A Note on Self-confirming Equilibrium and Stochastic Control

Pierpaolo Battigalli and Giacomo Lanzani

First version 2016, current version November 2024

Abstract

This pedagogical note relates the self-confirming equilibrium concept for decision problems with feedback (cf. Battigalli *et al.*, *J. Econ. Theory*, 183 (2019), 740-785) with the limit behavior of solutions to stochastic control problems (see Easley and Kiefer, *Econometrica*, 56 (1988), 1045-1064).

1 Introduction

This pedagogical note relates the self-confirming equilibrium concept for decision problems with feedback (cf. Battigalli *et al.*, 2015, 2019) with the limit behavior of solutions to stochastic control problems (see Easley and Kiefer, 1988). This complements the sections on self-confirming equilibrium of Chapters 6-8 of *Game Theory: Analysis of Strategic Thinking* by Battigalli, Catonini, and De Vito.[1] Focusing on stationary decision problems with uncertainty and discounting, we first show that, if the stochastic process of beliefs and actions of a rational decision maker converges in finite time, then the rest point reached must be a self-confirming equilibrium of the static decision problem. This simple result is included for its pedagogical value. Next, adapting results on stochastic control (e.g., Easley and Kiefer, 1988), we show that the process of beliefs and actions converges almost surely[2] to a self-confirming equilibrium of the static decision problem. To keep the probability analysis simple, we focus on finite decision problems where the set of possible probability measures over states of nature is also finite. A simple example with 2 actions and 2 probability models illustrates the analysis.

2 Self-confirming equilibrium in decision problems

The self-confirming equilibrium concept (SCE) for *static decision problems with feedback* can be framed by the following list of primitive elements (cf. Battigalli *et al.*, 2019):

[1] See also the comments in Chapter 9 on the relevance of SCEs of the appropriately defined strategic form of a sequential game with feedback.

[2] In finite or infinite time.

- $a \in A$, *actions*;
- $s \in S$, *states of nature*;
- $\sigma \in \Sigma \subseteq \Delta(S)$, posited set of *stochastic models*;
- $y \in Y$, *consequences* (e.g., monetary or consumption consequences);
- $g : A \times S \to Y$, *consequence function* (or *game form*);
- $m \in M$, *observed outcomes, or messages*;
- $f : A \times S \to M$, *feedback function*;
- $v : Y \to \mathbb{R}$, von Neumann-Morgenstern *utility function*.

Assumptions: We assume for simplicity that all the sets above are *finite*. Furthermore, we assume that the *decision problem with feedback satisfies* **observed payoffs** *(or, observed consequences), that is, there is a function* $\bar{g} : A \times M \to Y$ *such that*

$$\forall a \in A, \forall s \in S,\ g(a,s) = \bar{g}\left(a, f(a,s)\right).$$

Let $u\left(a,\sigma\right)$ denote the objective expected utility of action a given probability model σ, that is,

$$u(a,\sigma) = \sum_s v\left(g\left(a,s\right)\right)\sigma\left(s\right) = \sum_s v\left(\bar{g}\left(a, f(a,s)\right)\right)\sigma\left(s\right),$$

where the second equality follows from the assumption of observed payoffs. Similarly we let $u\left(a,\mu\right)$ denote the subjective expected payoff of a given belief $\mu \in \Delta\left(\Sigma\right)$:

$$u\left(a,\mu\right) = \sum_\sigma u(a,\sigma)\mu(\sigma),$$

where the expectation can be expressed as a summation because we assumed for simplicity that Σ is finite.[3] Note, each belief $\mu \in \Delta\left(\Sigma\right)$ yields a "**conjecture**" σ_μ (called "predictive distribution" in statistics) defined by

$$\sigma_\mu(s) = \sum_\sigma \sigma(s)\mu(\sigma)$$

for all $s \in S$. Note that, by definition, $u\left(a,\mu\right) = u\left(a,\sigma_\mu\right)$.

The reason why we work with beliefs about probability models instead of beliefs about states (conjectures) is twofold. First, we can write the equilibrium conditions more explicitly and transparently. Second, if we analyze the repeated decision problem with *i.i.d.* draws of states according to the unknown probability model σ^*, then

[3] For example, if we have an urn with n balls of k colors, the set of possible urn compositions is finite. If Σ is a compact subset of the simplex $\Delta\left(S\right)$, then $u(a,\mu) = \int_\Sigma u(a,\sigma)\mu\left(\mathrm{d}\sigma\right)$.

Bayesian learning can be expressed by a random sequence of beliefs over Σ updated according to Bayes rule.

Finally, for each $a \in A$, recall that $f_a : S \to M$ denotes the *section* of f at a ($f_a(s) = f(a,s)$ for all $s \in S$) and that $\sigma \mapsto \sigma \circ f_a^{-1}$ denotes the *pushforward map* from $\Delta(S)$ to $\Delta(M)$ induced by f_a, that is

$$\forall m \in M,\ \sigma \circ f_a^{-1}(m) = \sigma\left(\{s \in S : f(a,s) = m\}\right).$$

Definition: A *self-confirming equilibrium* (SCE) of the decision problem with feedback is a triple (a^*, μ^*, σ^*) such that

1. (subjective best reply) $a^* \in \arg\max_{a \in A} u(a, \mu^*)$,
2. (confirmed belief) $\mu^*\left(\left\{\sigma \in \Sigma : \sigma \circ f_{a^*}^{-1} = \sigma^* \circ f_{a^*}^{-1}\right\}\right) = 1$.

For each $\sigma^* \in \Sigma$, a *self-confirming equilibrium at* σ^* is a pair (a^*, μ^*) such that (a^*, μ^*, σ^*) is an SCE. We let $SCE(\sigma^*)$ denote the set of self-confirming equilibria at σ^*.

