The Annals of Applied Probability
2017, Vol. 27, No. 5, 2907–2925
DOI: 10.1214/16-AAP1271

REINFORCEMENT LEARNING FROM COMPARISONS: THREE ALTERNATIVES ARE ENOUGH, TWO ARE NOT

BY BENOÎT LASLIER[1] AND JEAN-FRANÇOIS LASLIER

Université Paris Diderot and CNRS, Paris School of Economics

This paper deals with two generalizations of the Polya urn model where, instead of sampling one ball from the urn at each time, we sample two or three balls. The processes are defined on the basis of the problem of finding the best alternative using pairwise comparisons which are not necessarily transitive: they can be thought of as evolutionary processes that tend to reinforce currently efficient alternatives. The two processes exhibit different behaviors: with three balls sampled, we prove almost sure convergence towards the unique optimal solution of the comparisons problem while, in some cases, the process with two balls sampled has almost surely no limit. This is an example of a natural reinforcement model with no exchangeability whose asymptotic behavior can be precisely characterized.

1. Introduction. In a pairwise comparison problem, one is given a set of alternatives, with data about how they compare to each other. In its purest form, on which we focus in the present paper, we simply have, for any pair of distinct alternatives, the information of which one "beats" the other. Such a data set is called a tournament. Basic results on this structure can be found in Moon [19].

If a chess player beats all the other players, he or she is clearly the best. If a candidate cannot be defeated under majority rule by any challenger, that "Condorcet" candidate can claim to be the best according to majority rule. However, if no alternative beats all the others, it is not clear how to define the best alternatives. The problem of choosing from pairwise comparisons has thus attracted the attention of scholars in various fields, most often from the axiomatic, normative, point of view (David [6]; Fishburn [7]; Rubinstein [26]; Laslier [14]; Brandt et al. [1]).

In the present paper, we tackle the same problem from an evolutionary perspective instead of an axiomatic one. We study the two following processes, both of which consist in playing the tournament among few random alternatives and reinforcing the winner. Consider an urn of infinite capacity, where balls have colors corresponding to the alternatives.

Two-alternative reinforcement. Sample two balls from the urn. If they have different colors, one "beats" the other; then add an extra ball of the winning color. If they have the same color, add an extra ball of this color.

Received June 2016; revised October 2016.
[1]Supported in part by EPSRC Grant EP/I03372X/1.
MSC2010 subject classifications. 60J20, 91A22, 91E40.
Key words and phrases. Learning, tournament, urn process.

Three-alternative reinforcement. Sample three balls from the urn. If the three balls have different colors that form a rock-paper-scissor like cycle, reinforce one of these colors at random. Otherwise, reinforce the winning color.

We obtain two very different asymptotic behaviors for these two processes.

(i) In three-alternative reinforcement, the process is able to discover the optimal solution of the tournament, that is, the unique probability distribution p^* which is, in expectation, defeated by no alternative. With probability one, the composition of the sampling urn, which defines the probability p_τ of choosing the various alternatives at time τ, tends to p^* when τ tends to infinity.

(ii) In two-alternative reinforcement, the process is not able to discover the optimal solution, unless the solution is degenerated, with one alternative defeating all the others. With probability one, the composition of the sampling urn, which gives the probability p_τ of choosing the alternatives, concentrates on the support of the optimal solution p^*. This means that all the alternatives which are played with zero probability in the optimal solution are chosen with a probability going to 0. However, the composition of the urn may have no limit, staying away from the optimal solution. In some cases, we even prove that with probability one it has no limit.

Notice that the problem of finding the optimal solution of a tournament game is not difficult from the computational point of view. It can be done in polynomial time with respect to its size (Brandt and Fisher [2]). We are not here interested in computation but in decentralized learning through evolutionary processes.

The negative result (ii) seems original in the context of random urn models. It echoes known results about the evolutionary instability of mixed equilibria in evolutionary game theory, for instance cycling with probability one is proven by Posch [24] in a reinforcement urn model for 2×2 games.

The positive result (i) seems more expected of an urn model. However, in the context of evolutionary game theory, one typically does not have almost sure convergence toward the optimal solution of a zero-sum game. For example in a study of imitation processes in Matching Pennies games, Hofbauer and Schlag [12] observe that there is always cycling even though players end up closer to the equilibrium if they sample several individuals before imitating.

The techniques we use to derive these results are standard in the field of adaptive processes with reinforcement (see Pemantle [23]). They belong to the family of martingales techniques. The main ingredient of the proof is the definition of a well-chosen function of the process whose values form a martingale [see equation (11)]. We use the convergence theorem for positive martingales to obtain some global asymptotic information about the process. In the case of three alternatives, we get fairly directly the convergence of the process while, for the case of two alternatives, the convergence theorem has to be complemented with a variance analysis to prove the nonconvergence. One should note that the processes are not exchangeable, so that other standard techniques do not apply.

The paper is organized as follows. Section 2 introduces the necessary notation about tournaments (2.1), the Markov chain induced by the play of small-size tournaments at each date (2.2), the tournament game which allows to define and to prove existence of the optimal solution (2.3), and some further preliminary material (2.4, 2.5). Section 3 starts by the definition of urns and of the adaptive processes (3.1). Then, in order to illustrate the argument in a simple way, a toy example is introduced and treated according to its deterministic approximation (3.2). The statement and proof of our main result on Three-alternative reinforcement is found in (3.3) and Two-alternative reinforcement is treated in (3.4), before a short conclusion (3.5).

2. Framework.

2.1. *Tournaments.* Let X be a finite set. A *tournament* T on X is a complete and antisymmetric binary relation. For any x and y in X, one and only one of the three possibilities occurs: $x = y$, xTy, or yTx. When xTy we often say that x beats y. Define the sets

$$T^+(x) = \{y \in X : xTy\}, \qquad T^-(x) = \{y \in X : yTx\}. \tag{1}$$

The binary relation T is fixed throughout this paper. It is sometimes easier to use the notation

$$\max\{x, y\} = \begin{cases} x, & \text{if } xTy \text{ or } x = y, \\ y, & \text{if } yTx. \end{cases}$$

An alternative which beats all other alternatives is called a *Condorcet winner*. A tournament can have a Condorcet winner or not, but cannot have two. The *Top-Cycle* of the tournament T is the smallest (by inclusion) nonempty set $Y \subseteq X$ such that

$$\forall x \in X \setminus Y, \qquad \forall y \in Y, \qquad yTx.$$

It is easily seen that such a set is unique and reduces to a singleton $\{c\}$ if and only if c is a Condorcet winner.[2]

2.2. *A Markov chain.* Let $\Delta(X)$ be the set of probability distributions on X and let $p \in \Delta(X)$. The support of p is denoted by $\mathrm{Supp}(p)$. Given p, define a sequence $(p^{[t]})_{t\in\mathbb{N}}$ of probability distributions on X derived from p in the following way:

$$p^{[0]} = p, \tag{2}$$

$$p^{[t+1]}(x) = p^{[t]}(x) \cdot p\big(T^+(x) \cup \{x\}\big) + p^{[t]}\big(T^+(x)\big) \cdot p(x), \tag{3}$$

for any $t \geq 0$, for any $x \in X$.

[2]The literature on tournaments and formal political science has shown that the Top-Cycle is usually a very large set (McKelvey [18]), and has proposed many refinements of this set (see [14] for a survey).

