



Rationally Biased Learning

Michel de Lara

► To cite this version:

| Michel de Lara. Rationally Biased Learning. 2020. hal-01581982v2

HAL Id: hal-01581982

<https://hal.science/hal-01581982v2>

Preprint submitted on 23 Oct 2020 (v2), last revised 22 Mar 2022 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Rationally Biased Learning

Michel DE LARA

CERMICS, Ecole des Ponts, Marne-la-Vallée, France

Abstract

Humans display a tendency to pay more attention to bad outcomes, often in a disproportionate way relative to their statistical occurrence. They also display euphorism, as well as a preference for the current state of affairs (status quo bias). Based on the analysis of optimal solutions of infinite horizon stationary optimization problems under imperfect state observation, we show that such human perception and decision biases can be grounded in a form of rationality. We also provide conditions (boundaries) for their possible occurrence and an analysis of their robustness. Thus, biases can be the product of rational behavior.

Keywords: pessimism bias, status quo bias, euphorism bias, probability overestimation, optimal behavior, imperfect state information.

1. Introduction

When we perceive sounds, we overestimate the change in level of rising level tones relative to equivalent falling level tones (Neuhoff (1998)). When we assess pros and cons in decision making, we weigh losses more than gains (Kahneman and Tversky (1979)). We are more frightened by a snake or a spider than by a passing car or an electrical shuffle. Such human assessments are qualified of biases, because they depart from physical measurements or objective statistical estimates. Thus, there is “bias” when a behavior is not aligned with a given “rationality benchmark” (like expected utility theory), as documented in the “heuristics and biases” literature (Kahneman et al. (1982); Gilovich et al. (2002)).

However, if such biases are found consistently in human behavior, they must certainly have a reason. Some scholars (see (Gigerenzer (2004, 2008); Hutchinson and Gigerenzer (2005))) claim that those “so-called bias” were in

fact advantageous in the type of environment where our ancestors lived and thrived (ecological or, rather, evolutionary, validity (Barkow et al. (1992); Boutang and De Lara (2015, 2019))). In this conception, the benchmark should be a measure of fitness reflecting survival and reproduction abilities, and the “bias” can be explained in two ways.

- *Bias by mismatch.* The bias can result from a timelag: because the modern environment has departed so much from the environment in which natural selection had time to shape our minds, human behavior displays a mismatch (cars are objectively more dangerous than spiders in our modern environment).
- *Bias by design.* But the bias can be the feature of an optimal strategy where optimality is measured in fitness (the genes of those who accurately estimated the change in level of rising level tones have, more often than the “overestimators”, finished in the stomach of a predator).

This last conception of “bias by design” is reflected, for example, in (Haselton and Nettle (2006)). In their attempt to understand “how natural selection engineers psychological adaptations for judgment under uncertainty”, Haselton and Nettle consider an individual who has to decide between a safe option (one known payoff) and a risky option (known bad and good payoffs). They define a critical probability and observe that, if the bad outcome has higher probability, the (optimal) individual should avoid taking risks and select the safe option. The interesting point is that the critical probability is the ratio of the difference between good and safe payoffs over the difference between good and bad payoffs. As a consequence, the higher the latter difference, the more the individual should avoid taking risks. The general conclusion is nicely expressed by Martie G. Haselton (on her personal webpage) when she claims that “selection has led to adaptations that are biased by design and functioned to help ancestral humans avoid particularly costly errors” and that “when the costs of false positive and false negative errors were asymmetrical over evolutionary history, selection will have designed psychological adaptations biased in the direction of the less costly error”.¹ However, the

¹Such asymmetry in costs is manifest in the so-called *life-dinner* principle of Richard Dawkins — “The rabbit runs faster than the fox, because the rabbit is running for his life while the fox is only running for his dinner” — and can exert a strong selection pressure (Dawkins and Krebs (1979)). Neuroscientist Joseph LeDoux has a nice way to express

above analysis is performed under the so-called “error management theory”, that is, supposing known the probability of the bad outcome. What happens when the individual does not know *a priori* the objective probability driving the occurrence of a bad outcome?

In this paper, we consider the classical problem of a decision maker (DM) faced with a repeated choice between a certain option (one known safe payoff) and a risky option (two known risky payoff values, but unknown probability of each). Regarding uncertainty, we are thus in the so-called ambiguity setting (and not in the risk setting). Regarding payoffs, we suppose that the three payoffs are ranked in such a way that the safe one stands between the two risky ones; thus, the lowest (risky) payoff reflects a bad outcome. We will show that a rational decision maker — in the sense of maximizing expected discounted utility (where the mathematical expectation involves a prior on the unknown probabilities) — can exhibit a behavior displaying “euphorism” and status quo biases, as well as, under suitable conditions, the pessimistic erroneous assessment of the best objective option (which goes against the so-called “maximization”, which consists in choosing the lottery with the highest expected payoff) and an overestimation bias for the probability of the bad outcome. Thus, in some settings (detailed in the paper), it is quite rational to pay more attention to bad outcomes than to good ones, and to exaggerate their importance, even if one aggregates uncertainties by means of a (balanced, risk-neutral) mathematical expectation, and aggregates payoffs by summation.

The questions that we examine have been extensively discussed in the literature. Many approaches are empirical, bearing on experiments with human subjects having to choose between lotteries (Erev et al. (2017, 2010); Barron and Erev (2003); Hertwig et al. (2004)). The experimental settings are manifold: continuous or discrete probability distribution of payoffs, known or unknown payoffs, known or unknown payoffs probabilities (risk or ambiguity), one shot decisions from description (decisions under risk) or one shot decisions from experience or repeated decisions from experience, known or unknown experiment termination, known or unknown foregone payoffs (full or partial feedback), etc. (in (Erev et al. (2017)), the “space of experimental paradigms” is described with 11 dimensions). Often, behaviors of

“bias by design” in his book *The Emotional Brain*: “It is better to have treated a stick as a snake than not to have responded to a possible snake” ((LeDoux, 1996, p.166)).

participants are compared with the “rational maximization” behavior, and then are possibly interpreted as biases. For instance, deviations from maximization may suggest that the participants behave as if they overweight or underweight rare events (Erev et al. (2017)). Biases are also analyzed as contingent on the experimental setting. For instance, whether decisions are from description or from experience may change the nature of a bias, like the overweighting or underweighting of rare events (Erev et al. (2017, 2010); Hertwig and Erev (2009); Barron and Erev (2003)). The predictive power of models and heuristics is also challenged, with discussions on boundaries and competitions between heuristics for choice prediction (Erev et al. (2010, 2017)).

Some authors propose a specific mathematical model of human behavior — that is, a specific strategy on how information is searched for and processed as time goes on — and explore its consequences by means of numerical simulations and theoretical analysis (Denrell (2005); Denrell and Le Mens (2007)).

Finally, some authors specify an intertemporal objective function and a class of solutions (strategies), then try to find an optimal strategy (Le Mens and Denrell (2011)). Others explore the properties of suboptimal solutions, like in (Denrell (2007)) where the problem is framed as a finite horizon one-armed bandit problem, but there does not seem to be a mathematically formulated intertemporal criterion to be optimized, and the focus is on analyzing special strategies, like rolling horizon strategy assuming that the estimate of the decision maker follows Bayes rule.

Our approach is neither experimental nor simulation-based, but theoretical in the sense that we do not propose and test a possible human strategy, but that we put forward a sequential (repeated) decision-making framework, then make a theoretical analysis to deduce features of an optimal strategy. To the difference of (Le Mens and Denrell (2011)), who also propose an optimal sequential decision-making framework, we do not rely on numerical resolution to capture features of an optimal strategy but only on theory. Thus, if, like others, we formulate the problem as one of optimization and analyze optimal solutions in terms of human psychology, our results are less precise but we aim at greater generality. Let us explain how.

It is well-known that the problem we address can be framed as a two-armed bandit problem, as there are only two decisions, as information is triggered by decision, as the criterion is intertemporal under unknown probabilities. There is a huge literature on armed bandit problems, and on its

celebrated solution (when suitable hypothesis are met, see (Gittins (1979))) by means of a dynamic allocation index (Gittins Index Theorem), in the mathematics and the psychology literature. However, this not the route we follow. Indeed, we revisit this type of problem as an instance of optimization problem under imperfect state observation, and we devote a whole part to discuss which of our results are robust w.r.t. (with respect to) to assumptions like stationarity, discounting, finite or infinite horizon. By doing so, we want to reveal features of optimal strategies that are more general than those obtained by means of the Gittins index strategy. Of course, some features — for instance that the information needed for optimal decisions can be summarized in a posterior that is updated following Bayes rule — are shared with this latter, but they do not depend on the mathematical expression of the index.

To finish, let us stress that his paper is about sampling, and certainly about adaptive sampling (as information is triggered by decision), but not about “selective sampling”, like in (Denrell (2005); Denrell and Le Mens (2007)) who analyze “sampling selection bias”, as we do not address social influence, social psychology or strategic interactions. Let us also point out that our contribution resorts to economics, focusing on optimality benchmarks and resulting optimal strategies, whereas, in the evolutionary literature, the discussion rather bears on the precision of Bayesian estimates of the unknown probability (Trimmer et al. (2011)).

The paper is organized as follows. In Sect. 2, we consider the problem of a decision maker faced with a repeated choice between a certain option (one known safe payoff) and a risky option (yielding either a bad or a good outcome, but with unknown probabilities). We set up a formal mathematical model of stochastic sequential decision-making — under (Bayesian) ambiguity regarding random sequences of bad and good outcomes (Bernoulli trials) — and we describe an optimal strategy and the behavior of the optimal DM. This section contains known results, but with new proofs that make it possible to assess the robustness of the findings. In Sect. 3, we prove and display features of the optimal strategy — optimally designed for a Bayesian criterion, that is, for a certain (subjective) probability distribution on sequences of bad and good outcomes — when it is implemented with a Bernoulli process under objective probabilities (objective environment). We distinguish two outputs of the optimal strategy — estimation of the unknown objective probability of the bad outcome, assesement of whether the objective

environment is prone to risk-taking or not. When one output is not what it would be were the objective probability distribution known, we deal with a bias. We summarize in Tables 2 and 3 our findings regarding consistency or discrepancy w.r.t. what would be optimal in the objective environment: “euphorism” and status quo biases, as well as boundaries and amplitudes of two effects, the pessimistic erroneous assessment of the best objective option and the overestimation of the probability of the bad outcome. In Sect. 4, we discuss the cognitive burden of implementing the optimal strategy (hence the possibility to be an outcome of natural selection), the robustness of our findings and possible psychological interpretations, and we compare our results with the literature. Sect. 5 concludes and Appendix A gathers technical results.

2. Repeated decision-making under unknown probability

In §2.1, we lay out mathematical ingredients to set up a model of sequential decision-making under unknown probability, and formulate an expected discounted payoff maximization problem. In §2.2, we analyze the structure of an optimal strategy, and then describe the behavior of a decision-maker who adopts such optimal strategy.

The results of this Section are not new: the structure of an optimal strategy and the induced behavior are well-known, although they are generally presented as a consequence of the Gittins Index Theorem, which is not the way we prove them. Nevertheless, this Section fixes vocabulary, notation and main properties that will be used to show our main results in Sect. 3, and also to discuss, in §4.2, which of our results are robust w.r.t. to assumptions like stationarity, discounting, finite or infinite horizon.

2.1. An expected discounted payoff maximization problem

In (Haselton and Nettle (2006)), the following situation is examined. To reach her/his destination, an individual has two options: a short risky route passes through a grassy land — possibly hiding a poisonous snake inflicting serious (though non lethal) pains — whereas a safe route makes a long costly detour. Two decisions are possible, with different costs. If one avoids the grass, one makes a detour that is costly in time, but one suffers no pain from the (possible) snake. If one passes through the grass (“trying”, “learning”, “experimenting”), the time spent is shorter but one can suffer pain (though

not lethal) if the snake is present. We will illustrate our mathematical setting with this story.

Sequential decision-making

We consider two possible outcomes (states of Nature) — a bad one \mathbf{B} and a good one \mathbf{G} — that we illustrate by \mathbf{B} = “a snake is in the grass”, and by \mathbf{G} the contrary. We suppose that, at discrete stages $t \in \mathbb{N}$, the DM makes a decision — either “avoid” and be prudent (α) or “experiment” and take risks (ε) — without knowing in advance the state of Nature occurring at that time — either bad (\mathbf{B}) or good (\mathbf{G}). We denote by $t = 0, 1, 2 \dots$ the stage corresponding to the beginning of the time interval $[t, t + 1[$. We denote by $\{\alpha, \varepsilon\}$ the set of decisions, and by $v_t \in \{\alpha, \varepsilon\}$ the action taken by the DM at the beginning of the time interval $[t, t + 1[$. We define the *sample space*

$$\mathbb{H}_\infty = \{\mathbf{B}, \mathbf{G}\}^{\mathbb{N}^*} = \{\mathbf{B}, \mathbf{G}\} \times \{\mathbf{B}, \mathbf{G}\} \times \dots, \quad (1)$$

with generic element an infinite sequence (w_1, w_2, \dots) of elements in $\{\mathbf{B}, \mathbf{G}\}$. For $t = 1, 2 \dots$, we denote by²

$$\mathbf{W}_t : \mathbb{H}_\infty \rightarrow \{\mathbf{B}, \mathbf{G}\}, \quad \mathbf{W}_t(w_1, w_2, \dots) = w_t, \quad (2)$$

the *state of Nature* realized at the beginning of the time interval $[t, t + 1[$, but that cannot be revealed before the end of $[t, t + 1[$.