Remark 1 *If (a^*, μ^*, σ^*) is a self-confirming equilibrium, then $a^* \in \arg\max_{a \in A} u(a, \sigma_{\mu^*})$ and conjecture σ_{μ^*} is confirmed, that is, $\sigma_{\mu^*} \circ f_{a^*}^{-1} = \sigma^* \circ f_{a^*}^{-1}$. Thus, if $\sigma_{\mu^*} \in \Sigma$, $(a^*, \delta_{\sigma_{\mu^*}}, \sigma^*)$ is a self-confirming equilibrium as well.*

Consider the infinite repetition of the static decision problem, where the sequence of states of nature is *i.i.d.* with unknown marginal distribution σ^*; the corresponding product measure on $S^{\mathbb{N}}$ is denoted $\sigma^{*,\infty}$. The Decision Maker (DM) is characterized by discount factor $\beta \in [0,1)$ and prior belief $\mu_0 \in \Delta(\Sigma)$. An optimal strategy[4] of DM is one that solves

$$\max_{\alpha} \mathbb{E}_{\mu_0}\left(\sum_{t=0}^{\infty} \beta^t u\left(\mathbf{a}_t^{\alpha}, \boldsymbol{\mu}_t^{\alpha}\right)\right),$$

where $\alpha = (\alpha_t)_{t \in \mathbb{N}_0}$ is the strategy and $(\mathbf{a}_t^{\alpha}, \boldsymbol{\mu}_t^{\alpha})_{t \in \mathbb{N}_0}$ is the sequence of random actions and Bayes-updated beliefs induced by α given the prior μ_0 (in general, we use **boldface** letters to denote random variables). One can show by standard compactness-continuity arguments that the set of optimal strategies is not empty. Furthermore, among the optimal strategies there is always at least one such that α_t is determined only by the belief at the beginning of period t, that is μ_t. For example, α_0 depends on the prior belief μ_0, and α_1 depends on the updated belief $\mu_1(\cdot|a_0, m_0)$, but—given this—it does not directly depend on the action a_0 and message m_0 of the first period. The strategies that depend only on updated beliefs are called **stationary**. Thus, if the optimal strategy given prior μ_0 is unique, then it must be stationary.

[4] A **strategy** in this case is a sequence $\alpha = (\alpha_t)_{t \in \mathbb{N}_0}$ with $\alpha_0 = a_0 \in A$, and $\alpha_t : (A \times M)^{t-1} \to A$ for each $t \in \mathbb{N}$; that is, α_t selects an action in period t as a function of the history of actions taken and messages observed in previous periods $\tau = 0, ..., t-1$.

According to Bayes rule, the belief at the beginning of period $t = 1$ after DM chose a_0 and observed m_0 at the end of period $t = 0$ is given by the formula

$$\begin{aligned}\mu_1(\sigma'|a_0, m_0) &= \frac{\mathbb{P}_{\mu_0,a_0}(m_0, \sigma')}{\mathbb{P}_{\mu_0,a_0}(m_0)} = \frac{\mathbb{P}_{\mu_0,a_0}(m_0|\sigma')\,\mathbb{P}_{\mu_0}(\sigma')}{\sum_\sigma \mathbb{P}_{\mu_0,a_0}(m_0|\sigma)\,\mathbb{P}_{\mu_0}(\sigma)} \\ &= \frac{\left(\sigma' \circ f_{a_0}^{-1}\right)(m_0) \times \mu_0(\sigma')}{\sum_\sigma \left(\sigma \circ f_{a_0}^{-1}\right)(m_0) \times \mu_0(\sigma)}\end{aligned}$$

for each $\sigma' \in \Sigma$, provided that the denominator is positive. Since the observed message depends on the random state $\mathbf{s}_0$, the belief $\boldsymbol{\mu}_1^\alpha$ at the beginning of period $t = 1$ is random as well; hence, also the action $\mathbf{a}_1^\alpha = \alpha_1(\boldsymbol{\mu}_1)$ is random. In general, for every realization $\mu \in \Delta(\Sigma)$ of the random updated belief $\boldsymbol{\mu}_t^\alpha$, the updated belief of period $t+1$ given action a and message m is given by the **Bayes map**

$$B : \Delta(\Sigma) \times A \times M \to \Delta(\Sigma)$$

defined by

$$B(\mu, a, m)(\sigma') = \frac{\left(\sigma' \circ f_a^{-1}\right)(m) \times \mu(\sigma')}{\sum_\sigma \left(\sigma \circ f_a^{-1}\right)(m) \times \mu(\sigma)}$$

for each $\sigma' \in \Sigma$. It is immediate to see that B is continuous with respect to its first argument at every point where the denominator is strictly positive.[5]

Stationary strategies can be described as functions $\alpha : \Delta(\Sigma) \to A$. Given a stationary strategy $\mu_{t-1} \mapsto \alpha(\mu_{t-1})$ $(t \in \mathbb{N})$, Bayes rule yields a stochastic process $(\mathbf{a}_t^\alpha, \boldsymbol{\mu}_t^\alpha)_{t\in\mathbb{N}_0} = (\alpha(\boldsymbol{\mu}_t^\alpha), \boldsymbol{\mu}_t^\alpha)_{t\in\mathbb{N}_0}$ of actions and beliefs. The objective probabilities of each finite sequence of actions and beliefs are determined by the true marginal measure over states. For example, let $a_0 = \alpha(\mu_0)$, then, for each belief $\bar{\mu}_1$,

$$\mathbb{P}_{\sigma^*}(\boldsymbol{\mu}_1^\alpha = \bar{\mu}_1) = \sigma^*\left(\{s \in S : \mu_1(\cdot|a_0, f_{a_0}(s)) = \bar{\mu}_1\}\right);$$

with this, for each action $\bar{a}_1$,

$$\mathbb{P}_{\sigma^*}(\mathbf{a}_1^\alpha = \bar{a}_1) = \mathbb{P}_{\sigma^*}(\bar{a}_1) = \sum_{\bar{\mu}_1 : \alpha_1(\bar{\mu}_1) = \bar{a}_1} \mathbb{P}_{\sigma^*}(\boldsymbol{\mu}_1^\alpha = \bar{\mu}_1).$$

In general, $\boldsymbol{\mu}_t^\alpha : S^{\{0,\ldots,t-1\}} \to \Delta(\Sigma)$ is a random belief with realizations $\boldsymbol{\mu}_t^\alpha(s_0, ..., s_{t-1})$ whose objective, but unknown probability is $\prod_{k=0}^{t-1} \sigma^*(s_k)$.[6]

[5] Pointwise convergence is an immediate consequence of the fact that $\mu_{t+1}(\sigma'|a_t, m_t)$ is a ratio between two compositions of continuos functions of μ_t. Since $\mu_{t+1}(\cdot|a_t, m_t)$ is a vector, pointwise convergence coincides with uniform convergence.

