_{t+1} - P_t = (1 - P_t)(1 - \beta)$; and, if arm 1 is chosen and a success is observed, then $P_{t+1} - P_t = -P_t(1 - \beta)$, for some exogenously given constant $\beta \in [0, 1)$.[5] Although the underlying updating rule verifies BERHR with constant lower bound $(1 - \beta)(\mu_2 - \mu_1) > 0$, it is argued below that, with strictly positive probability, arm 1 is chosen at all times. Therefore, the event $\{P_\infty = 0\}$ has positive probability.

To see how convergence to choose arm 2 may fail, observe that if $\beta = 0$, then $P_1 = 0$ with probability $(1 - P_0)\mu_1$. Thus, in this case, $\mathbb{P}(P_\infty = 0) \geq (1 - P_0)\mu_1$. If $\beta \in (0, 1)$, partition the set of time periods in subsets or *blocks* of consecutive time periods, with cardinalities $1, 2, 3, \ldots$, i.e., $\{1\}, \{2, 3\}, \{4, 5, 6\}, \ldots$. Now, fix $j > 1$ and consider the event in which arm 1 is chosen at all times until the last time of the $(j-1)^{th}$ block, with at least one success in each block. Conditional on that event, the probability of choosing arm 1 at all times in the block of length j is at least $(1 - \beta^{j-1}P_0)^j$. Furthermore, the probability of obtaining at least one success in the j^{th} block, given that arm 1 is chosen in all time periods in that block, is $1 - (1 - \mu_1)^j$. Therefore, the probability that arm 1 is chosen at all times until the last time period of the N^{th} block, and that at least one success occurs in each block is at least

$$\prod_{j=1}^{N}(1 - \beta^{j-1}P_0)^j \cdot \prod_{j=1}^{N}(1 - (1 - \mu_1)^j). \tag{2.2}$$

We now argue that the limit of the expression in (2.2), as N tends to ∞, is strictly positive, which, directly yields $\mathbb{P}(P_\infty = 0) > 0$. Since $\sum_{j=1}^{\infty}(1 - \mu_1)^j < \infty$, we have $\prod_{j=1}^{\infty}(1 - (1 - \mu_1)^j) > 0$.[6] Finally, the inequalities $(1 - \beta^{j-1}P_0)^j \geq 1 - j\beta^{j-1}P_0 > 0$ for all large enough j imply that the first product in (2.2) converges to a strictly positive number since $\sum_{j=1}^{\infty} j\beta^{j-1} < \infty$ by the ratio test.

[5] Equivalently, if arm 1 is chosen and a success is observed, then $P_{t+1} = \beta P_t$.

[6] We use that, for $a_j < 1$, the product $\prod_{j=1}^{\infty}(1 - a_j)$ converges to a strictly positive number if the sum $\sum_{j=1}^{\infty} a_j$ converges absolutely, which follows from taking logarithms and the limit comparison test.

Intuitively, when an action is chosen initially and sufficient successes are observed, there is a positive probability that this action is always chosen. However, Corollary 2.1 below reveals that this event would not occur if the updating rule had decreasing step-size. In this example, such an updating rule is obtained by replacing $(1-\beta)$ with $(1-\beta)/(t+2)$ when updating P_t to P_{t+1} for all $t \in \mathbb{N}_0$. Theorem 2.1 yields an analog statement.

As this example illustrates, the BERHR property does not guarantee convergence to optimality almost surely. This paper is mainly concerned with strengthening this property to obtain convergence to optimality with high probability or almost surely.

2.3 Convergence results

In this subsection, we analyze the asymptotic properties of the performance measure of updating rules that satisfy either BERHR or WBERHR. In particular, we provide sufficient conditions for updating rules that satisfy these properties with shrinking step-size, in a sense that we make precise below, to yield convergence to optimality almost surely.

Towards this end, we fix an updating rule Π and a sequence $\theta := \{\theta_t\}_{t\in\mathbb{N}_0}$ such that $\theta_t \in (0,1]$ is $\mathcal{F}_{[0,t]}$–measurable, which we call a *compressing sequence*. We now consider the updating rule $\Pi^\theta = \left\{\Pi_t^\theta\right\}_{t\in\mathbb{N}}$ given by the sequence of functions $\Pi_t^\theta : \Omega \times [0,1] \to [0,1]$ that satisfy

$$\Pi_t^\theta(\cdot, p) := p + \theta_{t-1}\left(\Pi_t(\cdot, p) - p\right) \tag{2.3}$$

for all $p \in [0,1]$ and $t \in \mathbb{N}$.