The interpretation is that $p^{[t]}$ is the distribution of a random variable $\xi(t) \in X$ such that $\xi(0)$ is chosen at random according to p and then, given that $\xi(t) = x$, $\xi(t+1)$ is the winner (according to T) of the comparison between x and some alternative y randomly chosen in X according to p. Therefore, $\xi(t+1) = x$ either because $\xi(t)$ was already equal to x and y was chosen in $T^+(x) \cup \{x\}$ (first term in the above formula), or because $\xi(t)$ was in $T^+(x)$ and x was chosen according to p (second term). We call p the "sampling" probability.

An equivalent description is a random walk on the graph with vertex set X and that has an oriented edge from x to y whenever yTx or $x = y$. The edge from x to y is given a weight $p(y)$ and the process is then the usual random walk on a directed weighted graph.

This process is usually considered with p uniform on X (Daniels [5], Ushakov [29], Levchenkov [17], Slutzky and Volij [28], Chebotarev and Shamis [3, 4]). We need the general version because, later in this paper, p will be endogenous. Given p, the stationary distribution for this finite Markov chain exists and is unique;[3] we denote it by $p^{[\infty]}$. It is characterized by the fact that $\mathrm{Supp}(p^{[\infty]}) \subseteq \mathrm{Supp}(p)$ and, for any x in $\mathrm{Supp}(p)$,

$$p^{[\infty]}\big(T^+(x)\big) \cdot p(x) = p^{[\infty]}(x) \cdot p\big(T^-(x)\big). \tag{4}$$

Notice that the inclusion $\mathrm{Supp}(p^{[\infty]}) \subseteq \mathrm{Supp}(p)$ may be strict; indeed, $p^{[\infty]}(x) = 0$ when $p(T^+(x)) = 0$, that is when x beats no alternative in the support of p. More exactly, $\mathrm{Supp}(p^{[\infty]})$ is the Top-Cycle of the restriction of T to $\mathrm{Supp}(p)$: by construction any state in the set is accessible from any other state. Thus, $\mathrm{Supp}(p^{[\infty]})$ does not exactly really depends on p but only on $\mathrm{Supp}(p)$. If p has full support, for instance in the usual case where p is uniform, $\mathrm{Supp}(p^{[\infty]}) = TC(T)$.

2.3. *The tournament game.* The tournament game is the two-player, symmetric, zero-sum game defined by the strategy set X and the payoff function $g(x, y) = +1$ if xTy, $g(x, y) = 0$ if $x = y$, and $g(x, y) = -1$ if yTx. For $p, q \in \Delta(X)$ two probability distributions on X, write

$$g(p, q) = \sum_{x,y \in X} g(x, y)p(x)q(y), \tag{5}$$

and for a Dirac distribution δ_x write $g(x, q)$ for $g(\delta_x, q)$. From the definition, g is clearly antisymmetric: $g(q, p) = -g(p, q)$.

Remarkably, such a game has a unique equilibrium; Fisher and Ryan [10] prove this using linear algebra and Laffond et al. [13] have a direct proof using a parity argument.[4] Here is the precise result that will be needed in the sequel.

[3] We state the results in this section without proofs. They are easily derived from elementary theory of finite Markov chains and have already been noticed for p uniform in the mentioned references.

[4] The tournament game has been studied by graph theorists (Fisher and Ryan [8, 10, 11]) and has more recently attracted attention of computer scientists (Rivest and Chen [25]). As a model of majority voting and two-party electoral competition, it studied in Social Choice theory and formal Political Science (Moulin [20], Myerson [21, 22], Laslier [15, 16]).

PROPOSITION 1. *There exists a unique* $p^* \in \Delta(X)$ *such that* $g(p^*, q) \geq 0$ *for all* $q \in \Delta(X)$. *This* p^*, *called the optimal strategy, is also characterized by the following*: *for all* $x \in X$,

$$p^*(x) > 0 \quad \Longleftrightarrow \quad g\big(x, p^*\big) = 0,$$
$$p^*(x) = 0 \quad \Longleftrightarrow \quad g\big(x, p^*\big) < 0.$$

The intuition here is that, due to the symmetry of the game, the value of the game is zero, thus at equilibrium no alternative can have a strictly positive payoff and any alternative with a strictly negative payoff will not be played. The support of the optimal strategy is called the *Bipartisan Set* of the tournament: $\mathrm{Supp}(p^*) = \mathrm{BP}(T)$. This set is a subset of the Top Cycle (an alternative outside the Top Cycle has payoff -1 against any alternative inside) and the inclusion is often strict. For instance, in totally random tournaments, the Top Cycle typically contains all the alternatives and the Bipartisan Set contains only half of them (Fisher and Reeves [9]).

2.4. *Two formulas*. Before we go further and explain the relation between the game optimal strategy and stationary probabilities, it is useful to state two technical formulas. The following lemma describes, in term of the payoff function g, the probabilities $p^{[1]}$ and $p^{[2]}$, obtained after sampling two or three alternatives with the Markov chain defined in Section 2.2.

LEMMA 2. *For any* $x \in X$,

$$p^{[1]}(x) = p(x) \cdot \big(1 + g(x, p)\big),$$
$$p^{[2]}(x) = p(x) \cdot \left(1 + \frac{3}{2} g(x, p) + \frac{1}{2} g(x, p)^2 + \frac{1}{2} \sum_{y \in X} p(y) g(x, y) g(y, p)\right).$$

PROOF. First, let us notice a useful equality. By definitions (1) and (5),

$$g(x, p) = p\big(T^+(x)\big) - p\big(T^-(x)\big), \tag{6}$$

and, since $p(T^+(x)) + p(T^-(x)) + p(x) = 1$, we get:

$$1 + g(x, p) = 2p\big(T^+(x)\big) + p(x). \tag{7}$$

One thus obtains

$$\begin{aligned} p^{[1]}(x) &= p(x) \cdot \big(2p\big(T^+(x)\big) + p(x)\big) \\ &= p(x) \cdot \big(1 + g(x, p)\big). \end{aligned}$$

For the second formula,

$$
\begin{aligned}
p^{[2]}(x) &= p(x)\cdot\big(1+g(x,p)\big)\cdot\Big(p(x)+\sum_{y\in T^+(x)}p(y)\Big)\\
&\quad + p(x)\sum_{y\in T^+(x)}p(y)\big(1+g(y,p)\big)\\
&= p(x)^2\big[1+g(x,p)\big]+p(x)\sum_{y\in T^+(x)}p(y)\big[\big(2+g(x,p)\big)+g(y,p)\big]\\
&= p(x)\cdot\Big(\Big(p(x)+2\sum_{y\in T^+(x)}p(y)\Big)+g(x,p)\Big(\frac{1}{2}p(x)+\sum_{y\in T^+(x)}p(y)\Big)\\
&\quad + \frac{1}{2}p(x)g(x,p)+\sum_{y\in T^+(x)}p(y)g(y,p)\Big)
\end{aligned}
$$

one easily concludes from equation (7). □

2.5. *Relation between optimal strategies and stationary probabilities.* We first observe that the game optimal strategy p^* satisfies a nice fixed-point property if we take $p^{[0]}=p^*$ as the sampling probability to build the Markov chain, and that only an optimal strategy can be such a fixed point.