[6] When we consider random limit beliefs, they are defined as random variables $\boldsymbol{\mu}_\infty : S^{\mathbb{N}_0} \to \Delta(\Sigma)$, where the set of infinite sequences of states $s^\infty = (s_0, s_1, ...) \in S^{\mathbb{N}_0}$ is endowed with the smallest sigma-algebra generated by the collection of subsets

$$\left\{s^\infty \in S^{\mathbb{N}_0} : s_0 = \bar{s}_0, ..., s_t = \bar{s}_t\right\},\ t \in \mathbb{N}_0,\ (\bar{s}_0, ..., \bar{s}_t) \in S^{t+1}$$

(called elementary "cylinders").

We are interested in the limit behavior and beliefs of a rational DM. Suppose, for the sake of the discussion, that a steady state (a^*, μ^*) is reached in finite time t along some possible path of the stochastic process (there are non-trivial examples of decision problems and paths where the steady state is reached in finite time). Then it must be the case that the Bayes-update of belief μ^* given action a^* is the same as μ^*, that is, for each $\sigma' \in \text{Supp}(\mu^*)$, and each m that can be observed with positive subjective probability under μ^* (i.e., each m such that $\sum_\sigma \left(\sigma \circ f_{a^*}^{-1}\right)(m) \times \mu^*(\sigma) > 0$)

$$B\left(\mu^*, a^*, m\right)(\sigma') = \frac{\left(\sigma' \circ f_{a^*}^{-1}\right)(m) \times \mu^*(\sigma')}{\sum_\sigma \left(\sigma \circ f_{a^*}^{-1}\right)(m) \times \mu^*(\sigma)} = \mu^*(\sigma').$$

In this case we say that μ^* is **invariant** at a^*.

Lemma 2 *Belief μ^* is invariant at action a^* if and only if all the stochastic models in Supp(μ^*) yield the same distribution on M given a^*, that is, $\sigma \circ f_{a^*}^{-1} = \sigma' \circ f_{a^*}^{-1}$ for all $\sigma, \sigma' \in$Supp(μ^*).*

Proof (If) Condition

$$\forall \sigma, \sigma' \in \text{Supp}(\mu^*),\ \sigma \circ f_{a^*}^{-1} = \sigma' \circ f_{a^*}^{-1},$$

means that the pushforward map $\sigma \mapsto \sigma \circ f_{a^*}^{-1}$ is constant on $\text{Supp}(\mu^*)$. Then, for each $\sigma' \in \text{Supp}(\mu^*)$ and each $m \in M$,

$$\sum_\sigma \left(\sigma \circ f_{a^*}^{-1}\right)(m) \times \mu^*(\sigma) = \left(\sigma' \circ f_{a^*}^{-1}\right)(m) \sum_\sigma \mu^*(\sigma) = \left(\sigma' \circ f_{a^*}^{-1}\right)(m)$$

(because $\sum_\sigma \mu^*(\sigma) = 1$). Therefore, for each m such that $\sum_\sigma \left(\sigma \circ f_{a^*}^{-1}\right)(m) \times \mu^*(\sigma) > 0$,

$$\frac{\left(\sigma' \circ f_{a^*}^{-1}\right)(m) \times \mu^*(\sigma')}{\sum_\sigma \left(\sigma \circ f_{a^*}^{-1}\right)(m) \times \mu^*(\sigma)} = \mu^*(\sigma').$$

(Only if) Suppose that, for every $\sigma' \in \text{Supp}(\mu^*)$ and every m such that $\sum_\sigma \left(\sigma \circ f_{a^*}^{-1}\right)(m) \times \mu^*(\sigma) > 0$,

$$\frac{\left(\sigma' \circ f_{a^*}^{-1}\right)(m) \times \mu^*(\sigma')}{\sum_\sigma \left(\sigma \circ f_{a^*}^{-1}\right)(m) \times \mu^*(\sigma)} = \mu^*(\sigma').$$

Since $\sigma' \in \text{Supp}(\mu^*)$ if and only if $\mu^*(\sigma') > 0$,

$$\forall \sigma' \in \text{Supp}(\mu^*),\ \left(\sigma' \circ f_{a^*}^{-1}\right)(m) = \sum_{\sigma \in \text{Supp}(\mu^*)} \left(\sigma \circ f_{a^*}^{-1}\right)(m) \times \mu^*(\sigma),$$

for every m, which implies that $\sigma \mapsto \sigma \circ f_{a^*}^{-1}$ is constant on $\text{Supp}(\mu^*)$. ■

Proposition 3 *Let α be an optimal strategy. Suppose that $\mu_0(\sigma^*) > 0$ and that the process $(\mathbf{a}_t^\alpha, \boldsymbol{\mu}_t^\alpha)_{t \in \mathbb{N}_0}$ converges in finite time T to (a^*, μ^*) along each path with prefix $(s_0, ..., s_{T-1})$ such that $\sigma^*(s_t) > 0$ for all $t \in \{0, ..., T-1\}$. Then (a^*, μ^*) is a self-confirming equilibrium at σ^*.*

Proof Let μ_t denote the realization of the updated belief at the beginning of any period t given $(s_0, ...s_{t-1})$. Since $\sigma^*(s_t) > 0$ for all $t \in \{0, ..., T-1\}$ and $\mu_0(\sigma^*) > 0$, the Bayesian updating formula implies that $\mu^*(\sigma^*) = \mu_T(\sigma^*) > 0$, that is, $\sigma^* \in \text{Supp}(\mu^*)$ (this follows from an easy induction argument, considering that

$$\mu_{t+1}(\sigma^*|a_t, m_t) = \frac{\left(\sigma^* \circ f_{a_t}^{-1}\right)(m_t) \times \mu_t(\sigma^*)}{\sum_\sigma \left(\sigma \circ f_{a_t}^{-1}\right)(m_t) \times \mu_t(\sigma)} > 0$$

if $\mu_t(\sigma^*) > 0$). Then, by Lemma 2, $\sigma \circ f_{a^*}^{-1} = \sigma^* \circ f_{a^*}^{-1}$ for all $\sigma \in \text{Supp}(\mu^*)$, that is,

$$\mu^*\left(\left\{\sigma \in \Sigma : \sigma \circ f_{a^*}^{-1} = \sigma^* \circ f_{a^*}^{-1}\right\}\right) = 1.$$