We say that the updating rule Π^θ is a *small step-size version* of Π. Let $P^\theta = \left\{P_t^\theta\right\}_{t\in\mathbb{N}_0}$ be the corresponding *small step-size version* of P; to wit, $P_0^\theta = P_0 \in [0,1]$ and $P_t^\theta = \Pi_t^\theta(P_{t-1}^\theta)$ for all $t \in \mathbb{N}$. If Π satisfies WBERHR with lower bound sequence δ, then Π^θ satisfies WBERHR with lower bound sequence $\{\theta_t\delta_t\}_{t\in\mathbb{N}_0}$ since

$$\mathbb{E}_t\left[\Pi_{t+1}^\theta(p)\right] - p = \theta_t\mathbb{E}_t[\Pi_{t+1}(p) - p] \geq \theta_t\delta_t p(1-p) \tag{2.4}$$

for all $p \in [0,1]$ and $t \in \mathbb{N}_0$. Furthermore, if Π satisfies BERHR and $\inf_{t\in\mathbb{N}_0}\{\theta_t\} > 0$, Π^θ satisfies BERHR as well. Hence, as before, we can define $P_\infty^\theta := \lim_{t\uparrow\infty} P_t^\theta$ for Π^θ. Compared to the updating rule Π with corresponding performance measure P, the updating rule Π^θ yields a performance measure P^θ that moves in the same direction as P, but a smaller magnitude.

We are now ready to state the first result of this paper:

Theorem 2.1. *Suppose the updating rule* Π *satisfies WBERHR with lower bound sequence* $\delta = \{\delta_t\}_{t\in\mathbb{N}_0}$, $\sum_{t=0}^\infty \delta_t^2 = \infty$, *and* $P_0 > 0$. *Then, for all* $\varepsilon > 0$, *the sequence* $\theta = \{\theta_t\}_{t\in\mathbb{N}_0}$ *with* $\theta_t := (1 \wedge \delta_t) \cdot c \in (0,1)$, *where* c *is a constant depending only on* P_0 *and* ε, *satisfies* $\mathbb{P}(P_\infty^\theta = 1) > 1 - \varepsilon$.

Theorem 2.1 considers updating rules that satisfy WBERHR and whose relative hazard rates either vanish slowly or do not vanish. It asserts that for any arbitrary lower bound on the probability of convergence to optimality, there exists a small step-size version of the underlying updating rule such that this bound holds.

The proof of Theorem 2.1 can be found in Appendix A. It is based on the idea of applying an increasing, concave, continuous, and bijective function $\phi : [0,1] \to [0,1]$ to the submartingale P such that $\phi(P_0) > 1-\varepsilon$. The new sequence $\{\phi(P_t)\}_{t\in\mathbb{N}_0}$, in general, is not a submartingale. However, WBERHR yields a positive lower bound on the expected differences $P_{t+1} - P_t$ and ϕ is locally approximately linear. Applying ϕ to the small step-size version P^θ, which, in each step, only varies in a small neighborhood, corresponds to applying an almost linear function to a submartingale with a positive lower bound on its expected change. Hence, $\left\{\phi(P_t^\theta)\right\}_{t\in\mathbb{N}_0}$ is a submartingale. Given that $P_\infty^\theta \in \{0,1\}$, one then obtains the statement, that is, $\mathbb{P}(P_\infty^\theta = 1) > 1-\varepsilon$. The main ideas of this discussion are contained in the proof of Lemma A.1.[7,8]

Sometimes, an updating rule can directly be interpreted as a small step-size version of some fictitious updating rule. The following result uses this idea:

Theorem 2.2. *Consider an $\{\mathcal{F}_{[0,t]}\}_{t\in\mathbb{N}_0}$–adapted stochastic process $P = \{P_t\}_{t\in\mathbb{N}_0}$, taking values in $[0,1]$, such that the following three conditions hold:*

1. Non-summable Relative Hazard Rates: *The sequence P satisfies*

$$\mathbb{E}_t[P_{t+1}] - P_t \geq \delta_t P_t(1-P_t)$$

for some $\mathcal{F}_{[0,t]}$–measurable random variable $\delta_t > 0$, for all $t \in \mathbb{N}_0$, with $\sum_{t=0}^{\infty} \delta_t = \infty$.

2. WBERHR Stretchable: *There exist a random variable $\tilde{\delta} > 0$ and a sequence $\theta = \{\theta_t\}_{t\in\mathbb{N}_0}$ of almost surely non-increasing $\mathcal{F}_{[0,t]}$–measurable random variables $\theta_t \in (0,1]$ such that*

$$-P_t \leq \frac{1}{\theta_t}(P_{t+1} - P_t) \leq 1 - P_t \tag{2.5}$$

and $\delta_t/\theta_t > \tilde{\delta}$ for all $t \in \mathbb{N}_0$.

3. Relatively Fast Shrinking: *The stopping time ρ, defined as*

$$\rho := \min\{t \in \mathbb{N}_0 : P_t \geq y\theta_t\} \ \ \textit{with} \ \ \min\emptyset := \infty, \tag{2.6}$$

[7] In Lemma A.1, however, instead of constructing a new submartingale $\{\phi(P_t)\}_{t\in\mathbb{N}_0}$, we are constructing a supermartingale, using the same ideas. This modified approach simplifies the arguments for Theorem 2.1 and the assertions below.

[8] Taking a different point of view, small step-size updating leads to a higher probability of converging to 1 due to the lack of additivity of standard deviation. With small step-size, one step is replaced by several steps. While the expected values of these steps are additive, the standard deviations add up only subadditively; thus the standard deviation–to–expected value ratio decreases; and hence, the probability of convergence to 1 increases.

is almost surely finite for all $y \in \mathbb{R}$.