PROPOSITION 3. *Let p^* be the optimal strategy for the tournament game, then $(p^*)^{[1]}=(p^*)^{[\infty]}=p^*$. Conversely, let p be such that $p^{[1]}=p$, then p is the optimal strategy for the tournament game restricted to the support of p.*

PROOF. By Lemma 2, $p^{*[1]}(x)=p^*(x)(1+g(x,p^*))$ and, by Proposition 1, either $p^*(x)=0$ or $g(x,p^*)=0$.

Conversely, if $p^{[1]}(x)=p(x)=p(x)(1+g(x,p))$ then $g(x,p)=0$ as soon as $p(x)>0$ and p is the optimal strategy on its support. □

3. Learning. With the previous background material in mind, we turn to the main result of this paper. Instead of considering re-sampling at each date according to a constant probability distribution, as is done in the previously described Markov chains, we describe learning processes where winning alternatives are reinforced at the level of the sampling probability. These processes can be implemented with random urns.

3.1. *Choice by reinforcement.* An *urn* on X is a list n of strictly positive integers $n(x)$, $x\in X$. The integer $n(x)$ is the "number of balls of color x in the urn n." The set of such urns on X is denoted by $\mathcal{N}$, formally

$$\mathcal{N}=\mathbb{N}_+^X.$$

To each, $n \in \mathcal{N}$ is associated the probability distribution $\tilde{n}$ on X defined by

$$\tilde{n}(x) = \frac{n(x)}{\sum_{y \in X} n(y)}.$$

When we write that the alternative x is picked in the urn n, we mean that x is picked in X according to the probability $\tilde{n}$.

A *random urn sequence* is a sequence U_τ, $\tau \in \mathbb{N}$ of random variables on $\mathcal{N}$ such that $U_{\tau+1}$ is defined conditionally on U_τ. Here are three examples:

1. Two-alternative reinforcement. Given a realization $n_\tau \in \mathcal{N}$ of U_τ, an alternative x is picked in X according to the probability distribution $\tilde{n}_\tau^{[1]}$, and one ball of color x is added to the urn: $n_{\tau+1}(x) = n_\tau(x) + 1$ and for all $y \neq x$, $n_{\tau+1}(y) = n_\tau(y)$. This means that two alternatives, say a and b are picked independently in the urn n_τ, and are compared according to T. The result of the comparison is $x = \max\{a, b\}$, that is, $x = a$ if $a = b$ or if aTb and $x = b$ if bTa. Alternative x is reinforced.
2. Three-alternative reinforcement. Same thing as above, with the probability distribution $\tilde{n}_\tau^{[2]}$. This means that three alternatives, say a, b and c are picked independently in X according to n_τ; a, b and c are compared according to T in sequence and one ball of color $x = \max\{\max\{a, b\}, c\}$ is added to the urn. Note that there are only two cases: ranked alternatives where we reinforce the top one or a cycle where we reinforce at random. Therefore, this description is equivalent to the one in the Introduction.
3. Fast reinforcement. Same thing as above, with the probability distribution $\tilde{n}_\tau^{[\infty]}$, the stationary distribution for T when sampling is done according to $\tilde{n}_\tau$.

Note that the first two examples can be concretely implemented easily, as described, but fast reinforcement cannot.

3.2. *Two motivating differential systems.* This section presents two deterministic differential systems inspired by Two- and Three-alternative reinforcement in the simplest tournament: "rock-paper-scissors", that is, a cycle of three alternatives. With this example we will see that Two-alternative reinforcement should not converge to the optimal probability, even for a simple tournament, while Three-alternative reinforcement should. The actual proofs in the next section will follow the same overall structure with technical changes for the general tournament, the discrete time and the probabilistic evolution, in particular the same logarithmic function will be used.

Consider a cycle of three alternatives A, B and C with ATB, BTC, CTA, and consider the following deterministic systems of differential equations (which should mimic the large time behavior of the urn). We write $a(t)$, $b(t)$ and $c(t)$ the "numbers" of balls of each type [which will be such that $a(t) + b(t) + c(t) = t$] and

$\tilde{a}(t) = a(t)/t$, $\tilde{b}(t) = b(t)/t$ and $\tilde{c}(t) = c(t)/t$ the corresponding probabilities. For Two-alternative reinforcement, we get

$$\begin{aligned} \frac{\mathrm{d}a}{\mathrm{d}t} &= \tilde{a}^2 + 2\tilde{a}\tilde{b}, \\ \frac{\mathrm{d}b}{\mathrm{d}t} &= \tilde{b}^2 + 2\tilde{b}\tilde{c}, \\ \frac{\mathrm{d}c}{\mathrm{d}t} &= \tilde{c}^2 + 2\tilde{c}\tilde{a} \end{aligned} \tag{8}$$

and we note that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(\ln\tilde{a} + \ln\tilde{b} + \ln\tilde{c}) &= \frac{\mathrm{d}}{\mathrm{d}t}(-3\ln t + \ln a + \ln b + \ln c) \\ &= -\frac{3}{t} + \frac{\tilde{a} + 2\tilde{b}}{t} + \frac{\tilde{b} + 2\tilde{c}}{t} + \frac{\tilde{c} + 2\tilde{a}}{t} \\ &= 0 \end{aligned}$$

so $(\tilde{a}, \tilde{b}, \tilde{c})$ cannot converge to the optimal probability independently of the state at finite time.

For Three-alternative reinforcement,

$$\begin{aligned} \frac{\mathrm{d}a}{\mathrm{d}t} &= \tilde{a}^3 + 3\tilde{a}^2\tilde{b} + 3\tilde{a}\tilde{b}^2 + 2\tilde{a}\tilde{b}\tilde{c}, \\ \frac{\mathrm{d}b}{\mathrm{d}t} &= \tilde{b}^3 + 3\tilde{b}^2\tilde{c} + 3\tilde{b}\tilde{c}^2 + 2\tilde{a}\tilde{b}\tilde{c}, \\ \frac{\mathrm{d}c}{\mathrm{d}t} &= \tilde{c}^3 + 3\tilde{c}^2\tilde{a} + 3\tilde{c}\tilde{a}^2 + 2\tilde{a}\tilde{b}\tilde{c} \end{aligned} \tag{9}$$

and for the same quantity

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(\ln\tilde{a} + \ln\tilde{b} + \ln\tilde{c}) &= -\frac{3}{t} + \frac{\tilde{a}^2 + 3\tilde{a}\tilde{b} + 3\tilde{b}^2 + 2\tilde{b}\tilde{c}}{t} + \cdots \\ &= \frac{\tilde{a}^2 + \tilde{b}^2 + \tilde{c}^2 - (\tilde{a}\tilde{b} + \tilde{b}\tilde{c} + \tilde{c}\tilde{a})}{t}. \end{aligned}$$

Simple calculus shows that this last term is positive except for $\tilde{a} = \tilde{b} = \tilde{c} = 1/3$. Then $\ln\tilde{a} + \ln\tilde{b} + \ln\tilde{c}$ is an increasing negative function, so it converges. It is not difficult to see, using the divergence of $\int 1/t\,\mathrm{d}t$, that this implies that $\tilde{a}^2 + \tilde{b}^2 + \tilde{c}^2 - (\tilde{a}\tilde{b} + \tilde{b}\tilde{c} + \tilde{c}\tilde{a})$ converges to 0, and then that $(\tilde{a}, \tilde{b}, \tilde{c})$ converges to $(1/3, 1/3, 1/3)$ (the details of the arguments will be given in the rigorous proof of the next section).