Strategies

At the beginning of each time interval $[t, t + 1[$, the DM can either “avoid” (decision α) — in which case the DM has no information about the state of Nature — or “experiment” (decision ε) — in which case the state of Nature \mathbf{W}_{t+1} (\mathbf{B} or \mathbf{G}) is revealed and experimented, at the end of the time interval $[t, t + 1[$.

We assume that the DM is not visionary and learns only from the past: she/he cannot know the future in advance, neither can the DM know the state of Nature (\mathbf{B} or \mathbf{G}) if the DM decides to avoid. We define the *observation sets* at stage $t = 0, 1, 2, 3 \dots$ by $\mathbb{Y}_0 = \{\partial\}$, where ∂ corresponds to no information (no observation at initial stage $t = 0$), and $\mathbb{Y}_t = \{\mathbf{B}, \mathbf{G}, \partial\}^t$ for $t = 1, 2, 3 \dots$. We define the *observation mapping* $\mathcal{O} : \{\alpha, \varepsilon\} \times \{\mathbf{B}, \mathbf{G}\} \rightarrow \{\mathbf{B}, \mathbf{G}, \partial\}$ by $\mathcal{O}(\varepsilon, \mathbf{B}) = \mathbf{B}$,

²We denote random variables by uppercase bold letters.

$\mathcal{O}(\varepsilon, \mathbf{G}) = \mathbf{G}$ and $\mathcal{O}(\alpha, \mathbf{B}) = \mathcal{O}(\alpha, \mathbf{G}) = \partial$. Thus, the observation at stage $t = 0, 1, 2 \dots$ if the DM makes decision $v_t \in \{\alpha, \varepsilon\}$ is $\mathbf{Y}_{t+1} = \mathcal{O}(v_t, \mathbf{W}_{t+1})$. This case is also known as the *partial feedback* case, where foregone payoffs are not revealed.

We allow the DM to accumulate past observations; therefore the decision v_t at stage t can only be a function of $\mathbf{Y}_1, \dots, \mathbf{Y}_t$ (the initial decision v_0 is made without information). A *policy at stage t* is a mapping $\mathcal{S}_t : \mathbb{Y}_t \rightarrow \{\alpha, \varepsilon\}$ that tells the DM what will be the next action in view of past observations. A *strategy* \mathcal{S} is a sequence $\mathcal{S} = (\mathcal{S}_0, \mathcal{S}_1, \dots)$ of policies. Given a strategy \mathcal{S} , decisions and observations are inductively given by

$$\mathbf{V}_0 = \mathcal{S}_0 \in \{\alpha, \varepsilon\}, \quad (3a)$$

$$\mathbf{Y}_{t+1} = \mathcal{O}(\mathbf{V}_t, \mathbf{W}_{t+1}) \in \{\mathbf{B}, \mathbf{G}, \partial\}, \quad \forall t = 0, 1, 2 \dots, \quad (3b)$$

$$\mathbf{V}_t = \mathcal{S}_t(\mathbf{Y}_1, \dots, \mathbf{Y}_t) \in \{\alpha, \varepsilon\}, \quad \forall t = 0, 1, 2 \dots. \quad (3c)$$

In the full feedback case, where foregone payoffs are revealed no matter what the decision made, we have $\mathbf{Y}_t = \mathbf{W}_t$, for $t = 0, 1, 2 \dots$.

Hypothesized probability

We introduce the one-dimensional simplex

$$\Sigma^1 = \{(p^{\mathbf{B}}, p^{\mathbf{G}}) \in \mathbb{R}^2 \mid p^{\mathbf{B}} \geq 0, p^{\mathbf{G}} \geq 0, p^{\mathbf{B}} + p^{\mathbf{G}} = 1\}. \quad (4)$$

The simplex Σ^1 is identified with the unit segment $[0, 1]$ by the mapping (measurable bijection with measurable inverse) $\Sigma^1 \ni (p^{\mathbf{B}}, p^{\mathbf{G}}) \mapsto p^{\mathbf{B}} \in [0, 1]$. For any $(p^{\mathbf{B}}, p^{\mathbf{G}}) \in \Sigma^1$, we denote by

$$\mathcal{B}(p^{\mathbf{B}}, p^{\mathbf{G}}) = \bigotimes_{t=0}^{\infty} (p^{\mathbf{B}} \delta_{\mathbf{B}} + p^{\mathbf{G}} \delta_{\mathbf{G}}) \quad (5)$$

the probability \mathbb{P} on the sample space \mathbb{H}_{∞} in (1) which makes the stochastic process $(\mathbf{W}_1, \mathbf{W}_2, \dots)$ of states of Nature, as in (2), a sequence of independent Bernoulli trials with marginals given by $\mathbb{P}\{\mathbf{W}_t = \mathbf{B}\} = p^{\mathbf{B}}$ and $\mathbb{P}\{\mathbf{W}_t = \mathbf{G}\} = p^{\mathbf{G}}$.

We suppose that the DM makes the assumption that the stochastic process $(\mathbf{W}_1, \mathbf{W}_2, \dots)$ is governed by $\mathcal{B}(p^{\mathbf{B}}, p^{\mathbf{G}})$, but that the DM does not know the probabilities $(p^{\mathbf{B}}, p^{\mathbf{G}})$. Moreover, we suppose that the DM is a Bayesian who makes the assumption that the unknown couple $(p^{\mathbf{B}}, p^{\mathbf{G}})$ is a random variable with a distribution $\bar{\pi}_0$ on the one-dimensional simplex Σ^1 in (4).

This is why we consider the extended sample space $\Sigma^1 \times \mathbb{H}_\infty = \Sigma^1 \times \{\mathbf{B}, \mathbf{G}\}^{\mathbb{N}^*}$ equipped with the probability distribution $\bar{\pi}_0(d(p^{\mathbf{B}}, p^{\mathbf{G}})) \otimes \mathcal{B}(p^{\mathbf{B}}, p^{\mathbf{G}})$, whose marginal distribution on the sample space \mathbb{H}_∞ in (1) we denote by $\mathbb{P}^{\bar{\pi}_0}$. Thus, for any measurable bounded function $g : \mathbb{H}_\infty \rightarrow \mathbb{R}$, we have that

$$\mathbb{E}^{\mathbb{P}^{\bar{\pi}_0}}[g] = \int_{\Sigma^1} \bar{\pi}_0(d(p^{\mathbf{B}}, p^{\mathbf{G}})) \mathbb{E}^{\mathcal{B}(p^{\mathbf{B}}, p^{\mathbf{G}})}[g] . \quad (6)$$

Instantaneous payoffs

Now, to compare strategies, we will make up a criterion, or an objective function for the DM. In an evolutionary interpretation, payoffs are measured in “fitness” unit, for instance “number of days alive” or “number of days in a reproductive state”, taken as proxies for the number of offspring. The payoffs depend both on the decision and on the state of Nature as in Table 1.

	bad state \mathbf{B}	good state \mathbf{G}
avoid α	avoidance payoff $U(\alpha, \mathbf{B}) = \mathcal{U}_\alpha$	avoidance payoff $U(\alpha, \mathbf{G}) = \mathcal{U}_\alpha$
experiment ε	low payoff $U(\varepsilon, \mathbf{B}) = \mathcal{U}^{\mathbf{B}}$	high payoff $U(\varepsilon, \mathbf{G}) = \mathcal{U}^{\mathbf{G}}$

Table 1: Instant payoffs according to decisions (rows avoid (α) or experiment (ε)) and states of Nature (columns bad \mathbf{B} or good \mathbf{G})

We assume that the payoffs attached to the couple (action, state) in Table 1 are ranked as follows:

$$\overbrace{U(\varepsilon, \mathbf{G}) = \mathcal{U}^{\mathbf{G}}}^{\text{high payoff}} > \underbrace{U(\alpha, \mathbf{B}) = U(\alpha, \mathbf{G}) = \mathcal{U}_\alpha}_{\text{avoidance (middle) payoff}} > \overbrace{U(\varepsilon, \mathbf{B}) = \mathcal{U}^{\mathbf{B}}}^{\text{low payoff}} . \quad (7)$$

In other words, avoiding yields more utility than a bad outcome but less than a good one.

Intertemporal criterion

As the payoffs in Table 1 are measured in “fitness”, we suppose that they are cumulative, like days in a healthy condition or number of offspring. This

is why we suppose that the DM can evaluate her/his lifetime performance using strategy \mathcal{S} by the discounted intertemporal payoff

$$j(\mathcal{S}, \mathbf{W}) = \sum_{t=0}^{+\infty} \rho^t U(\mathbf{V}_t, \mathbf{W}_{t+1}) , \quad (8)$$

where \mathbf{V}_t is given by (3). Beyond “fitness”, our analysis extends to the maximization of any objective function which can be expressed as an infinite sum over time of discounted payoffs.

The rationale behind using *discounted* intertemporal payoff is the following. Suppose that the DM makes decisions up to a random ultimate stage \mathbf{T} like, for instance, the DM’s lifetime (measured in number of decision stages). If we suppose that the random variable \mathbf{T} is independent of the randomness in the occurrence of a bad and good outcomes, and follows a (memoryless) Geometric distribution with values in $\{0, 1, 2, 3 \dots\}$, then it is easy to establish the equality $\sum_{t=0}^{+\infty} \rho^t U(\mathbf{V}_t, \mathbf{W}_{t+1}) = \mathbb{E}_{\mathbf{T}}[\sum_{t=0}^{\mathbf{T}} U(\mathbf{V}_t, \mathbf{W}_{t+1})]$, where the mathematical expectation $\mathbb{E}_{\mathbf{T}}$ is only w.r.t. the random variable \mathbf{T} . Then, we can interpret the discount factor $\rho \in [0, 1[$ in term of the expected value $\bar{\mathbf{T}}$ of the random number \mathbf{T} of stages during which the DM has to make decisions, by means of the equations $\bar{\mathbf{T}} = \rho/(1-\rho)$ and $\rho = \bar{\mathbf{T}}/(\bar{\mathbf{T}}+1)$. For instance, for an individual making daily decisions during a mean time of one year (resp. fifty years), we have $\bar{\mathbf{T}} = 365$ (resp. $\bar{\mathbf{T}} = 365 \times 50$), hence $\rho \approx 0.9972$ (resp. $\rho \approx 0.9999$).

Expected discounted payoff maximization problem

As the payoff (8) is contingent on the unknown scenario $\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2, \dots)$, it is practically impossible that a strategy \mathcal{S} performs better than another for all scenarios. We look for an *optimal strategy* \mathcal{S}^* , solution of

$$\mathbb{E}^{\mathbb{P}^{\pi_0}} [j(\mathcal{S}^*, \mathbf{W})] = \max_{\mathcal{S}} \mathbb{E}^{\mathbb{P}^{\pi_0}} [j(\mathcal{S}, \mathbf{W})] , \quad (9)$$

where $j(\mathcal{S}, \mathbf{W})$ is given by (8), and the probability \mathbb{P}^{π_0} , on the sample space \mathbb{H}_{∞} in (1), is defined by (6).

2.2. Structure of an optimal strategy and behavior of an optimal decision-maker

Let $\Delta(\Sigma^1)$ denote the set of probability distributions on the simplex Σ^1 in (4). For any $\pi \in \Delta(\Sigma^1)$, we define

$$\llbracket \pi \rrbracket = \int_{\Sigma^1} (p^B, p^G) \pi(d(p^B, p^G)) = (\llbracket \pi \rrbracket^B, \llbracket \pi \rrbracket^G) \in \Delta(\Sigma^1), \quad (10a)$$

$$\llbracket \pi \rrbracket^B = \int_{\Sigma^1} p^B \pi(d(p^B, p^G)) \in [0, 1], \quad (10b)$$

$$\llbracket \pi \rrbracket^G = \int_{\Sigma^1} p^G \pi(d(p^B, p^G)) \in [0, 1], \quad (10c)$$

that is, the mean of the random variable (p^B, p^G) under probability π , and the means of its two components (with $\llbracket \pi \rrbracket^B + \llbracket \pi \rrbracket^G = 1$). We also define the two *shift mappings* $\theta^B, \theta^G : \Delta(\Sigma^1) \rightarrow \Delta(\Sigma^1)$ by

$$(\theta^B \pi)(d(p^B, p^G)) = \frac{p^B}{\llbracket \pi \rrbracket^B} \pi(d(p^B, p^G)), \quad (11a)$$

$$(\theta^G \pi)(d(p^B, p^G)) = \frac{p^G}{\llbracket \pi \rrbracket^G} \pi(d(p^B, p^G)). \quad (11b)$$

Thus, $\theta^B \pi$ and $\theta^G \pi$, are absolutely continuous with respect to π . When $\llbracket \pi \rrbracket^B = 0$, that is, when $\pi = \delta_{(0,1)}$, we set $\theta^B \delta_{(0,1)} = \delta_{(0,1)}$ and, when $\llbracket \pi \rrbracket^G = 0$, that is, when $\pi = \delta_{(1,0)}$, we set $\theta^G \delta_{(1,0)} = \delta_{(1,0)}$.