Then, $\lim_{t\uparrow\infty} P_t = 1$.

The first two conditions of Theorem 2.2, Non-summable Relative Hazard Rates and WBERHR Stretchable, seem natural given our previous analysis since they provide an interpretation of P as a small step-size version of the performance measure of an updating rule that satisfies WBERHR. In particular, WBERHR Stretchable requires that if at each time t, the sequence change were $(1/\theta_t)(P_{t+1} - P_t)$ instead of $(P_{t+1} - P_t)$, then the resulting value of P_{t+1} would still lie in $[0, 1]$; and such an artificial sequence would satisfy WBERHR with strictly positive expected relative hazard rate (bounded by δ_t/θ_t, i.e., the original bound times $1/\theta_t$). The third condition, Relatively Fast Shrinking, requires that the fictitious compressing sequence θ tends to zero faster than P. Section B.4 of the online appendix illustrates that the Roth-Erev learning model, for example, satisfies all these conditions.

The proof of Theorem 2.2 is similar to the one of Theorem 2.1 and can be found in Appendix A as well. Recalling the informal discussion of the proof of Theorem 2.1 above, we now compress, in each period, a fictitious sequence, which allows us to increase the concavity (and thus the value of $\phi(P_t)$) in each step, without losing the submartingale property of the process $\{\phi(P_t)\}_{t\in\mathbb{N}_0}$. At some point in time, this submartingale is greater than $1 - \varepsilon$, for any arbitrarily given ε. This event occurs in finite time due to the assumption of Relatively Fast Shrinking. From this point on, the proof follows the one of Theorem 2.1.

A limitation of Theorem 2.1 is that it does not provide sufficient conditions for convergence to optimality almost surely and that, for any small step-size version of the updating rule, one cannot pin down the probability of this event, unless the probability measure $\mathbb{P}$ is known. In applications, however, $\mathbb{P}$ is typically assumed to be unknown. These issues are taken care of by Corollary 2.1, where we consider a reciprocally linearly decreasing compressing sequence:

Corollary 2.1. *Suppose the updating rule* Π *satisfies BERHR and* $P_0 > 0$. *Then the compressing sequence* $\theta = \{\theta_t\}_{t\in\mathbb{N}_0}$, *defined by* $\theta_t = 1/(t+2)$, *satisfies* $\mathbb{P}(P^\theta_\infty = 1) = 1$.

Proof. We check that the sequence P^θ of the statement satisfies the assumptions of Theorem 2.2. By (2.4) and the fact that $\sum_{t=0}^{\infty} \theta_t = \infty$, we obtain that P^θ satisfies the Non-summable Relative Hazard Rates property. The WBERHR Stretchable property follows from the definition of P^θ. Finally, Lemma A.3 in Appendix A yields the Relatively Fast Shrinking property of P^θ. □

The reciprocally linearly decreasing compressing sequence θ in Corollary 2.1 typically appears in stochastic approximation theory. Appendix A contains a discussion on the connections of this paper's results to related findings based on arguments from that literature.

2.4 Extended framework

The stochastic processes studied in economics (and other sciences) are often multivariate. For instance, in models of individual learning, these processes correspond to vectors of probabilities of choosing each action. In order to embed such models in this paper's framework, we introduce the *configuration space* $\mathfrak{S}$, i.e., a convex subset of $\mathbb{R}^D$, where $D \in \mathbb{N}$. The elements of $\mathfrak{S}$ are called *configurations*. In the setup of individual learning, D is the number of actions, $\mathfrak{S}$ is the simplex of dimension $D-1$, and a configuration is a vector of probabilities of choosing each action.

Next, we map any configuration to a value in $[0,1]$, measuring its "performance." We call this mapping the *aggregator* $\mathfrak{A} : \mathfrak{S} \to [0,1]$ and assume that $\mathfrak{A}$ is a weighted sum of the components of the configuration. For instance, we map the vector of probabilities of choosing each action into the probability of choosing an optimal action.

With a slight misuse of notation, we redefine an updating rule as a sequence $\Pi = \{\Pi_t\}_{t\in\mathbb{N}}$ of functions $\Pi_t : \Omega \times \mathfrak{S} \to \mathfrak{S}$ such that $\Pi_t(\cdot, \sigma)$ is $\mathcal{F}_{[0,t]}$–measurable for all $t \in \mathbb{N}$ and $\sigma \in \mathfrak{S}$. A pair $(\Pi, \mathfrak{A})$ of an updating rule and an aggregator is called a *system*. The performance measure $P = \{P_t\}_{t\in\mathbb{N}_0}$ is now iteratively defined as $P_t = \mathfrak{A}(\sigma_t)$, with $\sigma_{t+1} = \Pi_{t+1}(\sigma_t)$ for all $t \in \mathbb{N}_0$ and $\sigma_0 \in \mathfrak{S}$ exogenously given.