3.3. *Three-alternative reinforcement and martingale technique.* We will now prove the result about Three-alternative reinforcement.

THEOREM 4. *For any initial urn $n_0 \in \mathcal{N}$, the random urn sequence obtained by Three-alternative reinforcement is such that the realization n_τ, $\tau \in \mathbb{N}$ almost surely verifies*

$$\lim_{\tau \to \infty} \tilde{n}_\tau = p^*.$$

The same proof will actually also give the first part of the result about Two-alternative reinforcement, which we thus state now:

THEOREM 5. *For any initial urn $n_0 \in \mathcal{N}$, the random urn sequence obtained by Two-alternative reinforcement is such that the realization n_τ, $\tau \in \mathbb{N}$ almost surely verifies*

$$\forall x \in X, \qquad p^*(x) = 0 \quad \Rightarrow \quad \lim_{\tau \to \infty} \tilde{n}_\tau(x) = 0.$$

The proof relies mainly on the study of a well-chosen function of the state of the urn. For integers $0 < a < b$, let $\mathrm{LD}[a, b]$ denotes the following discrete approximation of $\log(\frac{a}{b})$:

$$\mathrm{LD}[a, b] = -\sum_{i=a}^{b-1} \frac{1}{i}. \tag{10}$$

Recall that at time $\tau \in \mathbb{N}$, $n_\tau(w)$ denotes the number of w-balls in the urn. Let A denote the number of balls in the initial urn n_0. The total number of balls is increasing by 1 at each time, so $\sum_w n_\tau(w) = A + \tau$. The probability of drawing a w-ball is $\tilde{n}_\tau(w) = n_\tau(w)/(A + \tau)$. Consider the quantity

$$\mu_\tau = \sum_{w \in X} \mathrm{LD}\big[n_\tau(w), A + \tau\big] \cdot p^*(w), \tag{11}$$

that is the expected value, according to the optimal probability p^*, of the discrete logarithm at time τ.

PROPOSITION 6. *For both Two-alternative and Three-alternative reinforcement, the sequence μ_τ, $\tau \in \mathbb{N}$ is a negative submartingale. More precisely we have, for Two-alternative reinforcement,*

$$\mathbb{E}[\mu_{\tau+1} - \mu_\tau \mid n_\tau] = \frac{g(p^*, \tilde{n}_\tau)}{A + \tau},$$

and for Three-alternative reinforcement,

$$\begin{aligned}\mathbb{E}[\mu_{\tau+1} - \mu_\tau \mid n_\tau] &= \frac{1}{A + \tau}\Bigg(g\big(p^*, \tilde{n}\big) + \frac{1}{2}\sum_{w \in X} g(w, \tilde{n})^2 p^*(w) \\ &\quad + \frac{1}{2}\sum_{v} \tilde{n}(v) g\big(p^*, v\big)\big(1 + g(v, \tilde{n})\big)\Bigg).\end{aligned}$$

PROOF. We will write p for $\tilde{n}_\tau$ and let i denote either 1 or 2. From τ to $\tau+1$, one and only one ball is added. This ball has type w with probability $p^{[i]}(w)$. Thus,

$$\begin{aligned}
&\mathbb{E}[\mu_{\tau+1} - \mu_\tau \mid n_\tau] \\
&\quad = -\sum_{w\in X} p^{[i]}(w)\Bigg(\sum_{v\neq w} \frac{1}{A+\tau}\cdot p^*(v) + \left[\frac{1}{A+\tau} - \frac{1}{n_\tau(w)}\right]\cdot p^*(w)\Bigg) \\
&\quad = \frac{-1}{A+\tau} + \sum_{w\in X} p^{[i]}(w)\frac{1}{n_\tau(w)}p^*(w) \\
&\quad = \frac{-1}{A+\tau} + \sum_{w\in X} \frac{p^{[i]}(w)}{p(w)}\frac{p^*(w)}{A+\tau},
\end{aligned}$$

where, in the last line, we used the definition $p(w) = n_\tau(w)/(A+\tau)$. Using the formula for $p^{[i]}$ of Lemma 2, it comes, for Two-alternative,

$$\begin{aligned}
&\mathbb{E}[\mu_{\tau+1} - \mu_\tau \mid n_\tau] \\
&\quad = \frac{-1}{A+\tau} + \sum_{w\in X} \big(1 + g(w,p)\big)\frac{p^*(w)}{A+\tau} \\
&\quad = \frac{g(p^*,p)}{A+\tau},
\end{aligned}$$

which is always nonnegative, thanks to the definition of p^*. Furthermore, $g(p^*,p) = 0$ implies, by the first part of Proposition 1, that $\mathrm{Supp}(p) \subseteq \mathrm{Supp}(p^*)$.

For Three-alternative, we have

$$\begin{aligned}
&\mathbb{E}[\mu_{\tau+1} - \mu_\tau \mid n_\tau] \\
&\quad = \frac{-1}{A+\tau} + \sum_{w\in X}\Bigg(1 + \frac{3}{2}g(w,p) + \frac{1}{2}g(w,p)^2 \\
&\qquad + \frac{1}{2}\sum_v p(v)g(w,v)g(v,p)\Bigg)\frac{p^*(w)}{A+\tau}.
\end{aligned}$$

One can rearrange

$$\begin{aligned}
&(A+\tau)\mathbb{E}[\mu_{\tau+1} - \mu_\tau \mid n_\tau] \\
&\quad = \frac{3}{2}g\big(p^*,p\big) + \frac{1}{2}\sum_{w\in X} g(w,p)^2 p^*(w) + \frac{1}{2}\sum_v p(v)g\big(p^*,v\big)g(v,p) \\
&\quad = g\big(p^*,p\big) + \frac{1}{2}\sum_{w\in X} g(w,p)^2 p^*(w) + \frac{1}{2}\sum_v p(v)g\big(p^*,v\big)\big(1+g(v,p)\big),
\end{aligned}$$

and all the terms in this sum are nonnegative. The sum can be 0 only if both $\mathrm{Supp}(p) \subseteq \mathrm{Supp}(p^*)$, and $g(w,p) = 0$ for all w in the support of p^*, which implies $p = p^*$ by the uniqueness in Proposition 1. □

We are now able to prove the two results of the beginning of the section.

PROOF OF THEOREMS 4 AND 5. We consider for this proof either Two- or Three-alternative reinforcement. We have

$$\mathbb{E}[\mu_\tau] = \mathbb{E}[\mu_0] + \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbb{E}[\mu_t - \mu_{t-1} \,\big|\, \mu_{t-1}]\right]. \tag{12}$$

By Proposition 6, μ_τ is a negative submartingale, so it converges almost surely to an integrable random variable μ_∞ (see Corollary VII.4.1 and VII.4.2 in [27]). Furthermore, the left-hand side is an increasing negative sequence so it converges to a finite value. In the right-hand side, the sum is an increasing sequence of positive random variables so by the monotonous convergence theorem (Theorem II.6.1 in [27]) we can take the limit inside the expectation. Hence, $\mathbb{E}[\sum_{t=1}^{\infty} \mathbb{E}[\mu_t - \mu_{t-1} \mid \mu_{t-1}]] = \lim \mathbb{E}[\mu_\tau] - \mathbb{E}[\mu_0]$ is finite and so $\sum_{t=1}^{\infty} \mathbb{E}[\mu_t - \mu_{t-1} \mid \mu_{t-1}]$ is almost surely finite.