Now we have to show that $a^* \in \arg\max_a u(a, \mu^*)$. First note that the normalized expected value of an optimal strategy at any given period depends only on the belief μ at the beginning of that period. According to the Bellman equation,

$$\begin{aligned}
&\mathbb{E}_{\mu^*}\left(\sum_{t=T}^{\infty} \beta^{t-T} u\left(\mathbf{a}_t^\alpha, \boldsymbol{\mu}_t^\alpha\right)\right) \\
= \; & V(\mu^*) = \max_{a \in A}\left\{u(a, \mu^*) + \beta \sum_{m:\mathbb{P}_{\mu^*,a}(m)>0} V\left(B\left(\mu^*, a, m\right)\right)\mathbb{P}_{\mu^*,a}(m)\right\} \\
= \; & u(a^*, \mu^*) + \beta \sum_{m:\mathbb{P}_{\mu^*,a^*}(m)>0} V\left(B\left(\mu^*, a^*, m\right)\right)\mathbb{P}_{\mu^*,a^*}(m) = u(a^*, \mu^*) + \beta V\left(\mu^*\right),
\end{aligned}$$

where the last equality holds because μ^* is invariant at $a^* = \mathbf{a}_T^\alpha(s_1, ..., s_{T-1})$, the action selected by optimal strategy α at time T given the realizations $(s_0, ..., s_{T-1})$. Thus,

$$\frac{u(a^*, \mu^*)}{1-\beta} = V\left(\mu^*\right).$$

Suppose, by way of contradiction, that $u(a^*, \mu^*) < u(\bar{a}, \mu^*)$ for some $\bar{a} \in A$. The continuation strategy of choosing $\bar{a}$ forever starting from period T would yield the subjective expected value

$$\begin{aligned}
&\sum_{\sigma \in \text{Supp}(\mu^*)} \mu^*(\sigma) \sum_{t=T}^{\infty} \beta^{t-T} u(\bar{a}, \sigma) \\
= \; & \frac{1}{1-\beta} \sum_{\sigma \in \text{Supp}(\mu^*)} \mu^*(\sigma)\, u(\bar{a}, \sigma) \\
= \; & \frac{1}{1-\beta} u(\bar{a}, \mu^*) > \frac{1}{1-\beta} u(a^*, \mu^*) = V\left(\mu^*\right),
\end{aligned}$$

thus contradicting the optimality of α. Therefore, it must be the case that $a^* \in \arg\max_a u(a, \mu^*)$. ■

The previous result provides a simple link between SCE and (Bayesian) learning for a rational DM: if the learning process converges in finite time, then the limit is

an SCE. Adapting results from the literature on repeated decision problems under uncertainty (e.g., Easley and Kiefer, 1988), one can prove that (1) the same is true also when convergence is only "asymptotic," and (2) convergence is almost sure:[7]

Theorem 4 *Suppose that* $\mu_0(\sigma^*) > 0$ *and* $\alpha : \Delta(\Sigma) \to A$ *is the uniquely optimal (hence, stationary) strategy given* β *and* μ_0*. Then the stochastic process* $(\boldsymbol{\mu}_t^\alpha)_{t\in\mathbb{N}_0}$ *converges* $\sigma^{\infty,*}$*-almost surely to a random limit* $\boldsymbol{\mu}_\infty^\alpha$*, that is,*

$$\mathbb{P}_{\sigma^{\infty,*}}\left(\lim_{t\to\infty}\boldsymbol{\mu}_t^\alpha = \boldsymbol{\mu}_\infty^\alpha\right) = 1$$

for some random belief $\boldsymbol{\mu}_\infty^\alpha : S^{\mathbb{N}_0} \to \Delta(\Sigma)$*. Suppose, furthermore, that the static best reply to the limit belief is* $\sigma^{\infty,*}$*-almost surely unique i.e.,*

$$\mathbb{P}_{\sigma^{\infty,*}}\left(\left|\arg\max_{a\in A} u\left(a, \boldsymbol{\mu}_\infty^\alpha\right)\right| = 1\right) = 1,$$

then the stochastic process $(\alpha(\boldsymbol{\mu}_t^\alpha), \boldsymbol{\mu}_t^\alpha)_{t\in\mathbb{N}}$ *converges* $\sigma^{\infty,*}$*-almost surely to a self-confirming equilibrium at* σ^* *(i.e.,* $\mathbb{P}_{\sigma^{\infty,*}}(\lim_{t\to\infty}(\alpha(\boldsymbol{\mu}_t^\alpha), \boldsymbol{\mu}_t^\alpha) \in SCE(\sigma^*)) = 1$*).*

Theorem 4 is proved in the Appendix. Here we sketch the main steps of the proof. First, we show that the stochastic process $(\boldsymbol{\mu}_t^\alpha)_{t\in\mathbb{N}_0}$ is a (necessarily bounded) martingale, that is, the expected value of the next-period subjective probability of any stochastic model σ is the current subjective probability of σ. Therefore, the Martingale Convergence Theorem implies the existence of a random limit $\boldsymbol{\mu}_\infty^\alpha : S^{\mathbb{N}_0} \to \Delta(\Sigma)$ such that $\mathbb{P}_{\sigma^{\infty,*}}$-almost surely $\boldsymbol{\mu}_t^\alpha \to \boldsymbol{\mu}_\infty^\alpha$. Next, using Lemma 2 and the continuity of the Bayes map, we show that, for every path s^∞ such that $\boldsymbol{\mu}_t^\alpha(s^\infty) \to \boldsymbol{\mu}_\infty^\alpha(s^\infty)$ and for every action a_∞ played infinitely often on this path, it must be the case that $\boldsymbol{\mu}_\infty^\alpha$ is a_∞-invariant. Finally, we show that, since the actions played infinitely often tend to have no experimentation value (because they are invariant at $\boldsymbol{\mu}_\infty^\alpha$) they can only be static best replies to this limit belief. Since by hypothesis the static best reply is unique, the thesis follows.

3 Example

Consider a DM that can choose between the action Up (U) and Down (D). Action U is *safe*, i.e. its consequence does not depend on the realized state of nature, whereas action D is *risky*, because it has state-contingent payoffs. Ex-post, the DM only observes his monetary consequences. The static decision problem with feedback can be represented as follows:

g, f	l	r
U	2	2
D	0	3

[7] In games with strategic opponents part (1) holds, but part (2) may fail.

Moreover, the DM is uncertain about the stochastic model. Specifically, he has a 2-point prior μ with $\Sigma = \text{Supp}\mu = \{\sigma^{uni}, \sigma^r\}$, where σ^{uni} is the uniform model, whereas σ^r assigns a higher probability $(\frac{5}{6})$ to r. Formally:

Model	l	r
σ^{uni}	$1/2$	$1/2$
σ^r	$1/6$	$5/6$

Prior	σ^{uni}	σ^r
μ	μ^{uni}	$1-\mu^{uni}$

As a consequence, not only action D is *objectively risky*, but it is also *subjectively* ambiguous, that is, it entails risks that are unknown to the DM. Finally, suppose we assume risk neutrality, that is $v(m) = m$.