The proof of the following Proposition 1 is to be found in Appendix A.1.

Proposition 1. *There exists an optimal strategy $\mathcal{S}^* = (\mathcal{S}_0^*, \mathcal{S}_1^*, \dots)$ solution of the optimization problem (9) made of stationary feedback policies of the form*

$$\mathcal{S}_t^*(\mathbf{Y}_1, \dots, \mathbf{Y}_t) = \widehat{\mathcal{S}}(\pi_t), \quad \forall t = 0, 1, 2, \dots, \quad (12)$$

where $\pi_t \in \Delta(\Sigma^1)$ is given by the dynamical equation

$$\pi_0 = \bar{\pi}_0 \quad \text{and} \quad \pi_{t+1} = f(\pi_t, \mathbf{Y}_{t+1}) = \begin{cases} \pi_t & \text{if } \mathbf{Y}_{t+1} = \partial, \\ \theta^B \pi_t & \text{if } \mathbf{Y}_{t+1} = B, \\ \theta^G \pi_t & \text{if } \mathbf{Y}_{t+1} = G. \end{cases} \quad (13)$$

Regarding the stationary feedback $\widehat{\mathcal{S}} : \Delta(\Sigma^1) \rightarrow \{\varepsilon, \alpha\}$, there exists a subset $\Pi_\alpha \subset \Delta(\Sigma^1)$, and its complementary subset $\Pi_\varepsilon = \Delta(\Sigma^1) \setminus \Pi_\alpha$, such that

- $\widehat{\mathcal{S}}(\pi) = \alpha$ (that is, select decision “avoid”) if $\pi \in \Pi_\alpha$,
- $\widehat{\mathcal{S}}(\pi) = \varepsilon$ (that is, select decision “experiment”) if $\pi \in \Pi_\varepsilon$.

Regarding the complementary subsets Π_α and Π_ε , there exists a function $V : \Delta(\Sigma^1) \rightarrow \mathbb{R}$ such that

$$\pi \in \Pi_\alpha \iff V(\pi) = \frac{\mathcal{U}_\alpha}{1 - \rho}, \quad \pi \in \Pi_\varepsilon \iff V(\pi) > \frac{\mathcal{U}_\alpha}{1 - \rho}. \quad (14)$$

The so-called *information state* $\pi_t \in \Delta(\Sigma^1)$ is the conditional distribution, with respect to $\mathbf{Y}_1, \dots, \mathbf{Y}_t$, of the first coordinate mapping on $\Sigma^1 \times \mathbb{H}_\infty$, that is, the *posterior* of (p^B, p^G) at stage t .

We call *optimal DM* a decision-maker who adopts the optimal strategy of Proposition 1. To describe the behavior of an optimal DM, we introduce the *first avoidance stage*, or *first prudent stage*, as the random variable defined by

$$\tau = \inf \{t = 0, 1, 2 \dots \mid \pi_t \in \Pi_\alpha\}. \quad (15)$$

In case $\pi_t \notin \Pi_\alpha$ for all stages $t = 0, 1, 2 \dots$, the convention is $\tau = \inf \emptyset = +\infty$.

The proof of the following Proposition 2 is to be found in Appendix A.2 (as we said at the beginning of this Section, the following result is not new, but we give a proof that does not rely on the Gittins Index Theorem).

Proposition 2. *The DM that follows the optimal strategy of Proposition 1 switches at most once from experimenting to avoiding. More precisely, her/his behavior displays one of the three following patterns, depending on the first avoidance stage τ in (15).*

- Infinite risky behavior:*
if $\tau = +\infty$, that is, if $\pi_t \in \Pi_\varepsilon$ for all stages $t = 0, 1, 2 \dots$, the optimal DM always experiments (taking risks), hence never avoids.
- No risky behavior:*
if $\tau = 0$, that is, if $\pi_0 \notin \Pi_\alpha$ (that is, $\pi_0 \in \Pi_\varepsilon$), the optimal DM avoids from the start and, from then on, the DM keeps avoiding (prudence) for all times.
- Finite risky behavior:*
if $1 \leq \tau < +\infty$, the optimal DM

- *experiments (taking risks) from $t = 0$ to $\tau - 1$, that is, as long as $\pi_t \in \Pi_\varepsilon$,*
- *switches to avoiding at stage $t = \tau$, that is, as soon as $\pi_t \in \Pi_\alpha$,*
- *from then on, keeps avoiding (prudence) for all times.*

Both optimal strategy and behavior are simpler in the full feedback case, where foregone payoffs are revealed no matter what the decision made. Indeed, in this case, the posterior π_t is updated as in (A.11), and then one chooses the risky option at stage t if and only if $\llbracket \pi_t \rrbracket^B \mathcal{U}^B + \llbracket \pi_t \rrbracket^G \mathcal{U}^G > \mathcal{U}_\alpha$. As a consequence, when under full feedback, one optimally selects the risky option, it is also optimal to do so under partial feedback because, by (A.8b), $\llbracket \pi_t \rrbracket^B \mathcal{U}^B + \llbracket \pi_t \rrbracket^G \mathcal{U}^G > \mathcal{U}_\alpha \implies V(\pi_t) \geq \llbracket \pi_t \rrbracket^B \mathcal{U}^B + \llbracket \pi_t \rrbracket^G \mathcal{U}^G > \mathcal{U}_\alpha$.

3. Conditions for biased or accurate assessments

Now, we show features of the optimal DM behavior that possess interesting psychological interpretations in terms of human biases: “euphorism” and status quo biases in §3.1; possible erroneous assessments of the objective best option and of the objective probabilities in §3.2.

3.1. “Euphorism” and status quo biases

Our analysis provides theoretical support to a mix of the so-called *status quo bias* — a preference for the current state of affairs documented in Samuelson and Zeckhauser (1988) — and to an inclination that we coin “euphorism” bias, related to the “stay-with-a-winner” rule — if an individual experiments a good outcome, it is rational to go on taking risks.

Proposition 3. *If the optimal DM experiments a good outcome, the DM will go on experimenting (“euphorism”). As a consequence, the experimenting phase (in case it exists) of the optimal DM can only stop when a bad outcome materializes: the switch from riskiness to prudence can only be triggered by the occurrence of a bad outcome.*

Therefore, the behavior of the optimal DM displays at most two consecutive phases of “status quo” — one (possibly empty) of experimenting, that is, taking risks, one (possibly empty) of prudence — with at most one switch; in particular, once prudent, this is forever.

Proof. First, we prove that, if the optimal DM experiments a good outcome, the DM will go on experimenting.

Suppose that, at stage t the optimal DM is experimenting. We will show that, if a good outcome \mathbf{G} materializes at the end of the interval $[t, t+1[$ (that is, if $\mathbf{Y}_{t+1} = \mathbf{G}$), then necessarily the optimal DM goes on experimenting at stage $t+1$. In what follows, the value function $V : \Delta(\Sigma^1) \rightarrow \mathbb{R}$ has been introduced in Proposition 1, and is defined in (A.5). We have that

$$V(\pi_{t+1}) = V(\theta^{\mathbf{G}}\pi_t)$$

because, as we supposed that $\mathbf{Y}_{t+1} = \mathbf{G}$, we have that $\pi_{t+1} = \theta^{\mathbf{G}}\pi_t$ by the dynamics (13)

$$\begin{aligned} &\geq V(\pi_t) && \text{(by the property } V \circ \theta^{\mathbf{G}} \geq V, \text{ shown afterward)} \\ &> \frac{\mathcal{U}_\alpha}{1-\rho} \end{aligned}$$

by (14), because, as we supposed that the optimal DM is experimenting at stage t , we have that $\pi_t \in \Pi_\varepsilon$. Thus, we have obtained that $V(\pi_{t+1}) > \frac{\mathcal{U}_\alpha}{1-\rho}$. By (14), we conclude that the optimal DM goes on experimenting at stage $t+1$ by Proposition 1.

We now prove that the value function (A.5) has the property

$$V \circ \theta^{\mathbf{G}} \geq V, \tag{16}$$

that is, the value function cannot decrease when the posterior changes following a good outcome. We will use this “stay-with-a-winner” property when we discuss the robustness of our findings in §4.2.

Before that, we recall that two random variables \mathbf{C} and \mathbf{D} , defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, are said to be *comonotonic* when we have that $(\mathbf{C}(\omega) - \mathbf{C}(\omega'))(\mathbf{D}(\omega) - \mathbf{D}(\omega')) \geq 0$, for any $(\omega, \omega') \in \Omega^2$. In that case, it is easily shown that $\mathbb{E}[\mathbf{CD}] \geq \mathbb{E}[\mathbf{C}]\mathbb{E}[\mathbf{D}]$, when \mathbf{C} and \mathbf{D} are square integrable.

By definition (A.5) of the value function V , to prove (16) it suffices to show that

$$\int_{\Sigma^1} L(x, v)(\theta^{\mathbf{G}}\pi)(dx) \geq \int_{\Sigma^1} L(x, v)\pi(dx), \quad \forall v \in \{\alpha, \varepsilon\}.$$

This is obvious for $v = \alpha$ since $L(x, \alpha) = \mathcal{U}_\alpha$ by (A.1e). For $v = \varepsilon$, we have that

$$\begin{aligned} \int_{\Sigma^1} L(x, \varepsilon)(\theta^G \pi)(dx) &= \int_{\Sigma^1} [p^B \mathcal{U}^B + p^G \mathcal{U}^G] \frac{p^G}{\llbracket \pi \rrbracket^G} \pi(d(p^B, p^G)) \\ &\quad \text{(by (A.1e) and (11))} \\ &\geq \frac{1}{\llbracket \pi \rrbracket^G} \int_{\Sigma^1} [p^B \mathcal{U}^B + p^G \mathcal{U}^G] \pi(d(p^B, p^G)) \int_{\Sigma^1} p^G \pi(d(p^B, p^G)) \end{aligned}$$

because the random variables $\mathbf{C} : \Sigma^1 \ni (p^B, p^G) \mapsto p^B \mathcal{U}^B + p^G \mathcal{U}^G$ and $\mathbf{D} : \Sigma^1 \ni (p^B, p^G) \mapsto p^G$ are comonotonic, since the function $p^G \mapsto p^B \mathcal{U}^B + p^G \mathcal{U}^G = p^G(\mathcal{U}^G - \mathcal{U}^B) + \mathcal{U}^B$ is increasing as a consequence of $\mathcal{U}^G > \mathcal{U}^B$ by (7)

$$\begin{aligned} &= \int_{\Sigma^1} [p^B \mathcal{U}^B + p^G \mathcal{U}^G] \pi(d(p^B, p^G)) \\ &\quad \text{(by the definition (10) of } \llbracket \pi \rrbracket^G \text{)} \\ &= \int_{\Sigma^1} L(x, \varepsilon) \pi(dx) \quad \text{(by (A.1e).)} \end{aligned}$$

We can prove in the same way that $V \circ \theta^B \leq V$, so that we have obtained

$$V \circ \theta^G \geq V \geq V \circ \theta^B. \quad (17)$$

The rest of the proof follows from Proposition 1. In particular, once the optimal DM selects the “avoid” option, the DM will never more experiment. Indeed, the optimal rule of Proposition 1 states that, once the optimal DM selects the “avoid” option, the DM does not observe the random outcomes, hence the DM no longer updates the posterior π_t because of the dynamics (13) so that the DM keeps avoiding.

This ends the proof. \square

3.2. Possible erroneous assessments of the objective best option and of the objective probabilities

3.2.1. Objective environment

We suppose that Nature produces bad and good outcomes that are sequences of independent Bernoulli trials governed by a given $(\bar{p}^B, \bar{p}^G) \in \Sigma^1$. Thus, we equip the sample space \mathbb{H}_∞ in (1) with the probability distribution $\mathbb{P}^{\delta_{(\bar{p}^B, \bar{p}^G)}} = \mathcal{B}(\bar{p}^B, \bar{p}^G)$ as in (5).

Definition 4. We call the couple $\bar{p} = (\bar{p}^B, \bar{p}^G) \in \Sigma^1$ the objective or true probabilities. We call environment the triplet (ρ, U, p) consisting of the discount factor $\rho \in [0, 1[$, the payoff function U in Table 1 (that is, avoidance payoff \mathcal{U}_α , low payoff \mathcal{U}^B and high payoff \mathcal{U}^G), and the objective probabilities $\bar{p} = (\bar{p}^B, \bar{p}^G)$.

We define the critical probability p_c by the ratio

$$p_c = \frac{\mathcal{U}^G - \mathcal{U}_\alpha}{\mathcal{U}^G - \mathcal{U}^B} = \frac{\text{relative payoff of avoidance}}{\text{relative payoff of bad outcome}} \in]0, 1[, \quad (18a)$$

so that we have the equivalence

$$\bar{p}^B < p_c \iff \bar{p}^B \mathcal{U}^B + \bar{p}^G \mathcal{U}^G > \mathcal{U}_\alpha . \quad (18b)$$

When $\bar{p}^B < p_c$ (resp. \geq) or, equivalently, when $\bar{p}^B \mathcal{U}^B + \bar{p}^G \mathcal{U}^G > \mathcal{U}_\alpha$ (resp. \leq), we say that the risky (resp. certain) option is the objectively best option and that the environment is prone to risk-taking (resp. prudence).