Let $f^{[1]}(p) = g(p^*, p)$ and $f^{[2]}(p) = g(p, p^*) + \sum_{w \in X} g(w, p)^2 p^*(x)$. The simplex $\Delta(X)$ is embedded in $\mathbb{R}^X$ so we use the L^∞ distance on it. With this distance, $f^{[i]}$ is continuous and $d(\tilde{n}_\tau, \tilde{n}_{\tau+1}) \le \frac{1}{A+\tau}$ almost surely. Denote by $B(p, \eta)$ the ball of center p and radius η.

Now consider a single realization of the urn process. Since X is a finite set, $\Delta(X)$ is compact, so let $\tilde{n}_\infty$ be an accumulation point for $\tilde{n}_\tau$. Looking for a contradiction, suppose $f^{[i]}(\tilde{n}_\infty) > 0$.

Since $f^{[i]}$ is continuous, let $\varepsilon, \eta > 0$ be such that $\forall p \in B(\tilde{n}_\infty, \eta)$, $f^{[i]}(p) > \varepsilon$. Let $(\phi(k))_{k \ge 0}$ be a subsequence such that $\forall k, \tilde{n}_{\phi(k)} \in B(\tilde{n}_\infty, \eta/2)$ and $\phi(k+1) > (1+\eta)\phi(k)$, we have

$$\sum_{t=0}^{\infty} \mathbb{E}[\mu_{t+1} - \mu_t \mid n_t] \ge \sum_{t=0}^{\infty} \frac{1}{2(A+t)} f^{[i]}(\tilde{n}_t) \tag{13}$$

$$\ge \sum_{k=0}^{\infty} \frac{1}{2(A+\phi(k))} \sum_{t=\phi(k)}^{\lfloor (1+\eta/2)\phi(k) \rfloor} f^{[i]}(\tilde{n}_t) \tag{14}$$

$$\ge \sum_{k=0}^{\infty} \frac{1}{2(A+\phi(k))} \left\lfloor \frac{\eta}{2} \phi(k) \right\rfloor \varepsilon. \tag{15}$$

The right-hand side of the last line is infinite. We already proved that the left-hand side is almost surely finite. It follows that $f^{[i]}(\tilde{n}_\tau) \to 0$ almost surely. We have seen in the proof of Proposition 6 that this fact implies exactly the theorems. □

3.4. *Two-alternative reinforcement and variance estimates.* In this section, we study in detail the Two-alternative reinforcement. The results are gathered in the next theorem, whose first item was already stated (in Theorem 5). The main idea is to study the variance of μ_∞ conditionally on the state at a large time t.

THEOREM 7. *For any initial urn $n_0 \in \mathcal{N}$, the random urn sequence obtained by Two-alternative reinforcement is such that*:

1. *almost surely, for all alternatives x such that $p^*(x) = 0$, $\tilde{n}_\tau(x) \to 0$ when $\tau \to \infty$;*
2. *with positive probability, the realized sequence $\tilde{n}_\tau, \tau \in \mathbb{N}$ has no limit as $\tau \to \infty$;*
3. *if T is such that $\forall x$, $p^*(x) > 0$ [in other words, $\mathrm{BP}(T) = X$] then, with probability one, $\tilde{n}_\tau, \tau \in \mathbb{N}$ has no limit.*

To simplify notation, in this section we let the process start at $\tau \neq 0$ so that τ always denote the number of ball in the urn (i.e., $A = 0$). We will also only consider Two-alternative reinforcement in this section. Recall the piece of notation $\mathrm{Supp}(p^*) = \mathrm{BP}(T)$; for ease of notation, we will often drop the argument. Also recall from the last section that μ_τ is a negative submartingale so it has an almost sure limit μ_∞. Let $\phi = \sum_{x \in \mathrm{BP}} p^*(x) \log p^*(x)$ and note that it is the value of μ_∞ if $\tilde{n}_\tau$ converges to p^*.

The first point is the following variance estimate.

LEMMA 8. *Let $\tau > 0$ and let $\varepsilon_\tau(x) = p^*(x)/\tilde{n}_\tau(x) - 1$. We have*

$$\mathbb{E}\big[(\mu_{\tau+1} - \mu_\tau)^2 \mid \mathcal{F}_\tau\big] = \frac{1}{\tau^2} \sum_{x \in \mathrm{BP}} \tilde{n}^{[1]}(x) \varepsilon_\tau(x)^2.$$

PROOF. This is a straightforward computation:

$$\begin{aligned}
\mathbb{E}\big[(\mu_{\tau+1} - \mu_\tau)^2 \mid \mathcal{F}_\tau\big] &= \sum_x \tilde{n}_\tau^{[1]}(x) \left(- \sum_y p^*(y) \frac{1}{\tau} + p^*(x) \frac{1}{n(x)} \right)^2 \\
&= \sum_x \tilde{n}_\tau^{[1]}(x) \frac{1}{\tau^2} \left(-1 + \frac{p^*(x)}{\tilde{n}(x)} \right)^2 \\
&= \frac{1}{\tau^2} \sum_x \tilde{n}^{[1]}(x) \varepsilon_\tau(x)^2.
\end{aligned}$$

□

The key point is that the factor $\frac{1}{\tau^2}$ makes the sum finite (once some control on ε is provided). Thus, the variance of μ_∞ conditioned on $\mathcal{F}_\tau$ will be of order ε^2/τ and, therefore, with hight probability μ_∞ will be close to μ_τ. If μ_τ is far enough from ϕ, then it follows that $\mu_\infty \neq \phi$.

We will first consider the case where $\mathrm{BP} \neq X$. In this case, we have $\mathbb{E}[\mu_{\tau+1} - \mu_\tau \mid \mathcal{F}_\tau] = \frac{g(p^*, \tilde{n}_\tau)}{\tau} \geq \frac{1}{\tau} g_0 \cdot \tilde{n}(\mathrm{BP}^c) > 0$ [where $g_0 = \inf_{y \in \mathrm{BP}^c} g(p^*, y)$] so we need an estimate of $\tilde{n}(\mathrm{BP})$.

LEMMA 9. *Suppose that there exists $q \in (0,1)$ and τ_0 such that, at each time $\tau \geq \tau_0$, the probability of adding a ball in BP^c is at most $q \cdot \tilde{n}_\tau(\mathrm{BP}^c)$. Then we have, for all $\tau \geq \tau_0$,*

$$\mathbb{E}\big[\tilde{n}_\tau(\mathrm{BP}^c) \mid \mathcal{F}_{\tau_0}\big] \leq \left(\prod_{t=\tau_0+1}^{\tau} \frac{t+q-1}{t}\right)\tilde{n}_{\tau_0}(\mathrm{BP}^c) \leq \left(\frac{\tau_0}{\tau}\right)^{1-q}\tilde{n}_{\tau_0}(\mathrm{BP}^c).$$