Recall that, without loss of generality, the optimal strategy can be chosen to be stationary. Therefore, we consider a strategy $\alpha^* \in A^{\Delta(\Sigma)}$ that solves

$$\max_{\alpha \in A^{\Delta(\Sigma)}} \mathbb{E}_\mu \left(\sum_{t=0}^{\infty} \beta^t u\left(\alpha\left(\boldsymbol{\mu}_t^\alpha\right), \boldsymbol{\mu}_t^\alpha\right) \right) = V(\mu) = \max_{a \in A} \left\{ u(a,\mu) + \beta \sum_{m:\mathbb{P}_{\mu,a}(m)>0} V\left(B(\mu,a,m)\right) \mathbb{P}_{\mu,a}(m) \right\}, \tag{1}$$

where the second equality follows from Bellman's equation. The last expression sheds light on the particular form assumed by the optimal strategy in this example. Indeed, note that, for every μ, action U does not allow any learning: given U, the DM receive with probability 1 a payoff (and message) equal to 2 and the posterior coincides with the prior. Formally:

$$\forall \mu \in \Sigma, \mathbb{P}_{\mu,U}(m : B(\mu,U,m) = \mu) = \mathbb{P}_{\mu,U}(\mathbf{m} = 2) = 1.$$

Therefore, if for some t the belief μ_t is such that the optimal strategy α^* prescribes action U, then he will stick to this action forever on, and the value will coincide with the discounted sum of the constant payoff 2:

$$\begin{array}{ll} \alpha(\mu) = U \Rightarrow & V(\mu) = u(U,\mu) + \beta \sum_{m:\mathbb{P}_{\mu,U}(m)>0} V\left(B(\mu,U,m)\right) \mathbb{P}_{\mu,U}(m) = u(U,\mu) + \beta V(\mu) \\ & V(\mu) = \frac{2}{1-\beta}. \end{array} \tag{2}$$

In other words, U has no experimentation value. On the other hand, playing action D has a positive experimentation value, since the different probabilities assigned to message 3 by models σ^{uni} and σ^r imply that the DM will end up with a posterior different from his prior.[8] As a consequence, we may imagine that the more the DM is patient (i.e., the more he is willing to trade-off current consumption for information that is valuable for future choices) the more action D becomes attractive. Indeed, this intuition is correct.

[8] Except, of course, for the degenerate cases $\mu = \delta_x$, $x \in \{\sigma^{uni}, \sigma^r\}$.

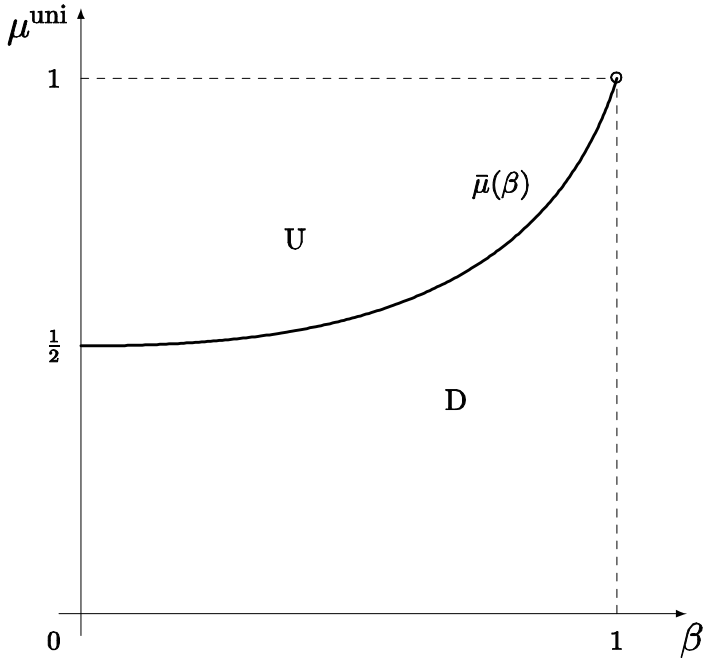

Figure1. Graph of the optimal policy.

Figure 1 depicts the optimal strategy as a function of the discount factor β and the belief in the uniform model μ^{uni} (recall that the optimal action for the uniform model is U). For each β, there is a threshold $\bar{\mu}(\beta)$ such that it is optimal to experiment (D) if $\mu^{uni} < \bar{\mu}(\beta)$ and it is optimal to stop experimentation (U) if $\mu^{uni} > \bar{\mu}(\beta)$. Thus, U is optimal in the upper region. The threshold function $\bar{\mu}(\beta)$ is increasing, because a more patient DM needs a higher belief in the uniform model σ^{uni} to stop experimenting with D. If $\mu_0^{uni} > \bar{\mu}(\beta)$, then the optimal strategy plays U forever. If $\mu_0^{uni} < \bar{\mu}(\beta)$, the optimal strategy starts with D. If the first message is $m_0 = 3$, then $\mu_1(\sigma^{uni}|D, 3) < \mu_0^{uni} < \bar{\mu}(\beta)$ and the optimal strategy keeps experimenting with D in period $t = 1$. If instead the first message is $m_0 = 0$, then $\mu_1(\sigma^{uni}|D, 3) > \mu_0^{uni}$. If the belief in the uniform model increases above $\bar{\mu}(\beta)$, then experimentation stops, otherwise it continues.

The objective probability of messages 0 and 3 given D depends on the true model σ. If the true model is σ^{uni} experimentation stops in finite time with probability 1 and the DM ends up using the objectively optimal action U even if the long-run belief does not assign probability 1 to σ^{uni} because μ_t becomes constant as soon as $\alpha(\boldsymbol{\mu}_t^{\alpha}) = U$. To see that in this case experimentation stops with probability 1 in finite time, suppose—by way of contradiction, that $\mathbb{P}_{\sigma^{uni}}(\forall t \in \mathbb{N}_0, \alpha(\boldsymbol{\mu}_t^{\alpha}) = D) > 0$, then the probability that the updated belief converges to the Dirac on σ^{uni} is also positive, that is, $\mathbb{P}_{\sigma^{uni}}(\lim_{t\to\infty} \boldsymbol{\mu}_t^{\alpha}(\sigma^{uni}) = 1) > 0$. Then, $\mathbb{P}_{\sigma^{uni}}(\exists t, \boldsymbol{\mu}_t^{\alpha}(\sigma^{uni}) > \bar{\mu}(\beta)) > 0$, which implies that experimentation must stop in finite time with positive probability, a contradiction.