All things being equal, the worse a bad outcome (that is, low payoff of bad outcome), the lower the critical probability (18a). When $p_c \approx 0$, the bad outcome is so bad that the positive difference between the payoff of the good outcome and the avoidance payoff is negligible w.r.t. the positive difference between the payoff of the good and the bad outcomes; hence, prudence (avoidance) is the best objective option for most of the values \bar{p}^B , since $\bar{p}^B \geq p_c \approx 0$. When $p_c \approx 1$, the good outcome is so good that avoiding the bad outcome costs almost as well as suffering it; taking risks is the best objective option for most of the values \bar{p}^B , since $\bar{p}^B < p_c \approx 1$.

3.2.2. Tentative formal definition of bias

We consider the situation where the optimal DM adopts the strategy of Proposition 1, optimal for a given prior beta distribution $\bar{\pi}_0$. More precisely, let $n_0^B > 0$ and $n_0^G > 0$ be two positive scalars. We suppose that the distribution $\bar{\pi}_0$ is the beta distribution $\beta(n_0^B, n_0^G)$ on the simplex Σ^1 in (4), that is, for any integrable function $\varphi : \Sigma^1 \rightarrow \mathbb{R}$, we have that

$$\int_{\Sigma^1} \varphi(p^B, p^G) \bar{\pi}_0 d(p^B, p^G) = \frac{\int_0^1 \varphi(p, 1-p) p^{n_0^B-1} (1-p)^{n_0^G-1} dp}{\int_0^1 p^{n_0^B-1} (1-p)^{n_0^G-1} dp} . \quad (19)$$

Now, we are equipped to formally define what we call a bias. On the one hand, the optimal strategy of Proposition 1, depends on the discount

factor $\rho \in [0, 1[$, on the payoff function U in Table 1, and on the prior beta distribution $\bar{\pi}_0 = \beta(n_0^B, n_0^G)$, but not on the objective probabilities (\bar{p}^B, \bar{p}^G) . In other words, optimality is w.r.t. the criterion (9), where the mathematical expectation is taken w.r.t. the (subjective) probability $\mathbb{P}^{\bar{\pi}_0}$, on the sample space \mathbb{H}_∞ in (1), as defined by (6). On the other hand, Nature produces bad and good outcomes that are sequences of independent Bernoulli trials governed by the objective probabilities $(\bar{p}^B, \bar{p}^G) \in \Sigma^1$. Would the DM know (\bar{p}^B, \bar{p}^G) , she/he would design a strategy maximizing the criterion (9), but where the mathematical expectation would be taken w.r.t. the (objective) probability $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} = \mathcal{B}(\bar{p}^B, \bar{p}^G)$ as in (5). We say that the optimal strategy of Proposition 1 displays a bias when one of its outputs is discrepant with what it would be if the objective probability distribution were known.

Thus, the probability distribution $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} = \mathcal{B}(\bar{p}^B, \bar{p}^G)$ and the stochastic process $(\mathbf{W}_1, \mathbf{W}_2, \dots)$ governed by $\mathcal{B}(\bar{p}^B, \bar{p}^G)$ play the role of a background reference objective environment against which one can assess the outputs of the optimal strategy (optimal for $\bar{\pi}_0$), and possibly qualify them of biased or not. We distinguish two outputs of the optimal strategy. One such output is $\llbracket \pi_t \rrbracket^B$, the mathematical expectation (10) of the random variable p^B . By optimally updating the posterior distribution π_t , the optimal DM also updates her/his estimate $\llbracket \pi_t \rrbracket^B$ of the unknown probability \bar{p}^B of the bad outcome B. Another output is whether $\llbracket \pi_t \rrbracket^B \mathcal{U}^B + \llbracket \pi_t \rrbracket^G \mathcal{U}^G > \mathcal{U}_\alpha$ or $\llbracket \pi_t \rrbracket^B \mathcal{U}^B + \llbracket \pi_t \rrbracket^G \mathcal{U}^G \leq \mathcal{U}_\alpha$, that is, how the DM assesses whether the environment is prone to risk-taking or to prudence.

3.2.3. Main result

The following Proposition 5 details what are the estimates $\llbracket \pi_t \rrbracket^B$ of the objective probability value \bar{p}^B that the optimal DM is forming during the course of learning, and how she/he assesses the environment. It also establishes how the probability of relevant events monotonically depend upon objective probabilities. To our knowledge, these results are new.

Proposition 5. *Let $(\mathbf{W}_1, \mathbf{W}_2, \dots)$ be a sequence of independent Bernoulli trials governed by the objective probability distribution $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} = \mathcal{B}(\bar{p}^B, \bar{p}^G)$, as in (5), where $(\bar{p}^B, \bar{p}^G) \in \Sigma^1$. Suppose that the DM adopts the corresponding strategy \mathcal{S}^* of Proposition 1, based on the observations $(\mathbf{Y}_1, \mathbf{Y}_2, \dots)$ inductively given by (3) (for $\mathcal{S} = \mathcal{S}^*$) and on the sequence of posteriors π_t given by the dynamics (13).*

Suppose also that the DM holds the prior beta distribution $\bar{\pi}_0 = \beta(n_0^B, n_0^G)$, where $n_0^B > 0$ and $n_0^G > 0$ are two positive scalars. Then, if we define the

numbers N_t^B and N_t^G of bad and good outcomes up to stage t by

$$N_0^B = N_0^G = 0, \quad N_t^B = \sum_{s=1}^t \mathbf{1}_{\{\mathbf{Y}_s = B\}}, \quad N_t^G = \sum_{s=1}^t \mathbf{1}_{\{\mathbf{Y}_s = G\}}, \quad t = 1, 2, \dots \quad (20)$$

the posterior π_t in Proposition 1 is the beta distribution

$$\pi_t = \beta(n_0^B + N_t^B, n_0^G + N_t^G), \quad (21)$$

on the simplex Σ^1 , whose conditional expectation (10) is given by the statistics

$$\llbracket \pi_0 \rrbracket^B = \frac{n_0^B}{n_0^B + n_0^G} \quad \text{and} \quad \llbracket \pi_t \rrbracket^B = \frac{n_0^B + N_t^B}{n_0^B + n_0^G + t}, \quad t = 1, 2, \dots \quad (22)$$

Moreover, here are the assessments of the objective probability value \bar{p}^B and of the environment made by the above optimal DM.

- a) Infinite learning $\tau = +\infty$. Infinite learning can only happen when $\bar{\pi}_0 \in \Pi_\varepsilon$. When $\tau = +\infty$, the optimal DM experiments forever and, the statistics $\llbracket \pi_t \rrbracket^B$ in (22) asymptotically reaches the objective probability value \bar{p}^B , almost surely under the objective probability distribution $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)}$, that is,

$$\lim_{t \rightarrow +\infty} \llbracket \pi_t \rrbracket^B = \bar{p}^B, \quad (23a)$$

or, in more precise terms,

$$\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} \left\{ \lim_{t \rightarrow +\infty} \llbracket \pi_t \rrbracket^B = \bar{p}^B, \quad \tau = +\infty \right\} = \mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} \{ \tau = +\infty \}. \quad (23b)$$

Then, the optimal DM asymptotically makes an accurate assessment of the objective best option as $\lim_{t \rightarrow +\infty} (\llbracket \pi_t \rrbracket^B \mathcal{U}^B + \llbracket \pi_t \rrbracket^G \mathcal{U}^G) = \bar{p}^B \mathcal{U}^B + \bar{p}^G \mathcal{U}^G$.

Infinite learning happens with probability $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)}(\pi_t \in \Pi_\varepsilon, \quad \forall t = 0, 1, 2, \dots)$, which goes up to 1 when the objective probability \bar{p}^B of the bad outcome B goes down to 0. As a consequence, an accurate estimation of the objective probability of a rare bad outcome is likely.

- b) No learning $\tau = 0$. No learning happens if and only if $\bar{\pi}_0 \in \Pi_\alpha$. When $\tau = 0$, the optimal DM never experiments and the DM initial estimate $\llbracket \pi_0 \rrbracket^B$ of the objective probability value \bar{p}^B satisfies

$$p_c \leq \llbracket \pi_0 \rrbracket^B, \quad (24)$$

where the critical probability p_c is defined in (18a). From the start, the optimal DM assesses that the environment is prone to prudence, as $\llbracket \pi_0 \rrbracket^B \mathcal{U}^B + \llbracket \pi_0 \rrbracket^G \mathcal{U}^G \leq \mathcal{U}_\alpha$.

- c) Finite learning $1 \leq \tau < +\infty$. *Finite learning can only happen when $\bar{\pi}_0 \in \Pi_\varepsilon$. When $1 \leq \tau < +\infty$, the optimal DM experiments till stage τ and the DM stops her/his estimation of the objective probability value \bar{p}^B at a value $\llbracket \pi_\tau \rrbracket^B$ which satisfies*

$$p_c \leq \llbracket \pi_\tau \rrbracket^B, \quad (25a)$$

or, in more precise terms,

$$\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} \left\{ p_c \leq \llbracket \pi_\tau \rrbracket^B, \ 1 \leq \tau < +\infty \right\} = \mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} \left\{ 1 \leq \tau < +\infty \right\}. \quad (25b)$$

When the optimal DM stops experimenting, she/he assesses that the environment is prone to prudence, as $\llbracket \pi_\tau \rrbracket^B \mathcal{U}^B + \llbracket \pi_\tau \rrbracket^G \mathcal{U}^G \leq \mathcal{U}_\alpha$.

Finite learning happens with probability $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} (\exists t = 1, 2, \dots, \pi_t \in \Pi_\alpha)$, which goes down to 0 when the objective probability \bar{p}^B of the bad outcome B goes down to 0. As a consequence, if the objective probability \bar{p}^B of the bad outcome B is low enough, in the sense that $\bar{p}^B \leq p_c$, when the experiment phase ends at $\tau < +\infty$, we have

$$\bar{p}^B \leq p_c \leq \llbracket \pi_\tau \rrbracket^B, \quad (26)$$

hence the DM will overestimate the objective probability \bar{p}^B of the bad outcome B, but this with a vanishing probability as $\bar{p}^B \downarrow 0$.

Proof. By the dynamics (13), we easily establish that (21) holds true. Equation (22) follows from property of beta distributions (19).

- a) By Proposition 1, when $\tau = +\infty$ it is optimal to select decision ε and experiment forever. Thus, the observations $(\mathbf{Y}_1, \mathbf{Y}_2, \dots)$ in (3) coincide with $(\mathbf{W}_1, \mathbf{W}_2, \dots)$, and we get that $N_t^B = \sum_{s=1}^t \mathbf{1}_{\{\mathbf{W}_s=B\}}$ and $N_t^G = \sum_{s=1}^t \mathbf{1}_{\{\mathbf{W}_s=G\}}$, for all $t = 1, 2, \dots$, by (20). As the random variables $(\mathbf{W}_1, \mathbf{W}_2, \dots)$ in (2) are i.i.d. under $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)}$, by the Law of large numbers we have that

$$\frac{n_0^B + N_t^B}{n_0^B + N_t^B + n_0^G + N_t^G} = \frac{n_0^B + \sum_{s=1}^t \mathbf{1}_{\{\mathbf{W}_s=B\}}}{n_0^B + n_0^G + t} \xrightarrow{t \rightarrow +\infty} \bar{p}^B, \quad \mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} - \text{p.s.}$$

By (22), asymptotically the statistics $\llbracket \pi_t \rrbracket^B$ reaches the objective probability value \bar{p}^B almost surely under the probability $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)}$.