PROOF. The first line comes from a straightforward induction:

$$\begin{aligned}\mathbb{E}\big[\tilde{n}_\tau(\mathrm{BP}^c) \mid \mathcal{F}_{\tau_0}\big] &= \frac{1}{\tau}\mathbb{E}\big[n_{\tau-1}(\mathrm{BP}^c) \mid \mathcal{F}_{\tau_0}\big] + \frac{1}{\tau}\mathbb{E}\big[n_\tau(\mathrm{BP}^c) - n_{\tau-1}(\mathrm{BP}^c) \mid \mathcal{F}_{\tau_0}\big] \\ &\leq \frac{\tau-1}{\tau}\mathbb{E}\big[\tilde{n}_{\tau-1}(\mathrm{BP}^c) \mid \mathcal{F}_{\tau_0}\big] + \frac{1}{\tau}\mathbb{E}\big[q\tilde{n}_{\tau-1}(\mathrm{BP}^c) \mid \mathcal{F}_{\tau_0}\big] \\ &\leq \frac{\tau+q-1}{\tau}\mathbb{E}\big[\tilde{n}_{\tau-1}(\mathrm{BP}^c) \mid \mathcal{F}_{\tau_0}\big].\end{aligned}$$

For the second line, we have

$$\begin{aligned}\log \prod_{t=\tau_0+1}^{\tau} \frac{t+q-1}{t} &= \sum_{\tau_0+1}^{\tau} \log\left(1 + \frac{q-1}{t}\right) \\ &\leq \sum_{\tau_0+1}^{\tau} \frac{q-1}{t} \\ &\leq (q-1)\log\left(\frac{\tau}{\tau_0}\right).\end{aligned}$$

□

Now we are able to prove the second point of the theorem.

PROOF OF THEOREM 7, CASE $\mathrm{BP}(T) \neq X$. Note that for $\delta > 0$ small enough and τ_0 big enough, the set

$$S = \left\{n \in \mathcal{N} \mid n(X) \geq \tau_0 \text{ and } \left|\phi - \sum_x p^*(x)\mathrm{LD}\big(n(x), n(X)\big)\right| \leq \delta\right\}$$

verifies:

- $\forall n \in S, \forall x \in \mathrm{BP}, |p^*(x)/\tilde{n}(x) - 1| \leq 1$,
- $\forall n \in S, \forall x \in \mathrm{BP}^c, 1 + g(x, \tilde{n}) \leq q$

for some $q \in (1 - \inf_{x \in \mathrm{BP}^c} g(p^*, x), 1)$.

Let us assume that $\tilde{n}_{\tau_0} \in S$ (which clearly happens with positive probability). Let T be the first time after τ_0 such that $\tilde{n}_\tau \notin S$; T is a stopping time so $\mu_{\tau \wedge T}$ is

still a submartingale, let us call it μ'_τ. Note that we have

$$\begin{aligned}\mathbb{E}[\mu_{\tau+1} - \mu_\tau \mid \mathcal{F}_\tau] &= \frac{g(p^*, \tilde{n}_\tau)}{\tau} \\ &\leq \frac{\tilde{n}_\tau(\mathrm{BP}^c)}{\tau}.\end{aligned}$$

By definition of the set S, up to time T, we can apply Lemma 9 so

$$\begin{aligned}\mathbb{E}\big[\mu'_{\tau+1} - \mu'_\tau \mid \mathcal{F}_{\tau_0}\big] &\leq \frac{1}{\tau}\mathbb{E}\big[\tilde{n}_\tau(\mathrm{BP}^c)\mathbf{1}_{\{\tau<T\}} \mid \mathcal{F}_{\tau_0}\big] \\ &\leq \tau_0^{1-q}\tau^{q-2}\tilde{n}_{\tau_0}(\mathrm{BP}^c).\end{aligned}$$

Now let μ'_∞ denote the almost sure limit of μ'. Note that if $|\phi - \mu'_\infty|(\omega) < \delta$ then $T(\omega) = \infty$, and thus $\mu_\infty(\omega) = \mu'_\infty(\omega)$. It is therefore enough to show that with positive probability $\phi - \delta < \mu'_\infty < \phi$.

Note that μ' is a bounded submartingale, so it also converges in L^1 and L^2 toward μ'_∞, thus

$$\begin{aligned}\mathbb{E}\big[\mu'_\infty - \mu'_{\tau_0} \mid \mathcal{F}_{\tau_0}\big] &= \sum_{t=\tau_0}^{\infty}\mathbb{E}\big[\mu'_{t+1} - \mu'_t \mid \mathcal{F}_{\tau_0}\big] \\ &\leq \sum_{t=\tau_0}^{\infty}\tau_0^{1-q}t^{q-2}\tilde{n}_{\tau_0}(\mathrm{BP}^c) \\ &\leq C\tilde{n}_{\tau_0}(\mathrm{BP}^c),\end{aligned}$$

and, putting together the variance formula of Lemma 8 and the bound $\varepsilon \leq 1$ up to time T,

$$\begin{aligned}\mathrm{Var}(\mu'_\infty \mid \mathcal{F}_{\tau_0}) &\leq \sum_{t=\tau_0}^{\infty}\mathbb{E}\big[(\mu'_{t+1} - \mu'_t)^2 \mid \mathcal{F}_{\tau_0}\big] \\ &\leq \sum_{t=\tau_0}^{\infty}\frac{1}{t^2}\mathbb{E}\Big[\sum_{x\in\mathrm{BP}}\tilde{n}^{[1]}(x)\varepsilon_t(x)^2\mathbf{1}_{t<T} \mid \mathcal{F}_{\tau_0}\Big] \\ &\leq C\frac{1}{\tau_0}.\end{aligned}$$

Finally, consider a τ_0 large enough so that $\frac{1}{\tau_0} \ll \delta$. It is clear that with a positive probability, $\tilde{n}_{\tau_0}(\mathrm{BP}^c) \ll \delta$ and $|\mu_{\tau_0} - \phi|$ is close to $\delta/2$. On this event, we see that $|\phi - \mu'_\infty|$ is a random variable with expectation close to $\delta/2$ and variance small with respect to δ, so $|\mu'_\infty - \phi|$ has a positive probability to be in $[\delta/4, 3\delta/4]$, which proves the theorem. □

Now we turn to the case where $X = \mathrm{BP}(T)$. The idea will be similar, with Lemma 8 being the core argument. The main simplification comes from the fact

that, in this case, $g(p^*, p) = 0$ for all p, so μ is a martingale and Lemma 9 will no longer be needed. However, in order to prove that μ_∞ is almost surely different from ϕ, we will need an almost sure lower bound on $|\phi - \mu_\tau|$ which will come from a precise analysis of the difference between LD and the real logarithm. Finally, since the almost sure bound that we will get will be much worse than the one we were able to have with positive probability, we will need to be more careful in our use of Lemma 8.

First, recall the following well-known approximation result.

PROPOSITION 10. *Let $k > 0$. There exists a constant γ (Euler's constant) such that*

$$\log(k+1) + \gamma - \frac{1}{2}\sum_{i=k+1}^{\infty}\frac{1}{i^2} \le \sum_{i=1}^{k}\frac{1}{i} \le \log(k+1) + \gamma - \frac{1}{2}\sum_{i=k+2}^{\infty}\frac{1}{i^2}. \tag{16}$$

Furthermore, when k tends to infinity,

$$\sum_{i=k+1}^{\infty}\frac{1}{i^2} \sim \frac{1}{k}. \tag{17}$$

This proposition implies the following corollary.