If instead the true model is σ^r and $\mu_0^{uni} < \bar{\mu}(\beta)$, the optimal strategy starts with

D and there are two possible limits: either (1) there are sufficiently many early realizations of message 3, experimentation (D) continues forever and $\lim_{t\to\infty} \boldsymbol{\mu}_t^{\alpha}(\sigma^r) = \lim_{t\to\infty}(1-\boldsymbol{\mu}_t^{\alpha}(\sigma^{uni})) = 1$ because the DM learns the true probabilities observing the empirical frequencies, or (2) there is a sufficiently high number of early unlucky realization of message 0 and eventually $\boldsymbol{\mu}_t^{\alpha}(\sigma^{uni}) > \bar{\mu}(\beta)$, so that the optimal strategy switches to the *objectively suboptimal* action U forever and $\boldsymbol{\mu}_t^{\alpha}$ becomes constant with $\boldsymbol{\mu}_t^{\alpha}(\sigma^{uni}) > \bar{\mu}(\beta)$.

4 Appendix

Proof of Theorem 4 For technical reasons, to show almost sure convergence of the random vector of beliefs in the probability space $(S^{\mathbb{N}_0}, \mathcal{G}, \sigma^*)$, where $\mathcal{G}$ is the σ-algebra generated by the elementary cylinders, we have to rely on a derived probability space. Consider the probability space $(S^{\mathbb{N}_0}, \mathcal{G}, p_{\mu_0})$ where p_{μ_0} is the (predictive) measure obtained from the prior μ_0. More specifically, for every $E \in \mathcal{G}$

$$p_{\mu_0}(E) = \sum_{\sigma\in\Sigma} \sigma^{\infty}(E)\,\mu_0(\sigma).$$

(Step 1: the belief process is a martingale) We show that, for every $\bar{\sigma} \in \Sigma$, $(\boldsymbol{\mu}_t^{\alpha}(\bar{\sigma}), \mathcal{F}_t)_{t\in\mathbb{N}_0}$ is a martingale in this probability space, where $\mathcal{F}_t$ is the sigma-algebra generated by the random variables $(\boldsymbol{\mu}_0^{\alpha}(\bar{\sigma}), ..., \boldsymbol{\mu}_t^{\alpha}(\bar{\sigma}))$.

(Proof of Step 1) By definition, $(\mathcal{F}_t)_{t\in\mathbb{N}_0}$ is a filtration and $(\boldsymbol{\mu}_t^{\alpha}(\bar{\sigma}))_{t\in\mathbb{N}_0}$ is adapted to $(\mathcal{F}_t)_{t\in\mathbb{N}_0}$. Moreover, for all $t \in \mathbb{N}_0$, we have $0 \leq \mathbb{E}_{p_{\mu_0}}(\boldsymbol{\mu}_t^{\alpha}(\bar{\sigma})) \leq 1$. Therefore, it only remains to show that for every $t \in \mathbb{N}_0$, we want to show that with probability $\mathbb{P}_{p_{\mu_0}}$ almost surely

$$\mathbb{E}_{p_{\mu_0}}\left(\boldsymbol{\mu}_{t+1}(\bar{\sigma})\,|\mathcal{F}_t\right) = \boldsymbol{\mu}_t(\bar{\sigma}).$$

Given the finiteness of S, for every $t \in \mathbb{N}_0$, there is a finite number of values that the random belief $\boldsymbol{\mu}_t$ takes with positive probability. So, it is enough to show that contingent of the realization of any such μ_t, the expected value of $\boldsymbol{\mu}_{t+1}$ is μ_t. Fix any μ_t in the support of the random belief $\boldsymbol{\mu}_t$: $\mathbb{P}_{p_{\mu_0}}(\boldsymbol{\mu}_t = \mu_t) > 0$. Recall that the Bayes map yields

$$(\mu_t, \alpha(\mu_t), m) \mapsto B(\mu_t, \alpha(\mu_t), m)(\bar{\sigma}) = \frac{\mu_t(\bar{\sigma})\left(\left(\bar{\sigma}\circ f^{-1}_{\alpha(\mu_t)}\right)(m)\right)}{\sum_{\sigma}\left(\left(\sigma\circ f^{-1}_{\alpha(\mu_t)}\right)(m)\right)\mu_t(\sigma)}$$

for each m deemed possible according to μ_t given action $\alpha(\mu_t)$, that is, each m such that the denominator is positive. With this,

$$\mathbb{E}_{p_{\mu_0}}\left(\boldsymbol{\mu}_{t+1}(\bar{\sigma})\,|\mu_t\right) = \sum_{m}\mathbb{P}_{p_{\mu_0}}(\mathbf{m}_t = m|\mu_t)\,B(\mu_t, \alpha(\mu_t), m)(\bar{\sigma})$$

$$= \sum_m \sum_\sigma \frac{\mu_0(\sigma)\,\sigma^\infty\left(\left\{s^\infty : \boldsymbol{\mu}_t(s^{t-1}) = \mu_t, s_t \in f^{-1}_{\alpha(\mu_t)}(m)\right\}\right)}{\sum_{\hat\sigma\in\Sigma}\mu_0(\hat\sigma)\,\hat\sigma^\infty\left(\{s^\infty : \boldsymbol{\mu}_t(s^{t-1}) = \mu_t\}\right)} B(\mu_t, \alpha(\mu_t), m)(\bar\sigma)$$

$$= \sum_m \sum_\sigma \sigma\left(f^{-1}_{\alpha(\mu_t)}(m)\right) \frac{\mu_0(\sigma)\,\sigma^\infty\left(\{s^\infty : \boldsymbol{\mu}_t(s^{t-1}) = \mu_t\}\right)}{\sum_{\hat\sigma\in\Sigma}\mu_0(\hat\sigma)\,\hat\sigma^\infty\left(\{s^\infty : \boldsymbol{\mu}_t(s^{t-1}) = \mu_t\}\right)} B(\mu_t, \alpha(\mu_t), m)(\bar\sigma)$$

$$= \sum_m \sum_\sigma \sigma\left(f^{-1}_{\alpha(\mu_t)}(m)\right) \mu_t(\sigma) \frac{\mu_t(\bar\sigma)\left(\left(\bar\sigma \circ f^{-1}_{\alpha(\mu_t)}\right)(m)\right)}{\sum_{\sigma'}\left(\left(\sigma' \circ f^{-1}_{\alpha(\mu_t)}\right)(m)\right)\mu_t(\sigma')}$$