The definitions of the properties WBERHR and BERHR generalize from updating rules to systems replacing (2.1) of Definition 2.1 with

$$\mathbb{E}_t[\mathfrak{A}(\Pi_{t+1}(\sigma))] - \mathfrak{A}(\sigma) \geq \delta_t \cdot \mathfrak{A}(\sigma)(1 - \mathfrak{A}(\sigma))$$

for all $\sigma \in \mathfrak{S}$, and $t \in \mathbb{N}_0$. The definition of small step-size versions of updating rules is adapted to this generalization, formally replacing p by σ in (2.3). We say that the system $(\Pi^\theta, \mathfrak{A})$ is a *small step-size version* of the system $(\Pi, \mathfrak{A})$ and redefine P^θ in an analogous way.

With the corresponding adjustments, all results of Section 2.3 generalize to this extended framework:

Remark 2.1. *Theorem 2.1 and Corollary 2.1 hold, mutatis mutandis, replacing the updating rule* Π *with the system* $(\Pi, \mathfrak{A})$ *and the initial performance measure* P_0 *with* $\mathfrak{A}(\sigma_0)$*. Theorem 2.2 holds as it is.*

3 Application to learning models

We provide several applications of our results in the online appendix. In a first application, we consider models of individual learning with partial information. That is, we consider an individual who every period chooses one action out of a finite set according to a probability distribution and observes a payoff realization of her choice. Upon observing this realization, she adjusts the probability of choosing each action according to a function mapping (potentially all) past realizations of her choices and obtained payoffs to the revised probability of choosing each action. In such a setup Börgers et al. (2004) identify conditions on these mappings such that, in every period, the conditional expected value of the change in the probability of choosing expected-payoff maximizing actions is positive. The conditions they identify yield learning models that satisfy the BERHR property. Their analysis, however, focuses on the change in the probability of choosing each action from one period to the next and hence, is silent about the asymptotic properties of these models. Instead, by explicitly extending their framework to an infinite horizon, we show that such learning models converge to payoff maximizing actions: (i) with high probability, when they exhibit small step-size (applying Theorem 2.1) and (ii) almost surely, when they exhibit linear shrinking step-size such as, for instance, one of the learning models considered in Sarin and Vahid (2004) (applying Corollary 2.1). We also provide a similar construction in a setting where individuals receive full information, i.e., observe both obtained and forgone payoffs. For details, see Section B.2 in the online appendix.

In a second application, we analyze models of social learning. We now consider a population of individuals who, in every period, choose one action within a finite set. Individuals adjust the probability of choosing each action upon observing their own payoff and also the actions chosen and payoffs obtained by a random sample of some of the other individuals. Each individual's revised probability of choosing each action is a weighted average of its previous probability and an *imitation component* that only places probability on the actions that the individual observed, as a function of the payoffs they yielded. The weight of the imitation component is called the *imitation rate*. Schlag (1998) considers a version of this model such that only one other individual's action and payoff is sampled and the imitation rate is 1, i.e., each individual chooses with positive probability only the action she chose or the action chosen by the individual she observed. He provides conditions on the imitation component that allow the average payoff of the population, in expected value, to increase in every period.[9] For finite populations, however, Schlag (1998) finds that, with positive probability, all individuals converge to choosing non-optimal actions. By instead assuming linearly decreasing imitation rates, we can apply Corollary 2.1 to prove that the event in which the whole population converge to choose optimal actions occurs almost surely.[10] For details, see Section B.3 in the online appendix.

In a third application, we consider the Roth-Erev model of individual learning (see Erev and Roth (1998)). In this model, in every period, each action has an "attraction" corresponding to the accumulated sum of payoffs that this action has yielded when it has been chosen, and the probability of choosing each

[9]These conditions require that the difference between the probability that an individual who chose action a switches to action b and the probability that an individual makes the opposite switch is an increasing linear function of the difference between the payoffs yielded by b and a.

[10]In Oyarzun and Ruf (2009) we generalize the conditions in Schlag (1998): if the function describing the net switching from a to b is strictly increasing in the payoff of b and strictly decreasing in the payoff of a (and satisfies some symmetry condition), then the fraction of the population who choose first-order stochastically dominant actions is strictly increasing in expectation and the convergence results of this paper can be applied to that setup as well.

action is proportional to its attraction. This model has embedded linearly shrinking step-size, therefore we can use Theorem 2.2 to analyze its asymptotical properties. Beggs (2005) and Hopkins and Posch (2005) prove that this model converges to payoff maximization using arguments based on stochastic approximation. In contrast, the proof we provide in the online appendix builds on the properties of the expected relative hazard rates of this learning model. Hence, our results provide a different interpretation of the convergence property of this model.[11] For details, see Section B.4 in the online appendix.

4 Discussion

The analysis of systems that satisfy WBERHR or BERHR can be the starting point for the study of slightly more complex dynamics. There are many other models in the literature with similar characteristics to those considered here that do not satisfy these properties. One example is the model of word-of-mouth social learning in Ellison and Fudenberg (1995). In their model, individuals sample $n \in \mathbb{N}$ other individuals out of a continuum population and choose the action that has the highest average payoff in their observed sample. Aggregate shocks (on top of individual specific shocks) of the payoffs yielded by the two available actions allow for randomness despite of the population's cardinality. For $n = 1$, their model satisfies BERHR, and hence, their findings are recovered by our results. In particular, the population may "herd" to the action with the lowest expected payoff with positive probability, and this probability goes to zero when there is enough inertia (which is equivalent to shrinking step-size in our analysis). For $n > 2$, however, their model does not satisfy WBERHR and thus our results tell us nothing about the asymptotic properties of their model. Future research could study related conditions on these systems that make it possible to analyze the asymptotic properties of the models in a general framework encompassing their findings for those cases.