COROLLARY 11. *Let T be a tournament on the set X such that $\mathrm{BP}(T) = X$. There exists $c > 0$ such that, for any urn $n \in \mathcal{N}$,*

$$\sum_{x} p^*(x)\mathrm{LD}\big(n(x), n(X)\big) \le \phi - \frac{c}{n(X)}. \tag{18}$$

Furthermore, writing $\varepsilon(x) = p^*(x)/\tilde{n}(x) - 1$, if we restrict ourselves to urns with $n(X)$ large enough (with ε staying bounded), the constant c can be taken as close as we want to

$$\frac{|X| - 1 + \sum_{x\in\mathrm{BP}}\varepsilon(x)}{2}. \tag{19}$$

PROOF. This is a straightforward computation using the definition of LD and Proposition 10. □

We also need a control of ε in term of μ.

LEMMA 12. *Almost surely, for any time τ,*

$$\sum_{x\in\mathrm{BP}}\big[\tilde{n}_\tau(x) + p^*(x)/2\big]\varepsilon_\tau(x)^2 \le \phi - \mu_\tau. \tag{20}$$

PROOF. We have, using Proposition 10 in the first line,

$$\phi - \mu_\tau \geq \sum_{x \in \mathrm{BP}} p^*(x)\big[\log p^*(x) - \log \tilde{n}_\tau(x)\big] \tag{21}$$

$$\geq \sum_{x \in \mathrm{BP}} p^*(x)\big(\varepsilon_\tau(x) + \varepsilon_\tau(x)^2/2\big) \tag{22}$$

$$\geq \sum_{x \in \mathrm{BP}} \big(p^*(x)/2 + \tilde{n}_\tau(x)\big)\varepsilon_\tau(x)^2 \tag{23}$$

by definition of ε. $\square$

Together, Lemma 8 and Lemma 12 have the following consequence.

LEMMA 13. *For any τ_0 large enough, if $|\phi - \mu_{\tau_0}| \geq \frac{5}{\tau_0 - 1}$ then*

$$\mathbb{P}\left(\forall \tau \geq \tau_0, |\phi - \mu_\tau| > \frac{1}{\tau_0}\right) \geq \frac{1}{10}.$$

PROOF. Let $d = |\phi - \mu_{\tau_0}|$ and let

$$T = \inf\{\tau > \tau_0 \mid \phi - \mu_\tau \geq 2d \text{ or } \phi - \mu_\tau \leq d/5\}.$$

T is a stopping time so $\mu'_\tau = \mu_{\tau \wedge T}$ is still a submartingale; by definition, it is also bounded so it converges almost surely and in all L^p. Let μ'_∞ denote its limit. We will show that $\mathbb{P}(\mu'_\infty \in (d/5, 2d)) > 1/10$. Note that, since μ makes vanishing steps, we are only interested in the behavior of μ close to ϕ, so we can restrict ourself to urns with a small ε.

As long as $\tau < T$, by definition we have $\phi - \mu_\tau \leq 2d$ and thus by Lemma 12, $\sum_x (p^*(x)/2 + \tilde{n}_\tau(x))\varepsilon_\tau(x)^2 \leq 2d$. Now note that for ε small, we have $\tilde{n}_\tau \simeq p^*$ and thus $\tilde{n}_\tau^{[1]} \simeq p^{*[1]} = p^*$. In particular, for any x, $p^*(x)/2 + \tilde{n}_\tau(x) \simeq \frac{3}{2}\tilde{n}_\tau^{[1]}(x)$. Therefore, for ε small enough, we have $p^*(x)/2 + \tilde{n}_\tau(x) > \tilde{n}_\tau^{[1]}(x)$ and thus $\sum_x \tilde{n}_\tau^{[1]}(x)\varepsilon_\tau(x)^2 \leq 2d$.

We can then use Lemma 8 to get

$$\forall \tau > \tau_0, \qquad \mathbb{E}\big[(\mu'_{t+1} - \mu'_t)^2 \mid \mathcal{F}\big] \leq \frac{2d}{\tau^2}. \tag{24}$$

Summing up to infinity (recall that μ' has constant expectation):

$$\mathrm{Var}(\mu'_\infty - \mu'_{\tau_0} \mid \mathcal{F}_{\tau_0}) = \sum_{\tau=\tau_0}^{\infty} \mathbb{E}\big[(\mu'_{\tau+1} - \mu'_\tau)^2 \mid \mathcal{F}_{\tau_0}\big] \tag{25}$$

$$\leq \sum_{\tau=\tau_0}^{\infty} \frac{2d}{\tau^2} \tag{26}$$

$$\leq 2d \frac{1}{\tau_0 - 1} \tag{27}$$

$$\leq \frac{2}{5}d^2, \tag{28}$$

where we used the hypothesis $d \geq \frac{5}{\tau_0 - 1}$ in the last line.

Note that, since $\mu_{T\wedge\tau}$ is bounded, $\mathbb{E}[\mu'_\infty] = d$. Moreover, notice that any random variable with expectation d and that never takes value in the interval $(d/2, 2d)$ has at least variance $d^2/2$. Since $\mathrm{Var}(\mu'_\infty \mid \mathcal{F}_\tau) \leq \frac{2}{5}d^2$, we have $\mathbb{P}(\mu' \in (d/2, 2d)) \geq 1/10$ and on this event, $T = \infty$, so $|\phi - \mu|$ has never reached $d/2 \geq \frac{1}{\tau_0}$. □

PROOF OF THEOREM 7, CASE $\mathrm{BP}(T) = X$. First, note that if $\tilde{n}_\tau$ converges, it has to be toward a fixed point. By Proposition 3, the fixed points of the dynamics are exactly the optimal strategies p^*_Y corresponding to all subtournaments $Y \subseteq X$. (This, of course, includes $p^* = p^*_X$ itself.)

We start by excluding convergence to any optimal strategy different from p^*_X. Let p^*_Y be such a probability measure, we claim that there exists some $x \in \mathrm{BP}$ such that $p^*_Y(x) = 0$. Indeed by contradiction, if $\mathrm{BP} \subset \mathrm{Supp}(p^*_Y)$ then by the first part of Proposition 1, $\forall x \in \mathrm{BP}$, $g(p^*_Y, x) = 0$ and in particular $g(p^*_Y, p^*_X) = 0$. By the second part of Proposition 1, this implies $\mathrm{Supp}(p^*_Y) = \mathrm{BP}$ and, therefore, $p^*_Y = p^*_X$ by uniqueness of the optimal strategy on BP. Now note that existence of $x \in \mathrm{BP}$ such that $p^*_Y(x) = 0$ implies that $\sum_x p^*_X(x) \log p^*_Y(x) = -\infty$. In particular, we see that any realisation of the urn such that $\tilde{n}_\tau \to p^*_Y$ has to satisfy $\mu_\tau \to -\infty$ but this is an event of probability 0 by the Markov inequality.

Let us now rule out convergence to p^*_X. We first consider the case $|X| \geq 12$. Then by Corollary 11, for any τ large enough, μ_τ almost surely verifies the hypothesis of Lemma 13. Fix any suitable τ_0, by Lemma 13, $\mathbb{P}(\tilde{n}_\tau \to p^*) \leq \mathbb{P}(\exists \tau \geq \tau_0 \mid |\phi - \mu_\tau| \leq 1/\tau_0) \leq 9/10$. On the event that $|\phi - \mu_\tau|$ does reach $1/\tau_0$ at some time τ_1, we can use Lemma 13 at time τ_1 to get $\mathbb{P}(\tilde{n}_\tau \to p^*) \leq (9/10)^2$. By induction, we get $\mathbb{P}(\tilde{n}_\tau \to p^*) = 0$ which proves the theorem.