Now, we show that the function $[0, 1] \ni \bar{p}^B \mapsto \mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)}(\pi_t \in \Pi_\varepsilon, \forall t = 0, 1, 2, \dots)$ is nonincreasing. For this purpose, it suffices to prove that, for any stage t , the function

$$[0, 1] \ni \bar{p}^B \mapsto \mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)}(\pi_s \in \Pi_\varepsilon, \forall s = 0, 1, 2, \dots, t) \quad (27)$$

is nonincreasing, and then let $t \rightarrow +\infty$. Now, the functions $\varphi_s(\pi) = \mathbf{1}_{\{\pi \in \Pi_\varepsilon\}}$ (the same function for all $s = 0, 1, 2, \dots$) satisfy the assumptions of Proposition 7 because

$$\begin{aligned} (\varphi_s \circ \theta^G)(\pi) &= \mathbf{1}_{\{\theta^G \pi \in \Pi_\varepsilon\}} \\ &= \mathbf{1}_{\{(V \circ \theta^G)(\pi) > \frac{\mathcal{U}_\alpha}{1-\rho}\}} && \text{(by (14))} \\ &\geq \mathbf{1}_{\{(V \circ \theta^B)(\pi) > \frac{\mathcal{U}_\alpha}{1-\rho}\}} && \text{(as } V \circ \theta^G \geq V \circ \theta^B \text{ by (17))} \\ &= \mathbf{1}_{\{\theta^B \pi \in \Pi_\varepsilon\}} && \text{(by (14))} \\ &= (\varphi_s \circ \theta^B)(\pi). \end{aligned}$$

We conclude, using Proposition 7, that the function (27) is nonincreasing. Finally, we easily establish that $\mathbb{P}^{\delta(0,1)}(\pi_t \in \Pi_\varepsilon, \forall t = 0, 1, 2, \dots) = 1$. Indeed, under the probability $\mathbb{P}^{\delta(0,1)}$, we have $\mathbf{W}_t = \mathbf{G}$ for all stage t almost-surely, hence $\pi_{t+1} = \theta^G \pi_t$ by the dynamics (13). Therefore, we get that

$$\begin{aligned} \pi_t \in \Pi_\varepsilon &\implies V(\pi_t) > \frac{\mathcal{U}_\alpha}{1-\rho} && \text{(by (14))} \\ &\implies V(\theta^G \pi_t) \geq V(\pi_t) > \frac{\mathcal{U}_\alpha}{1-\rho} && \text{(by (17))} \\ &\implies V(\pi_{t+1}) > \frac{\mathcal{U}_\alpha}{1-\rho} && \text{(since } \pi_{t+1} = \theta^G \pi_t \text{)} \\ &\implies \pi_{t+1} \in \Pi_\varepsilon. && \text{(by (14))} \end{aligned}$$

Since $\bar{\pi}_0 \in \Pi_\varepsilon$, we deduce that $\mathbb{P}^{\delta(0,1)}(\pi_t \in \Pi_\varepsilon, \forall t = 0, 1, 2, \dots) = 1$.

b) See the proof below.

c) Let us suppose that $\tau < +\infty$. We have that

$$\begin{aligned} \mathcal{U}_\alpha &= (1-\rho)V(\pi_\tau) && \text{(by definition (15) of } \tau \text{ and since } \tau < +\infty) \\ &\geq \llbracket \pi_\tau \rrbracket^B \mathcal{U}^B + \llbracket \pi_\tau \rrbracket^G \mathcal{U}^G && \text{(by the inequality (A.8b))} \\ &= -\llbracket \pi_\tau \rrbracket^B (\mathcal{U}^G - \mathcal{U}^B) + \mathcal{U}^G && \text{(since } \llbracket \pi_\tau \rrbracket^B + \llbracket \pi_\tau \rrbracket^G = 1 \text{ by (10).)} \end{aligned}$$

Rearranging the terms, and using (18a), we obtain that $\llbracket \pi_\tau \rrbracket^B \geq p_c = \frac{\mathcal{U}^G - \mathcal{U}_\alpha}{\mathcal{U}^G - \mathcal{U}^B}$.

The rest of the proof follows using the property that, if $\pi_0 \in \Pi_\varepsilon$, then

$$\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)}(\exists t = 1, 2, \dots, \pi_t \in \Pi_\alpha) = 1 - \mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)}(\pi_t \in \Pi_\varepsilon, \forall t = 0, 1, 2, \dots).$$

This ends the proof. \square

3.2.4. The case of environments prone to prudence

In Table 2, we sum up the results of Proposition 5 in the case of a prudence-prone environment. Apart from the “euphorism” and status quo biases, already discussed in §3.1, the optimal DM displays no bias in the sense that she/he makes the accurate assessment that the environment is prone to prudence, and that nothing can be said of how the estimate $\llbracket \pi_\tau \rrbracket^B$ is related to the objective probability \bar{p}^B (hence the empty box in Table 2).

3.2.5. The case of environments prone to risk-taking

In Table 3, we sum up the results of Proposition 5 in the case of an environment prone to risk-taking, and we point out two possible biases. On top of the “euphorism” and status quo biases, already discussed in §3.1, the optimal DM displays an additional form of pessimism bias. Indeed, the first (and last) time the optimal DM stops choosing the risky option, she/he will erroneously assess that the environment is prone to prudence, and will overestimate the probability of the bad outcome. More precisely, Proposition 5 yields the following *Biased Learning Theorem*.

Theorem 6 (Biased Learning Theorem). *Suppose that the assumptions of Proposition 5 are satisfied and that $\bar{\pi}_0 \in \Pi_\varepsilon$, so that learning happens, either infinite or finite. Suppose that the environment is prone to risk-taking, that is,*

$$\bar{p}^B < p_c. \quad (28)$$

Then, the optimal DM (of Proposition 1) can only display two behaviors.

1. *Either the optimal DM will experiment forever, and will accurately estimate asymptotically the objective probability \bar{p}^B of the bad outcome B; this experiment phase happens with a probability $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)}(\pi_t \in \Pi_\varepsilon, \forall t = 0, 1, 2, \dots)$ which goes up to 1 when the objective probability \bar{p}^B of the bad outcome B goes down to 0. In that case, we conclude that the more likely a bad outcome, the more likely the optimal DM makes an accurate estimation of its objective probability.*

Environment prone to prudence	Endless risk-taking $\tau = +\infty$	Risk-taking then endless prudence $1 \leq \tau < +\infty$
$\bar{p}^B \geq p_c$ \iff $\bar{p}^B \mathcal{U}^B + \bar{p}^G \mathcal{U}^G \leq \mathcal{U}_\alpha$	Behavior <i>discrepant</i> with the feature of the environment	Behavior <i>consistent</i> with the feature of the environment
	The more likely the bad outcome, the <i>lower the probability</i> of <i>discrepant</i> endless risk-taking: $\bar{p}^B \nearrow \implies$ $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} \{ \tau = +\infty \} \searrow$	The more likely the bad outcome, the <i>higher the probability</i> of <i>consistent</i> endless prudence: $\bar{p}^B \nearrow \implies$ $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} \{ \tau < +\infty \} \nearrow$
	<i>Accurate</i> asymptotic estimation of \bar{p}^B : $\lim_{t \rightarrow +\infty} \llbracket \pi_t \rrbracket^B = \bar{p}^B$	
	Asymptotic <i>accurate</i> assessment that the environment is prone to prudence: $\lim_{t \rightarrow +\infty} (\llbracket \pi_t \rrbracket^B \mathcal{U}^B + \llbracket \pi_t \rrbracket^G \mathcal{U}^G) \leq \mathcal{U}_\alpha$	<i>Accurate</i> assessment that the environment is prone to prudence: $\llbracket \pi_\tau \rrbracket^B \mathcal{U}^B + \llbracket \pi_\tau \rrbracket^G \mathcal{U}^G \leq \mathcal{U}_\alpha$

Table 2: Optimal behavior in an environment objectively prone to prudence

Environment prone to risk	Endless risk-taking $\tau = +\infty$	Risk-taking then endless prudence $1 \leq \tau < +\infty$
$\bar{p}^B < p_c$ \iff $\bar{p}^B \mathcal{U}^B + \bar{p}^G \mathcal{U}^G > \mathcal{U}_\alpha$	Behavior <i>consistent</i> with the feature of the environment	Behavior <i>discrepant</i> with the feature of the environment
	The more unlikely the bad outcome the <i>higher the probability</i> of <i>consistent</i> endless risk-taking: $\bar{p}^B \searrow \implies$ $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} \{ \tau = +\infty \} \nearrow$	The more unlikely the bad outcome the <i>lower the probability</i> of <i>discrepant</i> endless prudence: $\bar{p}^B \searrow \implies$ $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} \{ \tau < +\infty \} \searrow$
	$\bar{p}^B \downarrow 0 \implies$ $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} \{ \tau = +\infty \} \uparrow 1$	<i>Vanishing discrepancy:</i> $\bar{p}^B \downarrow 0 \implies$ $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)} \{ \tau < +\infty \} \downarrow 0$
	Asymptotic <i>accurate</i> estimation of the probability \bar{p}^B of the bad outcome: $\lim_{t \rightarrow +\infty} \llbracket \pi_t \rrbracket^B = \bar{p}^B$	<i>Overestimation</i> of the probability \bar{p}^B of the bad outcome: $\bar{p}^B < p_c \leq \llbracket \pi_\tau \rrbracket^B$
	Asymptotic <i>accurate</i> assessment that the environment is prone to risk-taking: $\lim_{t \rightarrow +\infty} (\llbracket \pi_t \rrbracket^B \mathcal{U}^B + \llbracket \pi_t \rrbracket^G \mathcal{U}^G) > \mathcal{U}_\alpha$	<i>Erroneous</i> assessment that the environment is prone to prudence: $\llbracket \pi_\tau \rrbracket^B \mathcal{U}^B + \llbracket \pi_\tau \rrbracket^G \mathcal{U}^G \leq \mathcal{U}_\alpha$

Table 3: Optimal behavior in an environment objectively prone to risk-taking

2. Or the DM will experiment during a finite number of stages and then stop experimenting forever; this stopping phase happens with a probability $\mathbb{P}^{\delta(\bar{p}^B, \bar{p}^G)}(\exists t = 1, 2, \dots, \pi_t \in \Pi_\alpha)$ which goes down to 0 when the objective probability \bar{p}^B of the bad outcome B goes down to 0. In that case, we conclude that, if the objective probability \bar{p}^B of the bad outcome B is so low that $\bar{p}^B \leq p_c$, then, when the experiment phase ends at $\tau < +\infty$, the optimal DM will overestimate the objective probability \bar{p}^B of the bad outcome B because of the inequalities

$$\bar{p}^B \leq p_c \leq \llbracket \pi_\tau \rrbracket^B. \quad (29)$$

However, such an overestimation happens with a vanishing probability as $\bar{p}^B \downarrow 0$.

Economists have made the point, coined the *Incomplete Learning Theorem*, that the optimal strategy (to maximize discounted expected utility) does not necessarily lead to exactly evaluate the unknown probability (Rothschild (1974); Easley and Kiefer (1988); Brezzi and Lai (2000)). Thus, optimality does not necessarily lead to perfect accuracy. Our results point to a *Biased Learning Theorem*, as we prove that the departure from accuracy displays a bias towards overestimation of the bad outcome when learning stops. However, learning stops (hence overestimation happens) with nonincreasing and vanishing probability as the objective probability of the bad outcome goes down to zero.

4. Discussion

In §4.1, we discuss the possible implementation of an optimal strategy by humans, and, in §4.2, the robustness of our findings. Finally, we discuss boundaries and compare our results with the literature in §4.3.

4.1. Possible implementation of an optimal strategy by humans

We discuss the data necessary to design an optimal strategy, as in Proposition 1, and the cognitive burden of implementing it, to see if they are insuperable impediments to its progressive selection during the course of human evolution.

How does the DM obtain the basic data needed to implement an optimal strategy?

The optimal strategy of Proposition 1 depends on the discount factor ρ in (8) and on the payoff function U in Table 1 (that is, avoidance payoff \mathcal{U}_α , low payoff \mathcal{U}^B and high payoff \mathcal{U}^G). Also, under the assumptions of Proposition 5, the DM holds the prior beta distribution $\bar{\pi}_0 = \beta(n_0^B, n_0^G)$ in (19), where $n_0^B > 0$ and $n_0^G > 0$ are two positive scalars.

We have already discussed, right after Equation (8), how the discount factor ρ can be related to the mean number of stages during which the DM has to make decisions. We suppose that the DM knows the avoidance (middle) payoff \mathcal{U}_α . Then, we suggest a way for the DM to jointly determine two integers n_0^B and n_0^G for the beta distribution $\beta(n_0^B, n_0^G)$, and both the low payoff \mathcal{U}^B and the base payoff \mathcal{U}^G . The DM starts by making a risky decision and

- either the DM first enjoys n good outcomes G — hence discovering the high payoff \mathcal{U}^G — before suffering a bad outcome B — hence discovering the low payoff \mathcal{U}^B ; in that case, the DM sets $n_0^G = n$ and $n_0^B = 1$;
- or the DM first suffers n bad outcomes B — hence discovering the low payoff \mathcal{U}^B — before enjoying a good outcome G — hence discovering the high payoff \mathcal{U}^G ; in that case, the DM sets $n_0^G = 1$ and $n_0^B = n$.

So, at the end of those $n_0^G + n_0^B$ trials, the DM disposes of the two payoffs \mathcal{U}^B and \mathcal{U}^G , as well as the two integer parameters $n_0^B > 0$ and $n_0^G > 0$.

What is the stage by stage cognitive load of the optimal DM?

We suppose that the DM holds the prior beta distribution $\bar{\pi}_0 = \beta(n_0^B, n_0^G)$, where n_0^B, n_0^G are integers. Then, we know from Proposition 5 that the posterior π_t is the beta distribution $\beta(n_0^B + N_t^B, n_0^G + N_t^G)$ as in (21). Thus, at each decision stage t , the cognitive load of the optimal DM is to keep track of the two integers $n_0^B + N_t^B$ and $n_0^G + N_t^G$ since, by Proposition 1, the optimal decision at stage t is function of the posterior π_t .

How can the DM make an optimal decision at stage t ?

By Proposition 1, the DM has to determine if the current posterior $\pi_t \in \Delta(\Sigma^1)$ either belongs to the subset $\Pi_\alpha \subset \Delta(\Sigma^1)$ or to the complementary subset $\Pi_\varepsilon = \Delta(\Sigma^1) \setminus \Pi_\alpha$, to make an optimal decision at stage t . As $\pi_t = \beta(n_0^B + N_t^B, n_0^G + N_t^G)$, the DM needs to identify in which of two subsets of \mathbb{N}^2

— that is, the couples of integers corresponding to the subsets Π_α and Π_ε of $\Delta(\Sigma^1)$ — does the couple $(n_0^B + N_t^B, n_0^G + N_t^G)$ belong.