Another possible extension is the study of properties of systems that satisfy (W)BERHR in games. Beggs (2005) proves that the Roth-Erev learning model leads individuals to converge to play with zero-probability actions eliminated by iterated deletion of dominated strategies. Tarres and Vandekerkhove (2012) show that the two-armed bandit algorithm converges to the arm that is optimal in average, even if its expected payoff is smaller in some periods. Their results suggest that our approach could be extended to analyze setups in which the set of optimal actions may not be the same in each single period, as it often is the case in learning in games. This topic deserves further attention in the future.

Lamberton and Pages (2008a,b) provide the rates of convergence for the two-armed bandit algorithm. Indeed, the trade-off between speed and the probability of achieving convergence to optimality is of particular interest in the literature of learning automata (see, e.g., Narendra and Thathachar (1989)) and

[11] Beggs (2005) and Hopkins and Posch (2005) assume that payoffs are bounded away from zero, which is not required in our proof.

hence, worth of further study in the setup of this paper.

In our analysis of the properties of the dynamics of choices in social learning models, our sampling assumptions may seem restrictive in some setups. For instance, observability may rule out network structures in which individuals may sample some other individuals in the network with zero probability (see, e.g., Bala and Goyal (1998)). It is intuitive, however, that the choices of individuals who are not sampled may be observed, after a number of periods, provided that there is a path of individuals connecting the individual who chose an action and another who could choose that action later via imitation. Analyzing the dynamics of the performance measure in such structures would require developing further the constructions provided above. This topic is left for future research.

A Proofs of the convergence results in Section 2.3

In this appendix, we provide the proofs of the statements in Section 2.3. Although the models we analyze resemble a typical setup of stochastic approximation theory in the spirit of Robbins and Monro (1951), Kiefer and Wolfowitz (1952), Kushner and Clark (1978), and Kushner and Yin (2003), the proofs in this appendix do not rely on standard techniques developed in that literature. For an excellent overview of that literature, we refer the reader to Benaïm (1999) and the references therein. Fudenberg and Kreps (1993), Hopkins and Posch (2005), and Benaïm and Faure (2012) provide examples where stochastic approximation techniques have been fruitfully applied to economic learning models. A similar approach to analyze bandit problems is developed by Lamberton et al. (2004), Lamberton and Pages (2008b), and Tarres and Vandekerkhove (2012). In the setup of our paper, we found that arguing from first principles and extending results in Norman (1968) and Lakshmivarahan and Thathachar (1976) was tractable for the generality of our statements.

We start with the core insight for the proofs of the statements in Section 2.3. Here, we strongly rely on the positivity of the lower bound sequence δ in the definition of the (W)BERHR condition. The following lemma is inspired by the ideas in Norman (1968) and Lakshmivarahan and Thathachar (1976):

Lemma A.1. *If the updating rule* Π *satisfies* (2.1) *for some* $p \in [0,1]$, $t \in \mathbb{N}_0$, *and* $\mathcal{F}_{[0,t]}$*–measurable* $\delta_t \geq 0$ *then*

$$\mathbb{E}_t\left[e^{-\frac{\gamma}{\theta_t}\Pi^{\theta}_{t+1}(p)}\right] \leq e^{-\frac{\gamma}{\theta_t}p}$$

for all $\mathcal{F}_{[0,t]}$*–measurable* $\gamma \in [0, 1 \wedge \delta_t]$ *and* $\mathcal{F}_{[0,t]}$*–measurable* $\theta_t \in (0,1]$.

Proof. We only need to show the statement for $\gamma > 0$. Thus, without loss of generality, assume that the event $\{\delta_t > 0\}$ occurs. Define the function $G_\gamma : [0,1] \to \mathbb{R}$ for all $\gamma \in (0, 1 \wedge \delta_t]$ by

$$G_\gamma(z) := z + \delta_t z(1-z) - \frac{1-e^{-\gamma z}}{1-e^{-\gamma}}, \tag{A.1}$$

and observe that $G_\gamma(0) = G_\gamma(1) = 0$ and that

$$\frac{\partial^2}{\partial z^2} G_\gamma(z) = -2\delta_t + \frac{\gamma^2}{1-e^{-\gamma}} \cdot e^{-\gamma z} \leq -2\delta_t + \frac{\gamma}{1-e^{-\gamma}} \cdot \gamma \leq 2(\gamma - \delta_t) \leq 0,$$

since $\gamma/(1-e^{-\gamma}) < 2$ for all $\gamma \in (0, 1 \wedge \delta_t]$. This yields $G_\gamma(z) \geq 0$. Similarly, we see that

$$\widetilde{G}_\gamma(z) := z - \frac{1-e^{-\gamma z}}{1-e^{-\gamma}} \leq 0 \tag{A.2}$$