For the case $3 \leq |X| \leq 11$ (a nontrivial tournament has at least 3 elements), consider any τ_0 large enough. By Corollary 11, we have $|\phi - \mu_{\tau_0}| \geq 1/\tau_0$. Let

$$T = \inf\left\{\tau > \tau_0 \mid |\phi - \mu_\tau| \leq \frac{1}{2\tau_0} \text{ or } |\phi - \mu_\tau| \geq \frac{5}{\tau_0}\right\}.$$

The event $\{T = \infty \text{ or } |\phi - \mu_T| \geq \frac{5}{\tau_0}\}$ has probability at least $1/9$ and if $|\phi - \mu_T| \geq \frac{5}{\tau_0}$ we can apply Lemma 13 at time T so the conclusion of Lemma 13 is still true with $1/\tau_0$ replaced by $1/2\tau_0$ and $1/10$ replaced by $1/90$. We can therefore use the same induction as before to prove the theorem. □

3.5. *Conclusion.* We have found the behavior of learning process designed to discover the "best" alternatives in a tournament. Learning is achieved through the following idea. The alternative that is considered as "good" at some date is reinforced for the future in the sense that one (slightly, and less and less) increases the

probability for this alternative to be considered: reinforcement updates the sampling, or "prior" probability. The test according to which an alternative is considered as a good one at time t rests on comparing a few randomly chosen alternatives.

We found a very different behavior between the processes where reinforcement occurs after sampling two or three alternatives. With three alternatives, the process converges almost surely to a well-defined limit that has a nice interpretation in term of the tournament game: it is the optimal strategy for this game. One can therefore say that this form of learning is "successful". With two alternatives, the picture is more complicated. The learning process "succeeds" in finding the bipartisan set (a set which has been argued to be more important in term of social choice than the numerical values of the optimal probabilities [15]), but not the optimal probabilities themselves. We conjecture that the almost sure nonconvergence happens for all tournaments and not only when $\mathrm{BP}(T) = X$.

Acknowledgement. Thanks to Bastien Mallein for useful discussions in the early stage of this study.

REFERENCES

[1] BRANDT, F., CHUDNOVSKY, M., KIM, I., LIU, G., NORIN, S., SCOTT, A., SEYMOUR, P. and THOMASSÉ, S. (2013). A counterexample to a conjecture of Schwartz. *Soc. Choice Welf.* **40** 739–743. MR3018395

[2] BRANDT, F. and FISCHER, F. (2008). Computing the minimal covering set. *Math. Social Sci.* **56** 254–268. MR2442206

[3] CHEBOTAREV, P. T. and SHAMIS, E. (1998). Characterizations of scoring methods for preference aggregation. *Ann. Oper. Res.* **80** 299–332.

[4] CHEBOTAREV, P. YU. and SHAMIS, E. (2006). Preference fusion when the number of alternatives exeeds two: Indirect scoring procedures. Preprint. Available at arXiv:math/060217v3 [math.OC].

[5] DANIELS, H. E. (1969). Round-Robin tournament scores. *Biometrika* **56** 295–299.

[6] DAVID, H. A. (1963). *The Method of Paired Comparisons*, Charles Griffin, London.

[7] FISHBURN, P. C. (1977). Condorcet social choice functions. *SIAM J. Appl. Math.* **33** 469–489. MR0449470

[8] FISHER, D. and RYAN, J. (1992). Optimal strategies for a generalized 'scissors, paper and stone' game. *Amer. Math. Monthly* **99** 935–942.

[9] FISHER, D. C. and REEVES, R. B. (1995). Optimal strategies for random tournament games. *Linear Algebra Appl.* **217** 83–85.

[10] FISHER, D. C. and RYAN, J. (1995). Tournament games and positive tournaments. *J. Graph Theory* **19** 217–236. MR1315439

[11] FISHER, D. C. and RYAN, J. (1995). Probabilities within optimal strategies for tournament games. *Discrete Appl. Math.* **56** 87–91. MR1311308

[12] HOFBAUER, J. and SCHLAG, K. (2000). Sophisticated imitation in cyclic games. *J. Evol. Econ.* **10** 523–543.

[13] LAFFOND, G., LASLIER, J.-F. and LE BRETON, M. (1993). The bipartisan set of a tournament game. *Games Econom. Behav.* **5** 182–201.

[14] LASLIER, J.-F. (1997). *Tournament Solutions and Majority Voting*. *Studies in Economic Theory* **7**. Springer, Berlin. MR1468987

[15] LASLIER, J.-F. (2000). Aggregation of preferences with a variable set of alternatives. *Soc. Choice Welf.* **17** 269–282. MR1746608

[16] LASLIER, J.-F. (2000). Interpretation of electoral mixed strategies. *Soc. Choice Welf.* **17** 283–292. MR1746609

[17] LEVCHENKOV, V. S. (1992). Social choice theory: A new insight. Discussion paper, Institute of Systems Analysis, Moscow.

[18] MCKELVEY, R. (1979). General conditions for global intransitivities in a formal voting model. *Econometrica* **47** 1085–1112.

[19] MOON, J. W. (1968). *Topics on Tournaments*. Holt, Rinehart and Winston, New York. MR0256919

[20] MOULIN, H. (1986). Choosing from a tournament. *Soc. Choice Welf.* **3** 271–291.

[21] MYERSON, R. B. (1993). Incentives to cultivate favored minorities under alternative electoral systems. *Am. Polit. Sci. Rev.* **87** 856–869.

[22] MYERSON, R. B. (1995). Analysis of democratic institutions: Structure, conduct and performance. *J. Econ. Perspect.* **9** 77–89.

[23] PEMANTLE, R. (2007). A survey of random processes with reinforcement. *Probab. Surv.* **4** 1–79.

[24] POSCH, M. (1997). Cycling in a stochastic learning algorithm for normal form games. *J. Evol. Econ.* **7** 193–207.

[25] RIVEST, R. L. and SHEN, E. (2010). An optimal single-winner preferential voting system based on game theory. Available at http://people.csail.mit.edu/rivest/gt/latest_conf.pdf.

[26] RUBINSTEIN, A. (1996). Why are certain properties of binary relations relatively more common in natural language? *Econometrica* **64** 343–355. MR1375737

[27] SHIRYAEV, A. N. (1995). *Probability. Graduate Text in Mathematics*, Springer, New York. MR3467826

[28] SLUTZKI, G. and VOLIJ, O. (2006). Scoring of web pages and tournaments—axiomatizations. *Soc. Choice Welf.* **26** 75–92.

[29] USHAKOV, I. A. (1976). The problem of choosing the preferred element: An application to sport games. In *Management Science in Sports* (R. E. Machol, S. P. Ladany and D. G. Morrison, eds.) 153–161. North-Holland, Amsterdam.

UNIVERSITÉ PARIS DIDEROT
BÂTIMENT SOPHIE GERMAIN
AVENUE DE FRANCE
75013 PARIS
FRANCE
E-MAIL: laslier@math.univ-paris-diderot.fr

CNRS, PARIS SCHOOL OF ECONOMICS
ECOLE NORMALE SUPÉRIEURE
48 BD. JOURDAN
75014 PARIS
FRANCE
E-MAIL: Jean-Francois.Laslier@ens.fr