It is hard to say if our mind can design — using the discount factor ρ , the avoidance (middle) payoff \mathcal{U}_α , the low payoff \mathcal{U}^B and the high payoff \mathcal{U}^G — and if our brain can hold such a “mental planar map” made of couples of integers. For instance, for an individual making daily decisions during a mean time of one year (resp. fifty years), this planar map would consist of $365^2 = 133,225$ (resp. $(5 \times 365)^2 \approx 333 \cdot 10^6$) couples of integers labelled with a binary label. Even if they are not astronomical, these numbers are huge.³

However, it is possible that a close suboptimal strategy be much more simply encoded by the following rule: if

$$\llbracket \pi_t \rrbracket^B \mathcal{U}^B + \llbracket \pi_t \rrbracket^G \mathcal{U}^G = \frac{n_0^B + N_t^B}{n_0^B + n_0^G + t} \mathcal{U}^B + \frac{n_0^G + N_t^G}{n_0^B + n_0^G + t} \mathcal{U}^G > \mathcal{U}_\alpha$$

then make the risky decision $\mathbf{V}_t = \varepsilon$, else avoid. A DM adopting this strategy would be more prudent than the optimal DM because, by (A.8b), we have that

$$\frac{\llbracket \pi_t \rrbracket^B \mathcal{U}^B + \llbracket \pi_t \rrbracket^G \mathcal{U}^G}{1 - \rho} > \frac{\mathcal{U}_\alpha}{1 - \rho} \implies V(\pi_t) > \frac{\mathcal{U}_\alpha}{1 - \rho} \implies \pi_t \in \Pi_\varepsilon.$$

4.2. Robustness of the results obtained

We discuss which of our results are robust w.r.t. to assumptions like stationarity (of the primitive random variables, of the payoffs), discounting, finite or infinite horizon.

³We can easily arrive at astronomical figures with general policies. Indeed, recall that a policy at stage t is a mapping $\mathcal{S}_t : \{\mathbf{B}, \mathbf{G}, \partial\}^t \rightarrow \{\alpha, \varepsilon\}$ that tells the DM what will be the next action in view of past observations. Disregarding the irrelevant “observation” ∂ , a policy at stage t is a mapping from a set of cardinal 2^t towards a binary set. If an interval $[t, t + 1[$ represents one day, the storage of policies for one year would be astronomically prohibitive. This is why the existence of a stationary feedback optimal policy seems good news. However, the argument of such policy is now an element of $\Delta(\Sigma^1)$, that is, a probability distribution on the one-dimensional simplex. Equivalently, being able to implement the optimal strategy of Proposition 1 amounts to being able to characterize the two complementary subsets Π_α and Π_ε , which is out of question except with simple rules.

“Euphorism” and status quo biases

The property that, if the optimal DM experiments a good outcome, she/he will go on taking risks is a consequence of the “stay-with-a-winner” property (16): the value function cannot decrease when the posterior changes following a good outcome. Screening the proof of (16) shows that this property only depends the ranking (7) of payoffs, hence that our finding (“euphorism” bias) is robust w.r.t. nonstationarity (as long as it does not change the ranking), absence of discounting, and finite or infinite horizon.

The property that, if one selects the prudent decision once, one will no longer make risky decisions afterwards is a consequence of both the stopping of posterior updating and of the stationarity of the avoidance domain Π_α . We discuss both of them. The property that the posterior is a sufficient information state for optimization is quite robust, as it holds true under nonstationarity, absence of discounting, and finite or infinite horizon ((Bertsekas and Shreve, 1996, Chap. 10)). The stopping of posterior updating follows from the information structure, as avoidance freezes observation, hence is robust. However, it is stationarity and infinite horizon that lead to status quo. Indeed, were the avoidance domain Π_α dependent on the stage t that we could no longer conclude to status quo. Thus, the status quo bias is less robust than the “euphorism” bias.

Pessimistic erroneous assessment of the environment and overestimation bias for the probability of the bad outcome

The property that the probability of a bad outcome is overestimated when the risky phase stops (hence the erroneous assessment that the environment is prone to prudence) comes from the inequality (A.8b), itself a consequence of stationarity, discounting and infinite horizon. In this sense, it is less robust than the “euphorism” bias.

4.3. Boundaries and comparison with the literature

Regarding the status quo bias, it is the consequence of the existence of a prudent decision which freezes observation. So, the boundaries of such a phenomenon are rather narrow. For instance, the status quo bias theoretically vanishes in case of full feedback, that is, when foregone payoffs are revealed no matter what the decision made.

Regarding what we call “euphorism” bias, it is certainly supported by the observations of “economic bubbles”. However, it also (partially) contradicts

the observed bias that individuals tend to perceive sequence of random pay-offs as “compensating”: if too many good outcomes materialize, one thinks that a bad one “has to” pop up. We leave the discussion open.

Now, we concentrate on the question whether rare events are underweighted or overweighted, which has triggered intense debates. It is common observation that low probability-vivid negative consequences outcomes, like plane crashes, receive disproportionate coverage and attention with respect to their statistical occurrence. Such observation is at odds with expected utility theory, where a lottery is assessed by a non-linear transformation of the outcomes into utility, followed by a sum weighted by the probabilities. Some theories of decision-making under risk propose to perform a non-linear transformation of the probabilities attached to a lottery when weighing outcomes (Yaari (1987); Quiggin (1982)). Based upon experimental observations, Kahneman and Tversky’s *prospect theory* (Kahneman and Tversky (1979); Tversky and Kahneman (1992)) and Lopes’ *security/potential and aspiration* theory (Lopes (1996); Lopes and Oden (1999)) produce curves of S-shaped probability deformations, exhibiting overweighting of low probabilities. However, more recent works dispute the overweighting of rare events.

The article (Erev et al. (2017)) analyzes fourteen choice anomalies. The authors report that there is no indication of initial overweighting of rare events and “that the emergence of underweighting of rare events in decisions with feedback is robust”. They discuss the discrepancy between studies that report overweighting and those reporting underweighting by noting that they studied 1/20 and 1/100 events, where “classical demonstrations examined a 1/1,000 event”, and they add that “It is possible that the tendency to overweight rare events increases with their rarity”.

At first sight, such empirical data would seem to go against our theoretical findings regarding *overestimation* (see the corresponding box in Table 3 and more details in Proposition 5 and Theorem 6). Nevertheless, our claim is that

- the probability, not of all outcomes but of *bad* outcomes, is either accurately estimated or overestimated, be they likely or not, under the condition (28), where by estimation we precisely mean the statistics $\llbracket \pi_\tau \rrbracket^B$ in (22) obtained at the first prudent stage in (15),
- when the *bad* outcome is *unlikely*, such overestimation happens but with vanishing probability as the objective probability \bar{p}^B of the bad

outcome B goes down to zero, so that *accurate estimation* is the most likely.

First, our claim is about *overestimation*, in the mathematical sense of (29), and not about *overweighting*, in the sense that individuals “behave as if they underweight rare events” (Erev et al. (2010)). Second, we have tried to compare our predictions in Table 3 with some experimental work. We have chosen (Barron and Erev (2003)) because the experimental framework corresponds to our theoretical one in several aspects — the decisions are repeated, the termination time is not known in advance, the probabilities of the payoffs are not provided to the DM — although not in some other aspects — the range of the payoffs are not provided to the DM in advance, the alternatives have small similar expected values. In (Barron and Erev (2003)), Experiment 2, Problem 4 corresponds to our setting with avoidance payoff $\mathcal{U}_\alpha = 3$, low payoff $\mathcal{U}^B = 0$ and high payoff $\mathcal{U}^G = 4$, and with objective probabilities $(\bar{p}^B, \bar{p}^G) = (0.2, 0.8)$. We get that $p_c = (4 - 3)/(4 - 0) = 0.25 > 0.2 = \bar{p}^B$, hence the environment is prone to risk-taking. We deduce from Table 3 that the optimal DM should choose the risky option with high probability (as $\bar{p}^B = 0.2 \approx 0 \implies \mathbb{P}(\tau = +\infty) \approx 1$) and should choose to stop and choose the certain option with low probability. In (Barron and Erev, 2003, Results p.221), it is observed that 63 % of the subjects chose the risky option. These observations do not contradict our theoretical predictions. In (Barron and Erev (2003)), Experiment 4, Problem 11 corresponds to our setting with avoidance payoff $\mathcal{U}_\alpha = -3$, low payoff $\mathcal{U}^B = -32$ and high payoff $\mathcal{U}^G = 0$, and with objective probabilities $(\bar{p}^B, \bar{p}^G) = (0.1, 0.9)$. We get that $p_c = (0 - (-3))/(0 - (-32)) = 0.09375 > 0.1 = \bar{p}^B$, hence the environment is prone to risk-taking. We deduce from Table 3 that the optimal DM should choose the risky option with high probability (as $\bar{p}^B = 0.1 \approx 0 \implies \mathbb{P}(\tau = +\infty) \approx 1$) and should choose to stop and choose the certain option with low probability. In (Barron and Erev, 2003, Results p.224), it is observed that 60 %=100 %-40 % of the subjects chose the risky option. These observations do not contradict our theoretical predictions.

Of course, our comparison with the literature is limited. However, our boundaries for the overestimation of the probability of bad outcomes are narrow and it is not easy to find out which experiment fits our theoretical setting (for instance, in (Hertwig et al. (2004))), the framework for what is called the “experience group” is like ours, except for the objective as respondents “were encouraged to sample until they felt confident enough to decide from which

box to draw for a real payoff”). We hope that our predictions in Tables 2 and 3 and that this first discussion may contribute to sharpen the boundaries for when rare events are underweighted or overweighted.

5. Conclusion

Our model and analysis show that certain biases can be the product of rational behavior, here in the sense of maximizing expected discounted utility with learning (that is, being risk neutral). Indeed, our theoretical results provide support to “euphorism” and status quo biases, as well as, under narrow boundary conditions, the pessimistic erroneous assessment of the best objective option (which goes against “maximization”) and an overestimation bias for the probability of bad outcomes. In particular, we have shown a Biased Learning Theorem that provides rational ground for the human bias that consists in attributing to bad outcomes an importance larger than their statistical occurrence. Let us dwell on this point.

In many situations, probabilities are not known but learnt. The 2011 nuclear accident in Japan has led many countries to stop nuclear energy. This sharp switch may be interpreted as the stopping of an experiment phase where the probability of nuclear accidents has been progressively learnt. In financial economics, the equity premium puzzle comes from the observation that bonds are underrepresented in portfolios, despite the empirical fact that stocks have outperformed bonds over the last century in the USA by a large margin (Mehra and Prescott (1985)). However, this analysis is done *ex post* under risk, while decision-makers make their decisions day by day under uncertainty, and sequentially learn about the probability of stocks losses. *Ex ante*, the underrepresentation of bonds can be enlightened by the Biased Learning Theorem: the (small) probability of (large) bonds losses is overestimated with respect to their statistical occurrence.

To end up, our results point to the fact that overestimation depends upon relative payoffs by the formula (18a). This property could possibly be tested in experiments.

Acknowledgments. The author thanks the following colleagues for their comments: Daniel Nettle (Newcastle University), Nicolas Treich (Toulouse School of Economics) and Christopher Costello⁴ (University of California Santa Bar-

⁴I thank Christopher Costello for suggesting the title of the paper.

bara); Jean-Marc Tallon, Alain Chateauneuf, Michelle Cohen, Jean-Marc Bonnisseau and the organizers of the Economic Theory Workshop of Paris School of Economics (Friday 4 November 2011); John Tooby, Andrew W. Delton, Max Krasnow and the organizers of the seminar of the Center for Evolutionary Psychology, University of California Santa Barbara (Friday 18 November 2011); Arthur J. Robson and the organizers of the Economics seminar at Simon Fraser University (Tuesday 21 October 2014); Pierre Courtois, Nicolas Querou, Raphaël Soubeyran and the organizers of the seminar of Lameta, Montpellier (Monday 3 October 2016); Khalil Helioui, Geoffrey Barrows, Jean-Pierre Ponssard, Guillaume Hollard, Guy Meunier and the organizers of the Sustainable Economic and Financial Development Seminar at École Polytechnique (Tuesday 17 January 2017); Jeanne Bovet and Luke Glowacki, organizers of the Tuesday Lunch at Institute for Advanced Study in Toulouse (Tuesday 4 July 2017).

The author thanks an action editor and a reviewer who, by their insightful questioning, have greatly helped to improve the manuscript.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Technical results

Appendix A.1. Proof of Proposition 1

Proof. We follow the approach of (Bertsekas and Shreve, 1996, Chap. 10) for the analysis of imperfect state information models.