for all $z \in [0, 1]$ and $\gamma \in (0, 1]$, which yields that

$$\mathbb{E}_t \left[\frac{1-e^{-\gamma \Pi_{t+1}(p)}}{1-e^{-\gamma}} \right] \geq \mathbb{E}_t \left[\Pi_{t+1}(p) \right] \geq G_\gamma(p) + \frac{1-e^{-\gamma p}}{1-e^{-\gamma}} \geq \frac{1-e^{-\gamma p}}{1-e^{-\gamma}},$$

for all $p \in [0, 1]$, where the first inequality follows from (A.2) with $z = \Pi_{t+1}(p)$ and the second inequality follows from (A.1) and (2.1). This yields

$$\mathbb{E}_t \left[e^{-\frac{\gamma}{\theta_t}\left(\Pi^\theta_{t+1}(p) - p\right)} \right] = \mathbb{E}_t \left[e^{-\gamma(\Pi_{t+1}(p) - p)} \right] \leq 1,$$

which proves the statement. □

Now, we provide the proof of Theorem 2.1 by applying the previous lemma:

Proof of Theorem 2.1. Fix the smallest integer $\widetilde{\gamma} = \widetilde{\gamma}(P_0, \varepsilon)$ such that $e^{-\widetilde{\gamma} P_0} < \varepsilon$. Define the compressing sequence $\{\theta\}_{t \in \mathbb{N}_0}$ by $\theta_t := (\delta_t \wedge 1)/\widetilde{\gamma} \leq 1$, observe that $\sum_{t=0}^\infty \theta_t \delta_t = \infty$, and define the process $M = \{M_t\}_{t \in \mathbb{N}_0}$ by $M_t := e^{-\widetilde{\gamma} P^\theta_t}$ for all $t \in \mathbb{N}_0$. We start by observing that M is a supermartingale since, for fixed $t \in \mathbb{N}_0$, we have

$$\mathbb{E}_t \left[M_{t+1} \right] = \mathbb{E}_t \left[e^{-\widetilde{\gamma} P^\theta_{t+1}} \right] = \mathbb{E}_t \left[e^{-\frac{(\delta_t \wedge 1)}{\theta_t} P^\theta_{t+1}} \right] \leq e^{-\frac{(\delta_t \wedge 1)}{\theta_t} P^\theta_t} = M_t,$$

where the inequality follows from Lemma A.1. Thus, M_t converges to some random variable $M_\infty \in [0, 1]$ and we obtain

$$\mathbb{P}\left(P^\theta_\infty = 1 \right) = \mathbb{E}[P^\theta_\infty] \geq 1 - \mathbb{E}[M_\infty] \geq 1 - M_0 \geq 1 - \varepsilon,$$

where we have used (2.4) and Lemma 2.1 in the first equality. □

The proof of Theorem 2.2 is similar:

Proof of Theorem 2.2. Fix $\varepsilon \in (0, 1)$ and observe that there exists a constant $\delta \in (0, 1)$ such that the event $A := \{\widetilde{\delta} \geq \delta\} \subset \{\inf_{t \in \mathbb{N}_0}\{\delta_t/\theta_t\} \geq \delta\}$ satisfies $\mathbb{P}(A) \geq 1 - \varepsilon/2$. Define $y := -\log(\varepsilon/2)/\delta > 0$ and the process $M = \{M_t\}_{t \in \mathbb{N}_0}$ by

$$M_t := e^{-\frac{\delta}{\theta_{t \wedge \rho}} P_t} \mathbf{1}_{\{\min_{u \in \{0, \ldots, t\}}\{\delta_u/\theta_u\} \geq \delta\}}$$

for all $t \in \mathbb{N}_0$, where the stopping time ρ is given in (2.6). As in the proof of Theorem 2.1 we start by showing that M is a supermartingale. Towards this end, fix some $t \in \mathbb{N}_0$ and assume, without loss of generality, that we are on the event $\{\min_{u\in\{0,\ldots,t\}}\{\delta_u/\theta_u\} \geq \delta\}$. Define now $\widetilde{P}_t = P_t$ and $\widehat{P}_{t+1} = \widetilde{P}_t + (P_{t+1} - P_t)/\theta_t \in [0,1]$ and observe that $\mathbb{E}_t[\widehat{P}_{t+1}] - \widetilde{P}_t \geq \delta \widetilde{P}_t(1-\widetilde{P}_t)$. We then obtain that

$$\mathbb{E}_t\left[M_{t+1}\right] \leq \mathbb{E}_t\left[e^{-\frac{\delta}{\theta_{(t+1)\wedge\rho}}P_{t+1}}\right] \leq \mathbb{E}_t\left[e^{-\frac{\delta}{\theta_{t\wedge\rho}}P_{t+1}}\right] = \mathbb{E}_t\left[e^{-\frac{\delta\theta_t/\theta_{t\wedge\rho}}{\theta_t}P_{t+1}}\right] \leq e^{-\frac{\delta}{\theta_{t\wedge\rho}}P_t} = M_t,$$

where we have used Lemma A.1, in which we interpret (P_t, P_{t+1}) as the "compressed version" of $(\widetilde{P}_t, \widehat{P}_{t+1})$. Thus, as in the proof of Theorem 2.1, M_t converges to some random variable $M_\infty \in [0,1]$.