$$= \sum_m \mu_t(\bar\sigma)\left(\left(\bar\sigma \circ f^{-1}_{\alpha(\mu_t)}\right)(m)\right) \sum_\sigma \frac{\sigma\left(f^{-1}_{\alpha(\mu_t)}(m)\right)\mu_t(\sigma)}{\sum_{\sigma'}\left(\left(\sigma' \circ f^{-1}_{\alpha(\mu_t)}\right)(m)\right)\mu_t(\sigma')}$$

$$= \sum_m \mu_t(\bar\sigma)\left(\left(\bar\sigma \circ f^{-1}_{\alpha(\mu_t)}\right)(m)\right) = \mu_t(\bar\sigma) \sum_m \left(\left(\bar\sigma \circ f^{-1}_{\alpha(\mu_t)}\right)(m)\right) = \mu_t(\bar\sigma).$$

where the first equality comes from the definition of expected value, the second by the definition of p_{μ_0}, the third from the fact that the environment is i.i.d., the fourth from the definition of conditional probability, whereas the remaining equalities are immediate.

(Step 2: the belief process converges) The stochastic process $(\boldsymbol{\mu}^\alpha_t)_{t\in\mathbb{N}_0}$ converges $\sigma^{\infty,*}$-almost surely to a random limit $\boldsymbol{\mu}^\alpha_\infty$.

(Proof of Step 2) By step 1, the sequence of random subjective probability of models $(\boldsymbol{\mu}^\alpha_t(\cdot))_{t\in\mathbb{N}_0}$ is a martingale. Moreover, for every $t \in \mathbb{N}_0$, $0 \leq \boldsymbol{\mu}^\alpha_t(\bar\sigma) \leq 1$, and therefore, $(\boldsymbol{\mu}^\alpha_t(\bar\sigma))_{t\in\mathbb{N}_0}$ is a uniformly bounded martingale. By the Martingale Convergence Theorem (see, e.g. Billingsley Theorem 35.5), the limit random variable $\boldsymbol{\mu}^\alpha_\infty(\bar\sigma)$ exists p_{μ_0}-almost surely. Since the result holds for every $\bar\sigma \in \Sigma$, we have that $(\boldsymbol{\mu}^\alpha_t)_{t\in\mathbb{N}_0}$ converges p_{μ_0}-almost surely to a random limit $\boldsymbol{\mu}^\alpha_\infty$. Indeed, since Σ is finite, $(\boldsymbol{\mu}^\alpha_t)_{t\in\mathbb{N}_0}$ is a $|\Sigma|$-dimensional vector-valued stochastic process. Therefore, the convergence of $(\boldsymbol{\mu}^\alpha_t(\bar\sigma))_{t\in\mathbb{N}_0}$ for every $\bar\sigma \in \Sigma$ implies the convergence of the vector. Moreover, since for every $t \in \mathbb{N}_0$ $\boldsymbol{\mu}^\alpha_t$ belongs to the $|\Sigma|$-dimensional simplex, a compact set, also the limit belongs to the simplex.

Therefore, there exists a set $\hat E \in \mathcal{G}$ such that $p_{\mu_0}\left(\hat E\right) = 1$ and $\lim_{t\to\infty}(\boldsymbol{\mu}^\alpha_t(s^\infty))_{t\in\mathbb{N}_0} = \boldsymbol{\mu}^\alpha_\infty(s^\infty)$ for every $s^\infty \in \hat E$. Note that, since $\mu_0(\sigma^*) > 0$

$$\begin{aligned} p_{\mu_0}\left(\hat E\right) &= 1 \Rightarrow p_{\mu_0}\left(S^{\mathbb{N}_0}\backslash\hat E\right) = 0 \\ &\Rightarrow \sum_{\sigma\in\Sigma}\sigma^\infty\left(S^{\mathbb{N}_0}\backslash\hat E\right)\mu_0(\sigma) = 0 \\ &\Rightarrow \sigma^{*,\infty}\left(S^{\mathbb{N}_0}\backslash\hat E\right) = 0 \Rightarrow \sigma^{*,\infty}\left(\hat E\right) = 1. \end{aligned}$$

Therefore, $(\boldsymbol{\mu}^\alpha_t)_{t\in\mathbb{N}_0}$ converges $\sigma^{\infty,*}$-almost surely to a random limit $\boldsymbol{\mu}^\alpha_\infty$.

(Step 3: invariance of limit belief) For every s^∞ define as $a^\infty(s^\infty)$ the set of actions played infinitely often (i.o.) along this path. We have that

$$\forall a \in a^\infty(s^\infty),\ \boldsymbol{\mu}_\infty^\alpha\left(\left\{\sigma \in \Sigma : \sigma \circ f_a^{-1} = \sigma^* \circ f_a^{-1}\right\}\right)(s^\infty) = 1$$

$\sigma^{\infty,*}$-almost surely.

(Proof of step 3) Consider the set $\bar{E}$ obtained as the intersection between $\hat{E}$ and:

$$\left\{s^\infty : \forall (a,\bar{s}) \in a^\infty(s^\infty) \times \text{Supp}\sigma^*, \left(\alpha\left(\boldsymbol{\mu}_{t-1}^\alpha\right), \mathbf{s}_t\right)(s^\infty) = (a,\bar{s}) \text{ i.o.}\right\}.$$

It can be shown that the latter set has probability 1.[9]

Consider any sample path $s^\infty \in \bar{E}$. Suppose by way of contradiction that $\boldsymbol{\mu}_\infty^\alpha(s^\infty)$ is not invariant for some $a \in a^\infty(s^\infty)$, that is, there exists $m \in M$ such that $(\sigma \circ f_a^{-1})(m) \neq (\sigma^* \circ f_a^{-1})(m)$ for some $\sigma \in \text{Supp}\boldsymbol{\mu}_\infty^\alpha(s^\infty)$. It follows that there exists $\bar{s} \in \text{Supp}\sigma^*$ such that $(\sigma \circ f_a^{-1})(f(a,\bar{s})) \neq (\sigma^* \circ f_a^{-1})(f(a,\bar{s}))$ for some $\sigma \in \text{Supp}\boldsymbol{\mu}_\infty^\alpha(s^\infty)$. Indeed,

$$\left(\sigma \circ f_a^{-1}\right)(m) \neq \left(\sigma^* \circ f_a^{-1}\right)(m)$$

for some $m \in M$, and since $\sigma \circ f_a^{-1}$ and $\sigma^* \circ f_a^{-1}$ are elements of the $|M|$-dimensional simplex, this implies that there are $m' \in M$ and $m'' \in \text{Supp}(\sigma^* \circ f_a^{-1})$ such that

$$\left(\sigma \circ f_a^{-1}\right)(m') > \left(\sigma^* \circ f_a^{-1}\right)(m') \text{ and } \left(\sigma \circ f_a^{-1}\right)(m'') < \left(\sigma^* \circ f_a^{-1}\right)(m'').$$