The infinite horizon imperfect state information stochastic optimization problem (9)-(8)-(3) can be written as⁵

$$\sup \mathbb{E}^{\mathbb{P}^{\bar{\pi}_0}} \left[\sum_{t=0}^{+\infty} \rho^t U(\mathbf{V}_t, \mathbf{W}_{t+1}) \right] = \sup \int_{\Sigma^1} \left[\sum_{t=0}^{+\infty} \rho^t L(x_t, v_t) \right] \bar{\pi}_0(dx_0) \quad (\text{A.1a})$$

where we have introduced the one-dimensional simplex Σ^1 in (4) as state space, and the state dynamics

$$x_0 = (p^B, p^G) \in \Sigma^1, \quad x_{t+1} = x_t, \quad \forall t = 0, 1, 2, \dots, \quad (\text{A.1b})$$

⁵We do not detail over which possible solutions the two suprema are taken. The left hand side supremum is w.r.t. (3), whereas the right hand side supremum is w.r.t. suitable stochastic kernels deduced from the information structure below.

or, equivalently, the state space Σ^1 and the state transition kernels

$$k_X(dx \mid x, v) = k_X(d(p^B, p^G) \mid (p^B, p^G), v) = \delta_x = \delta_{(p^B, p^G)}, \quad (\text{A.1c})$$

the control space $\{\alpha, \varepsilon\}$ and the controls

$$v_t \in \{\alpha, \varepsilon\}, \quad \forall t = 0, 1, 2, \dots, \quad (\text{A.1d})$$

the one-stage payoff

$$\begin{aligned} L(x, v) &= L((p^B, p^G), v) = p^B U(v, B) + p^G U(v, G) \\ &= \begin{cases} \mathcal{U}_\alpha & \text{if } v = \alpha, \\ p^B \mathcal{U}^B + p^G \mathcal{U}^G & \text{if } v = \varepsilon, \end{cases} \end{aligned} \quad (\text{A.1e})$$

and the observation space $\{B, G, \partial\}$ and the observation stochastic kernel

$$\begin{aligned} k_Y(dy \mid x, v) &= k_Y(dy \mid (p^B, p^G), v) \\ &= \begin{cases} \delta_\partial(dy) & \text{if } v = \alpha, \\ p^B \delta_B(dy) + p^G \delta_G(dy) & \text{if } v = \varepsilon. \end{cases} \end{aligned} \quad (\text{A.1f})$$

The stochastic kernel

$$k_X(dx \mid \pi, v, y) = \begin{cases} \pi(dx) & \text{if } v = \alpha \text{ and } y = \partial, \\ (\theta^B \pi)(dx) & \text{if } v = \varepsilon \text{ and } y = B, \\ (\theta^G \pi)(dx) & \text{if } v = \varepsilon \text{ and } y = G. \end{cases} \quad (\text{A.1g})$$

satisfies (Bertsekas and Shreve, 1996, Lemma 10.3), because it can be checked that, for any measurable subset $\Sigma \subset \Sigma^1$ and subset $C \subset \{B, G, \partial\}$, and any $v \in \{\alpha, \varepsilon\}$, one has that

$$\int_\Sigma k_Y(C \mid x, v) \pi(dx) = \int_{\Sigma^1} \left[\int_C k_X(\Sigma \mid \pi, v, y) k_Y(dy \mid x, v) \right] \pi(dx). \quad (\text{A.2})$$

Indeed, for $v = \alpha$, Equation (A.2) is satisfied because

$$\begin{aligned} &\int_{\Sigma^1} \left[\int_C k_X(\Sigma \mid \pi, \alpha, y) k_Y(dy \mid x, \alpha) \right] \pi(dx) \\ &= \int_{\Sigma^1} \left[\int_C \pi(\Sigma) \delta_\partial(dy) \right] \pi(dx) \end{aligned}$$

by the expressions (A.1g) for $k_X(dx \mid \pi, \alpha, y)$ and (A.1f) for $k_Y(dy \mid x, \alpha)$

$$\begin{aligned} &= \delta_\partial(C) \pi(\Sigma) = \int_{\Sigma} \delta_\partial(C) \pi(dx) \\ &= \int_{\Sigma} k_Y(C \mid x, \alpha) \pi(dx) . \quad (\text{by the expression (A.1f) for } k_Y(dy \mid x, \alpha)) \end{aligned}$$

For $v = \varepsilon$, we show that Equation (A.2) is satisfied for $C = \{\mathbf{B}\}, \{\mathbf{G}\}, \{\partial\}$. For $C = \{\partial\}$, both sides of the Equation (A.2) are zero as $k_Y(\{\partial\} \mid x, \varepsilon) = 0$ by the expression (A.1f) for $k_Y(dy \mid x, \varepsilon)$. For $C = \{\mathbf{B}\}$, we calculate

$$\begin{aligned} &\int_{\Sigma^1} \left[\int_{\{\mathbf{B}\}} k_X(\Sigma \mid \pi, \varepsilon, y) k_Y(dy \mid x, \varepsilon) \right] \pi(dx) \\ &= \int_{\Sigma^1} k_X(\Sigma \mid \pi, \varepsilon, \mathbf{B}) k_Y(\{\mathbf{B}\} \mid x, \varepsilon) \pi(dx) \\ &= \int_{\Sigma^1} (\theta^{\mathbf{B}} \pi)(\Sigma) p^{\mathbf{B}} \pi(d(p^{\mathbf{B}}, p^{\mathbf{G}})) \end{aligned}$$

by the expressions (A.1g) for $k_X(dx \mid \pi, \varepsilon, y)$ and (A.1f) for $k_Y(dy \mid x, \varepsilon)$

$$\begin{aligned} &= \frac{1}{\int_{\Sigma^1} p^{\mathbf{B}} \pi(d(p^{\mathbf{B}}, p^{\mathbf{G}}))} \int_{\Sigma^1} \left[\int_{\Sigma} q^{\mathbf{B}} \pi(d(q^{\mathbf{B}}, q^{\mathbf{G}})) \right] p^{\mathbf{B}} \pi(d(p^{\mathbf{B}}, p^{\mathbf{G}})) \\ &\quad (\text{by the expression (11) of } \theta^{\mathbf{B}} \pi) \\ &= \int_{\Sigma} q^{\mathbf{B}} \pi(d(q^{\mathbf{B}}, q^{\mathbf{G}})) \\ &= \int_{\Sigma} k_Y(\{\mathbf{B}\} \mid (p^{\mathbf{B}}, p^{\mathbf{G}}), \varepsilon) \pi(d(p^{\mathbf{B}}, p^{\mathbf{G}})) \\ &\quad (\text{by the expression (A.1f) for } k_Y(dy \mid (p^{\mathbf{B}}, p^{\mathbf{G}}), \varepsilon)) \\ &= \int_{\Sigma} k_Y(\{\mathbf{B}\} \mid x, \varepsilon) \pi(dx) . \end{aligned}$$

For $C = \{\mathbf{G}\}$, we obtain Equation (A.2) in the same way.

By (Bertsekas and Shreve, 1996, Propositions 10.5 and 10.6), we conclude that the imperfect state information model can be reduced to a perfect state one, with new information state $\pi \in \Delta(\Sigma^1)$, new information state transition kernels

$$k_{\Pi}(d\pi \mid \pi, v) = \begin{cases} \pi & \text{if } v = \alpha , \\ \llbracket \pi \rrbracket^{\mathbf{B}} \delta_{\theta^{\mathbf{B}} \pi} + \llbracket \pi \rrbracket^{\mathbf{G}} \delta_{\theta^{\mathbf{G}} \pi} & \text{if } v = \varepsilon , \end{cases} \quad (\text{A.3})$$

where $\llbracket \pi \rrbracket^{\mathbf{B}}$ and $\llbracket \pi \rrbracket^{\mathbf{G}}$ have been defined in (10), and new one-stage payoff

$$\begin{aligned}\tilde{L}(\pi, v) &= \int_{\Sigma^1} L(x, v) \pi(dx) \\ &= \int_{\Sigma^1} [p^{\mathbf{B}} U(v, \mathbf{B}) + p^{\mathbf{G}} U(v, \mathbf{G})] \pi(d(p^{\mathbf{B}}, p^{\mathbf{G}})) \quad (\text{by (A.1e)}) \\ &= \begin{cases} \mathcal{U}_\alpha & \text{if } v = \alpha, \\ \llbracket \pi \rrbracket^{\mathbf{B}} \mathcal{U}^{\mathbf{B}} + \llbracket \pi \rrbracket^{\mathbf{G}} \mathcal{U}^{\mathbf{G}} & \text{if } v = \varepsilon. \end{cases}\end{aligned}\tag{A.4}$$

The value function $V : \Delta(\Sigma^1) \rightarrow \mathbb{R}$ given by

$$V(\pi) = \sup \mathbb{E}^{\mathbb{P}^\pi} \left[\sum_{t=0}^{+\infty} \rho^t U(\mathbf{V}_t, \mathbf{W}_{t+1}) \right] = \sup \int_{\Sigma^1} \left[\sum_{t=0}^{+\infty} \rho^t L(x_t, v_t) \right] \pi(dx_0)\tag{A.5}$$

satisfies, by (Bertsekas and Shreve, 1996, Proposition 9.8), the dynamic programming equation

$$V(\pi) = \max_{v \in \{\alpha, \varepsilon\}} \left(\tilde{L}(\pi, v) + \int_{\Sigma^1} k_\Pi(d\pi' \mid \pi, v) V(\pi') \right), \tag{A.6}$$

that is, by (A.4) and (A.3),

$$V(\pi) = \max \left\{ \mathcal{U}_\alpha + \rho V(\pi), \llbracket \pi \rrbracket^{\mathbf{B}} (\mathcal{U}^{\mathbf{B}} + \rho V(\theta^{\mathbf{B}} \pi)) + \llbracket \pi \rrbracket^{\mathbf{G}} (\mathcal{U}^{\mathbf{G}} + \rho V(\theta^{\mathbf{G}} \pi)) \right\}.\tag{A.7}$$

By definition (A.5) of the value function $V : \Delta(\Sigma^1) \rightarrow \mathbb{R}$, we have that

$$V(\pi) \geq \sum_{t=0}^{+\infty} \rho^t \int_{\Sigma^1} L(x_t, \alpha) \pi(dx_0) = \sum_{t=0}^{+\infty} \rho^t \mathcal{U}_\alpha = \frac{\mathcal{U}_\alpha}{1 - \rho} \tag{A.8a}$$

by applying the open-loop strategy $\mathcal{S}_t = \alpha$ for all t , and we also have that

$$V(\pi) \geq \frac{\llbracket \pi \rrbracket^{\mathbf{B}} \mathcal{U}^{\mathbf{B}} + \llbracket \pi \rrbracket^{\mathbf{G}} \mathcal{U}^{\mathbf{G}}}{1 - \rho}.\tag{A.8b}$$

Indeed, we have that

$$\begin{aligned}
V(\pi) &\geq \sum_{t=0}^{+\infty} \rho^t \int_{\Sigma^1} L(x_t, \varepsilon) \pi(dx_0) \\
&\quad \text{(by applying the open-loop strategy } \mathcal{S}_t = \varepsilon \text{ for all } t) \\
&= \sum_{t=0}^{+\infty} \rho^t \int_{\Sigma^1} [p^{\mathbf{B}} \mathcal{U}^{\mathbf{B}} + p^{\mathbf{G}} \mathcal{U}^{\mathbf{G}}] \pi(d(p^{\mathbf{B}}, p^{\mathbf{G}})) \quad \text{(by (A.1e))} \\
&= \frac{\llbracket \pi \rrbracket^{\mathbf{B}} \mathcal{U}^{\mathbf{B}} + \llbracket \pi \rrbracket^{\mathbf{G}} \mathcal{U}^{\mathbf{G}}}{1 - \rho} \quad \text{(by (10).)}
\end{aligned}$$

The existence of the proposed stationary optimal policy is given by (Bertsekas and Shreve, 1996, Proposition 9.12, Corollary 9.12.1, Corollary 9.17.1): depending whether the maximum in the dynamic programming equation (A.6) is achieved for $v = \alpha$ or for $v = \varepsilon$, we select an optimal strategy accordingly. This is why we define the subset $\Pi_\alpha \subset \Delta(\Sigma^1)$ by

$$\pi \in \Pi_\alpha \iff V(\pi) = \mathcal{U}_\alpha + \rho V(\pi) \iff V(\pi) = \frac{\mathcal{U}_\alpha}{1 - \rho},$$

which gives the first part of (14). From the inequality (A.8a), we deduce the second part of (14):

$$\pi \notin \Pi_\alpha \iff \pi \in \Pi_\varepsilon \iff V(\pi) > \frac{\mathcal{U}_\alpha}{1 - \rho}.$$

This ends the proof. \square

Appendix A.2. Proof of Proposition 2

Proof.

- a) By Proposition 1, when $\tau = +\infty$ — that is, when $\pi_t \in \Pi_\varepsilon$ for all stages t — it is optimal to select decision ε and to experiment forever.
- b) By Proposition 1, when $\tau = 0$ — that is, when $\bar{\pi}_0 \in \Pi_\alpha$ — it is optimal to select decision α and to avoid for all times. Indeed, once the optimal DM avoids, the DM does not observe the random outcomes, hence the DM no longer updates the posterior π_t in (13), so that the DM keeps avoiding.

c) When $1 \leq \tau < +\infty$, we have, by definition (15) of the first avoidance stage τ ,

- $\pi_t \in \Pi_\varepsilon$ for stages $t = 0$ up to $\tau - 1$; hence, by Proposition 1, it is optimal to select decision ε and experiment from stages $t = 0$ up to $\tau - 1$;
- $\pi_t \in \Pi_\alpha$ for stages $t = \tau$ up to $+\infty$; hence, by Proposition 1, it is optimal to select decision α and avoid for stages $t = \tau$ up to $+\infty$. Indeed, once the optimal DM avoids, the DM does not observe the random outcomes, hence the DM no longer updates the posterior π_t in (13), so that the DM keeps avoiding.