As $\rho < \infty$ almost surely by assumption, we obtain from an argument similar to the one in Lemma 2.1 that

$$\begin{aligned}\mathbb{P}\left(P_\infty = 1\right) = \mathbb{E}[P_\infty] &\geq 1 - \mathbb{E}\left[e^{-\frac{\delta}{\theta_\rho}P_\infty}\right] \geq 1 - \mathbb{E}[M_\infty] - \mathbb{P}(A^C) \geq \mathbb{P}(A) - \mathbb{E}[M_\rho] \\ &\geq \mathbb{P}(A) - \mathbb{E}\left[e^{-\frac{\delta}{\theta_\rho}P^\theta_\rho}\right] \geq 1 - \frac{\varepsilon}{2} - e^{-\delta y} = 1-\varepsilon,\end{aligned}$$

similarly to the proof of Theorem 2.1, where $A^C := \Omega \setminus A$. As ε was chosen arbitrarily, we obtain the statement. $\square$

For the proof of Corollary 2.1, we make the following useful observation:

Lemma A.2. *For an updating rule Π with corresponding performance sequence $\{P_t\}_{t\in\mathbb{N}_0}$, we have $P^\theta_t \geq P_0/(t+1)$ for all $t \in \mathbb{N}_0$ for the compressing sequence θ defined by $\theta_t = 1/(t+2)$.*

Proof. Assume that we have shown $P^\theta_t \geq P_0/(t+1)$ for some fixed $t \in \mathbb{N}_0$. Then

$$P^\theta_{t+1} = P^\theta_t + \frac{1}{t+2}\left(\Pi_{t+1}\left(P^\theta_t\right) - P^\theta_t\right) \geq P^\theta_t - \frac{P^\theta_t}{t+2} = \frac{(t+1)P^\theta_t}{t+2} \geq \frac{P_0}{t+2},$$

and the statement follows by induction. $\square$

Simple computations then yield the following core conclusion:

Lemma A.3. *Let the updating rule Π satisfy WBERHR with performance sequence $\{P_t\}_{t\in\mathbb{N}_0}$ and $P_0 > 0$ and let θ denote the compressing sequence defined by $\theta_t := 1/(t+2)$ for all $t \in \mathbb{N}_0$. Fix $y > 0$ and define the stopping time ρ as*

$$\rho := \min\left\{t \in \mathbb{N}_0 : P^\theta_t \geq \frac{y}{t+2}\right\} \text{ with } \min\emptyset := \infty. \tag{A.3}$$

Then $\mathbb{P}(\rho < \infty) = 1$.

Proof. We shall show the statement with y in (A.3) replaced by NP_0, where $N = [y/P_0] + 1$ with $[\cdot]$ here denoting the largest integer smaller than the argument. Define $N_k := N^{2k} - 1$ for all $k \in \mathbb{N}_0$. Since

$\mathbb{P}(\rho > t)$ is non-increasing in t, it is sufficient to show that $\mathbb{P}(\rho > N_k) \downarrow 0$ as $k \uparrow \infty$. This follows if there exists some constant $c \in [0, 1)$ independent of k such that

$$p_1 := \mathbb{P}(\rho > N_{k+1}) \leq c\mathbb{P}(\rho > N_k) =: cp_0 \tag{A.4}$$

for all $k \in \mathbb{N}$. Towards this end, Lemma A.2 and the submartingale property of $\{P_t\}_{t\in\mathbb{N}_0}$ yield

$$\begin{aligned}\frac{P_0}{N^{2k}}p_0 &\leq \mathbb{E}[P^\theta_{N_k}\mathbf{1}_{\{\rho>N_k\}}] \leq \mathbb{E}\left[\mathbb{E}_{N_k}[P^\theta_{N_{k+1}\wedge\rho}]\mathbf{1}_{\{\rho>N_k\}}\right] = \mathbb{E}\left[P^\theta_{N_{k+1}}\mathbf{1}_{\{\rho>N_{k+1}\}}\right] + \mathbb{E}\left[P^\theta_\rho\mathbf{1}_{\{N_k<\rho\leq N_{k+1}\}}\right]\\ &\leq \frac{NP_0}{N^{2(k+1)}+1}p_1 + \left(\frac{NP_0}{N^{2k}+1} + \frac{1}{N^{2k}+1}\right)(p_0 - p_1) \leq \frac{P_0}{N^{2k}}\left(\frac{p_1}{N} + \left(N + \frac{1}{P_0}\right)(p_0 - p_1)\right)\end{aligned}$$

since $P^\theta_{t+1} \leq P^\theta_t + \theta_t$ and $p_0 \geq p_1$. Sorting terms, we have

$$p_1\left(N + \frac{1}{P_0} - \frac{1}{N}\right) \leq p_0\left(N + \frac{1}{P_0} - 1\right)$$

and, thus, (A.4). □

Lamberton et al. (2004) prove a generalized version of Corollary 2.1 for the special case of the two-armed bandit algorithm in their Theorem 1(c) and Corollary 2. Their argument can be summarized in three steps. First, they establish a version of Lemma 2.1 above. Then, via a coupling argument, they relate the two-armed bandit algorithm to another one where both arms have the same distribution. Third, they use martingale methods to prove that the later algorithm does not get trapped in zero. Putting the three steps together they obtain a version of Corollary 2.1 above.