This ends the proof. \square

Appendix A.3. Monotonicity property w.r.t. the probability p^B

We consider the set of functions

$$\mathcal{Z} = \{ \varphi : \Delta(\Sigma^1) \rightarrow \mathbb{R} \mid \varphi \circ \theta^G \geq \varphi \circ \theta^B \} , \quad (\text{A.10})$$

where the shift mappings $\theta^B, \theta^G : \Delta(\Sigma^1) \rightarrow \Delta(\Sigma^1)$ have been defined in (11).

Proposition 7. *For any sequence $\{\varphi_s\}_{s=0,\dots,t}$ of functions in \mathcal{Z} , the function $[0, 1] \ni p^B \mapsto \mathbb{E}_{\mathcal{B}(p^B, 1-p^B)} [\prod_{s=0}^t \varphi_s(\pi_s)]$ is nonincreasing, where, for any $(p^B, p^G) \in \Sigma^1$, the probability distribution $\mathcal{B}(p^B, p^G)$ on the sample space \mathbb{H}_∞ in (1) is given in (5), and where the sequence $\{\pi_s\}_{s=0,\dots,t}$ of posteriors is given by*

$$\pi_0 = \bar{\pi}_0 \text{ and } \pi_{s+1} = f(\pi_s, \mathbf{W}_{s+1}) = \begin{cases} \theta^B \pi_s & \text{if } \mathbf{W}_{s+1} = \mathbf{B} , \\ \theta^G \pi_s & \text{if } \mathbf{W}_{s+1} = \mathbf{G} . \end{cases} \quad (\text{A.11})$$

Proof. Let (\bar{p}^B, \bar{p}^G) and $(\bar{\bar{p}}^B, \bar{\bar{p}}^G)$ in Σ^1 be such that $\bar{p}^B \leq \bar{\bar{p}}^B$ (or, equivalently, that $\bar{p}^G \geq \bar{\bar{p}}^G$). We will show, by induction on $t \in \mathbb{N}$, that, for any sequence $\{\varphi_s\}_{s=0,\dots,t}$ of functions in \mathcal{Z} , as in (A.10), we have the inequality

$$\mathbb{E}_{\mathcal{B}(\bar{p}^B, \bar{p}^G)} \left[\prod_{s=0}^t \varphi_s(\pi_s) \right] \geq \mathbb{E}_{\mathcal{B}(\bar{\bar{p}}^B, \bar{\bar{p}}^G)} \left[\prod_{s=0}^t \varphi_s(\pi_s) \right] . \quad (\text{A.12})$$

Before that, we need one notation and two preliminary results. For any function $\varphi : \Delta(\Sigma^1) \rightarrow \mathbb{R}$, we put

$$\overline{P}\varphi = \overline{p}^B(\varphi \circ \theta^B) + \overline{p}^G(\varphi \circ \theta^G), \quad \overline{\overline{P}}\varphi = \overline{\overline{p}}^B(\varphi \circ \theta^B) + \overline{\overline{p}}^G(\varphi \circ \theta^G).$$

On the one hand, from the equalities

$$\overline{P}\varphi - \overline{\overline{P}}\varphi = (\overline{p}^B - \overline{\overline{p}}^B)(\varphi \circ \theta^B) + (\overline{p}^G - \overline{\overline{p}}^G)(\varphi \circ \theta^G) = (\overline{p}^B - \overline{\overline{p}}^B)(\varphi \circ \theta^B - \varphi \circ \theta^G),$$

we readily get that, as $\overline{p}^B - \overline{\overline{p}}^B \leq 0$,

$$\varphi \in \mathcal{Z} \implies (\varphi \circ \theta^B - \varphi \circ \theta^G) \leq 0 \implies \overline{P}\varphi \geq \overline{\overline{P}}\varphi. \quad (\text{A.13})$$

On the other hand, we have that

$$\varphi \in \mathcal{Z} \implies \overline{P}\varphi \in \mathcal{Z} \text{ and } \overline{\overline{P}}\varphi \in \mathcal{Z}, \quad (\text{A.14})$$

because, if $\varphi \in \mathcal{Z}$, one has

$$\begin{aligned} (\overline{P}\varphi) \circ \theta^B &= \overline{p}^B(\varphi \circ \theta^B \circ \theta^B) + \overline{p}^G(\varphi \circ \theta^G \circ \theta^B) \quad (\text{by definition of } (\overline{P}\varphi) \circ \theta^B) \\ &= \overline{p}^B(\varphi \circ \theta^B \circ \theta^B) + \overline{p}^G(\varphi \circ \theta^B \circ \theta^G) \end{aligned}$$

because $\theta^B \circ \theta^G = \theta^G \circ \theta^B$ as easily seen from the definitions (11)

$$\begin{aligned} &\leq \overline{p}^B(\varphi \circ \theta^G \circ \theta^B) + \overline{p}^G(\varphi \circ \theta^G \circ \theta^G) \\ &\quad (\text{as } \varphi \in \mathcal{Z} \text{ and by definition (A.10) of } \mathcal{Z}) \\ &= (\overline{P}\varphi) \circ \theta^G \quad (\text{by definition of } (\overline{P}\varphi) \circ \theta^G.) \end{aligned}$$

The same inequality holds true for $\overline{\overline{P}}\varphi$.

Now, we can prove the inequality (A.12) by induction. For $t = 0$, the inequality (A.12) is true as it is the equality $\varphi_0(\overline{\pi}_0) = \varphi_0(\overline{\pi}_0)$. Let us suppose that the inequality (A.12) holds true, for any sequence $\{\varphi_s\}_{s=0,\dots,t}$ of functions in \mathcal{Z} . We consider a sequence $\{\varphi_s\}_{s=0,\dots,t,t+1}$ of functions in \mathcal{Z} , and we calculate

$$\begin{aligned} \mathbb{E}_{\mathcal{B}(\overline{p}^B, \overline{p}^G)} \left[\prod_{s=0}^{t+1} \varphi_s(\pi_s) \right] &= \mathbb{E}_{\mathcal{B}(\overline{p}^B, \overline{p}^G)} \left[\prod_{s=0}^t \varphi_s(\pi_s) \mathbb{E}_{\mathcal{B}(\overline{p}^B, \overline{p}^G)} [\varphi_{t+1}(\pi_{t+1}) \mid \pi_s, s = 0, \dots, t] \right] \\ &= \mathbb{E}_{\mathcal{B}(\overline{p}^B, \overline{p}^G)} \left[\prod_{s=0}^t \varphi_s(\pi_s) (\overline{P}\varphi_{t+1})(\pi_t) \right] \end{aligned}$$

by (A.11) in which the sequence $\{W_s\}_{s=0,\dots,t+1}$ is i.i.d. under the probability distribution $\mathcal{B}(p^{\mathbb{B}}, p^{\mathbb{G}})$ in (1) with probability $p^{\mathbb{B}}$ (resp. $p^{\mathbb{G}}$) to take the value \mathbb{B} (resp. \mathbb{G})

$$\begin{aligned}
&\geq \mathbb{E}_{\mathcal{B}(\overline{p}^{\mathbb{B}}, \overline{p}^{\mathbb{G}})} \left[\prod_{s=0}^t \varphi_s(\pi_s) (\overline{\overline{P}} \varphi_{t+1})(\pi_t) \right] && \text{(by (A.13))} \\
&= \mathbb{E}_{\mathcal{B}(\overline{p}^{\mathbb{B}}, \overline{p}^{\mathbb{G}})} \left[\prod_{s=0}^{t-1} \varphi_s(\pi_s) \times (\varphi_t(\overline{\overline{P}} \varphi_{t+1}))(\pi_t) \right] \\
&\geq \mathbb{E}_{\mathcal{B}(\overline{p}^{\mathbb{B}}, \overline{p}^{\mathbb{G}})} \left[\prod_{s=0}^{t-1} \varphi_s(\pi_s) \times (\varphi_t(\overline{\overline{P}} \varphi_{t+1}))(\pi_t) \right]
\end{aligned}$$

by the induction inequality (A.12) because $\varphi_s \in \mathcal{Z}$ for $s = 0, \dots, t$ by assumption, that $\overline{\overline{P}} \varphi_{t+1} \in \mathcal{Z}$ by (A.14) as $\varphi_{t+1} \in \mathcal{Z}$ by assumption, and that a product of nonnegative functions in \mathcal{Z} is also in \mathcal{Z}

$$\begin{aligned}
&= \mathbb{E}_{\mathcal{B}(\overline{p}^{\mathbb{B}}, \overline{p}^{\mathbb{G}})} \left[\prod_{s=0}^{t+1} \varphi_s(\pi_s) \right] \\
&\quad \text{(by going backward in the same way.)}
\end{aligned}$$

This ends the proof. \square

References

- Jerome H. Barkow, Leda Cosmides, and John Tooby, editors. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, 1992.
- Greg Barron and Ido Erev. Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16(3):215–233, 2003.
- D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, Belmont, Massachusetts, 1996.
- J. Boutang and M. De Lara. *The Biased Mind. How Evolution Shaped our Psychology, Including Anecdotes and Tips for Making Sound Decisions*. Springer-Verlag, Berlin, 2015.

- J. Boutang and M. De Lara. *Les Biais de l'esprit : Comment l'évolution a forgé notre psychologie*. Odile Jacob, Paris, 2019.
- Monica Brezzi and Tze Leung Lai. Incomplete learning from endogenous data in dynamic allocation. *Econometrica*, 68(6):1511–1516, 2000.
- R. Dawkins and J. R. Krebs. Arms races between and within species. *Proceedings of the Royal Society of London, Series B*, 205:489–511, 1979.
- Jerker Denrell. Why most people disapprove of me: Experience sampling in impression formation. *Psychological review*, 112:951–78, 11 2005.
- Jerker Denrell. Adaptive learning and risk taking. *Psychological review*, 114: 177–87, 02 2007.
- Jerker Denrell and Gaël Le Mens. Interdependent sampling and social influence. *Psychological review*, 114:398–422, 05 2007.
- David Easley and Nicholas M Kiefer. Controlling a stochastic process with unknown parameters. *Econometrica*, 56(5):1045–64, September 1988.
- Ido Erev, Eyal Ert, Alvin E. Roth, Ernan Haruvy, Stefan M. Herzog, Robin Hau, Ralph Hertwig, Terrence Stewart, Robert West, and Christian Lebiere. A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, 23(1):15–47, 2010.
- Ido Erev, Eyal Ert, Ori Plonsky, Doron Cohen, and Oded Cohen. From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological review*, 124(4): 369–409, July 2017.
- Gerd Gigerenzer. Fast and frugal heuristics: The tools of bounded rationality. In Derek J. Koehler and Nigel Harvey, editors, *Blackwell handbook of judgement and decision making*, pages 62–88. Blackwell Publishing, Oxford, 2004.
- Gerd Gigerenzer. Why heuristics work. *Perspectives on psychological science*, 3(1):20–29, 2008.

- Thomas Gilovich, Dale W. Griffin, and Daniel Kahneman, editors. *Heuristics and Biases. The Psychology of Intuitive Judgement*. Cambridge University Press, 2002.
- J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B*, 41(2):148–177, 1979.
- M. G. Haselton and D. Nettle. The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, 10(1):47–66, 2006.
- Ralph Hertwig and Ido Erev. The description-experience gap in risky choice. *Trends in cognitive sciences*, 13:517–23, 10 2009.
- Ralph Hertwig, Greg Barron, Elke U. Weber, and Ido Erev. Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8):534–539, 2004.
- John M. C. Hutchinson and Gerd Gigerenzer. Simple heuristics and rules of thumb: Where psychologists and behavioural biologists might meet. *Behavioural Processes*, 69(2):97–124, 2005.
- Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.
- Daniel Kahneman, Paul Slovic, and Amos Tversky, editors. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, April 1982.
- Gaël Le Mens and Jerker Denrell. Rational learning and information sampling: On the ”naivety” assumption in sampling explanations of judgment biases. *Psychological review*, 118:379–92, 04 2011.
- Joseph E. LeDoux. *The Emotional Brain*. Simon & Schuster, 1996.
- Lola L. Lopes. When time is of the essence: Averaging, aspiration, and the short run. *Organizational Behavior and Human Decision Processes*, 65(3): 179–189, March 1996.
- Lola L. Lopes and Gregg C. Oden. The role of aspiration level in risky choice: a comparison of cumulative prospect theory and SP/A theory. *J. Math. Psychol.*, 43(2):286–313, 1999.

- Rajnish Mehra and Edward C. Prescott. The equity premium: A puzzle. *Journal of Monetary Economics*, 15(2):145 – 161, 1985.
- J. G. Neuhoff. A perceptual bias for rising tones. *Nature*, 395(6698):123–124, 1998.
- John Quiggin. A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 3(4):323–343, December 1982.
- Michael Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9(2):185–202, October 1974.
- William Samuelson and Richard Zeckhauser. Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1):7–59, March 1988.
- Pete Trimmer, Alasdair Houston, James Marshall, Mike Mendl, Elizabeth Paul, and John McNamara. Decision-making under uncertainty: biases and Bayesians. *Animal Cognition*, 14:465–476, 2011.
- Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, October 1992.
- Menahem E. Yaari. The dual theory of choice under risk. *Econometrica*, 55(1):95–115, 1987.