In contrast, the proofs provided here rely deeply on the bounds on the relative hazard rates. In particular, Lemma A.1 is related to the fact that a concave function of a submartingale is still a submartingale, as long as the original submartingale only moves in small steps, and is "sufficiently drifted," which the BERHR condition guarantees. Lemma A.3, which is used in the proof of Corollary 2.1, can be interpreted as the statement that the performance sequence does not get trapped in zero, similar to the third step in Lamberton et al. (2004).

References

Bala, V. and Goyal, S. (1998). Learning from neighbours. *Review of Economic Studies*, 65:595–621.

Bass, F. (1969). A new product growth for model consumer durables. *Management Science*, 15:215–227.

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

Benaïm, M. (1999). Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités XXXIII*, pages 1–68. Springer, Berlin.

Benaïm, M. and Faure, M. (2012). Stochastic approximation, cooperative dynamics and supermodular games. *Annals of Applied Probability*, 22(5):2133–2164.

Bills, A. G. (1934). *General Experimental Psychology.* Longmans, Green and Co, New York.

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.

Choi, J., Laibson, D., Madrian, B., and Metrick, A. (2009). Reinforcement learning and savings behavior. *Journal of Finance*, 64(6):2515–2534.

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

Durham, S., Flournoy, N., and Li, W. (1998). A sequential design for maximizing the probability of a favourable response. *Canadian Journal of Statistics*, 26(3):479–495.

Ellison, G. and Fudenberg, D. (1995). Word of mouth communication and social learning. *Quarterly Journal of Economics*, 110:93–125.

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.

Fudenberg, D. and Kreps, D. M. (1993). Learning mixed equilibria. *Games and Economic Behavior*, 5:320–367.

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

Jackson, M. and Yariv, L. (2011). Diffusion, strategic interaction, and social structure. In Benhabib, J., Bisin, A., and Jackson, M., editors, *Handbook of Social Economics*. North Holland.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23(3):462–466.

Kushner, H. and Clark, D. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer, Berlin.

Kushner, H. J. and Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, Berlin.

Lakshmivarahan, S. and Thathachar, M. (1976). Bounds on the convergence probabilities of learning automata. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-6:756–763.

Lamberton, D. and Pages, G. (2008a). How fast is the bandit? *Stochastic Analysis and Applications*, 26:603–623.

Lamberton, D. and Pages, G. (2008b). A penalized bandit algorithm. *Electronic Journal of Probability*, 13:341–373.

Lamberton, D., Pages, G., and Tarres, P. (2004). When can the two-armed bandit algorithm be trusted? *Annals of Applied Probability*, 14:1424–1454.

Laruelle, S. and Pages, G. (2013). Randomized urn models revisited using stochastic approximation. *Annals of Applied Probability*, 23(4):1409–1436.

Narendra, K. and Thathachar, M. (1989). *Learning Automata: An Introduction*. Prentice Hall, New Jersey.

Narendra, K. S. and Thathachar, M. A. (1974). Learning automata – a survey. *IEEE Transactions on Systems, Man and Cybernetics*, 4:323–334.

Newell, A. and Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In Anderson, J. R., editor, *Cognitive Skills and Their Acquisition*. Lawrence Eribaum Associates, Hillsdale.

Norman, M. F. (1968). On the linear model with two absorbing barriers. *Journal of Mathematical Psychology*, 5(2):225–241.

Nowak, M. A. (2006). *Evolutionary Dynamics*. Harvard University Press, Cambridge.

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

Oyarzun, C. and Ruf, J. (2014). Online appendix to: Convergence in models with bounded expected relative hazard rates. *manuscript*.

Oyarzun, C. and Sarin, R. (2013). Learning and risk aversion. *Journal of Economic Theory*, 148:196–225.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, pages 400–407.

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.

Shapiro, I. J. and Narendra, K. S. (1969). Use of stochastic automata for parameter self-optimization with multimodal performance criteria. *IEEE Transactions on Systems, Man and Cybernetics*, 5:352–360.

Tarres, P. and Vandekerkhove, P. (2012). On ergodic two-armed bandits. *Annals of Applied Probability*, 22:457–476.

Torkestani, J. and Meybodi, M. (2009). Solving the minimum spanning tree problem in stochastic graphs using learning automata. *International Conference on Information Management and Engineering*, DOI 10.1109:643–647.

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.

Viswanathan, R. and Narendra, K. (1972). A note on the linear reinforcement scheme for variable-structure stochastic automata. *IEEE Transactions on Systems, Man and Cybernetics*, 2:292–294.

Weinstock, S., North, A. J., Brody, A. L., and Loguidice, J. (1965). Probability learning in the **T** maze with noncorrection. *Journal of Comparative and Physiological Psychology*, 60:76–81.

Young, P. (2009). Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American Economic Review*, 99:1899–1924.