Therefore, Lemma 2 implies $B(\boldsymbol{\mu}_\infty^\alpha(s^\infty), a, f(a,\bar{s})) \neq \boldsymbol{\mu}_\infty^\alpha(s^\infty)$, and the continuity of the Bayes map implies that there exist $\varepsilon > 0$ and $\delta > 0$ such that:

$$||\boldsymbol{\mu}_{t-1}^\alpha(s^\infty) - \boldsymbol{\mu}_\infty^\alpha(s^\infty)|| \leq \delta \Rightarrow ||B\left(\boldsymbol{\mu}_{t-1}^\alpha(s^\infty), a, f(a,\bar{s})\right) - \boldsymbol{\mu}_\infty^\alpha(s^\infty)|| \geq \varepsilon.$$

Indeed, suppose $||B(\boldsymbol{\mu}_\infty^\alpha(s^\infty), a, \bar{s}) - \boldsymbol{\mu}_\infty^\alpha(s^\infty)|| = 2\varepsilon > 0$. There exists $\delta > 0$ such that

$$||\boldsymbol{\mu}_{t-1}^\alpha(s^\infty) - \boldsymbol{\mu}_\infty^\alpha(s^\infty)|| \leq \delta \Rightarrow ||B\left(\boldsymbol{\mu}_{t-1}^\alpha(s^\infty), a, f(a,\bar{s})\right) - B(\boldsymbol{\mu}_\infty^\alpha(s^\infty), a, f(a,\bar{s}))|| \leq \varepsilon$$

Therefore,

$$\begin{aligned} 2\varepsilon &\leq ||B(\boldsymbol{\mu}_\infty^\alpha(s^\infty), a, \bar{s}) - \boldsymbol{\mu}_\infty^\alpha(s^\infty)|| \\ &\leq ||B\left(\boldsymbol{\mu}_{t-1}^\alpha(s^\infty), a, f(a,\bar{s})\right) - \boldsymbol{\mu}_\infty^\alpha(s^\infty)|| \\ &\quad + ||B\left(\boldsymbol{\mu}_{t-1}^\alpha(s^\infty), a, f(a,\bar{s})\right) - B(\boldsymbol{\mu}_\infty^\alpha(s^\infty), a, f(a,\bar{s}))|| \end{aligned}$$

If we subtract ε from the LHS and $||B\left(\boldsymbol{\mu}_{t-1}^\alpha(s^\infty), a, f(a,\bar{s})\right) - B(\boldsymbol{\mu}_\infty^\alpha(s^\infty), a, f(a,\bar{s}))||$ from the RHS the inequality continues to hold and

$$\varepsilon \leq ||B\left(\boldsymbol{\mu}_{t-1}^\alpha(s^\infty), a, f(a,\bar{s})\right) - \boldsymbol{\mu}_\infty^\alpha(s^\infty)||.$$

[9] See the Proof of Claim 1 below.

Since $s^\infty \in \bar{E}$, there exists a subsequence of periods t_n such that, for every $n \in \mathbb{N}$, $\left(\alpha\left(\boldsymbol{\mu}^\alpha_{t_n-1}\right), \mathbf{s}_{t_n}\right)(s^\infty) = (a, \bar{s})$ and $||\boldsymbol{\mu}^\alpha_{t_n-1}(s^\infty) - \boldsymbol{\mu}^\alpha_\infty(s^\infty)|| \leq \delta$. Therefore, for every t_n

$$||\boldsymbol{\mu}^\alpha_{t_n}(s^\infty) - \boldsymbol{\mu}^\alpha_\infty(s^\infty)|| \geq \varepsilon,$$

but this contradicts the convergence of the beliefs shown in Step 2.

(Step 4: the limit action is a short-run best reply to the limit belief) Consider any path s^∞ such that $|\arg\max_{a\in A} u(a, \boldsymbol{\mu}^\alpha_\infty(s^\infty))| = 1$ and any $a^* \in a^\infty(s^\infty)$, then

$$a^* = \arg\max_{a\in A} u(a, \boldsymbol{\mu}^\alpha_\infty(s^\infty))$$

(Proof of Step 4) Fix s^∞ as above and denote the myopic best reply as:

$$a_1 = \arg\max_{a\in A} u(a, \boldsymbol{\mu}^\alpha_\infty(s^\infty)).$$

It is immediate to see that there exist an open ball of radius ε centered in $\boldsymbol{\mu}^\alpha_\infty(s^\infty)$, $N(\boldsymbol{\mu}^\alpha_\infty(s^\infty), \varepsilon)$ such that[10]

$$\forall \mu \in N(\boldsymbol{\mu}^\alpha_\infty(s^\infty), \varepsilon), \forall a' \in A \backslash \{a_1\}, \; u(a_1, \mu) - u(a', \mu) > 0.$$

Fix $\bar{a} \in a^\infty(s^\infty) \backslash \{a_1\}$, we show that there exists a δ such that

$$|\boldsymbol{\mu}^\alpha_\infty(s^\infty) - \mu_t| < \delta \Rightarrow \alpha(\mu_t) \neq \bar{a}.$$

Since the space of actions and states is finite, $|u(a, s) - u(a', s')| \leq K$ for some K in $\mathbb{R}$. Since $\beta < 1$ there exist $\tau \in \mathbb{N}$ such that

$$\frac{u(a_1, \boldsymbol{\mu}^\alpha_\infty(s^\infty)) - u(\bar{a}, \boldsymbol{\mu}^\alpha_\infty(s^\infty))}{2} > \beta^\tau \frac{K}{1-\beta}.$$

Moreover, by Step 3 and Lemma 2,

$$B(\boldsymbol{\mu}^\alpha_\infty(s^\infty), \bar{a}, \cdot) = \boldsymbol{\mu}^\alpha_\infty(s^\infty).$$

Therefore, for every $\varepsilon > 0$ there exists a $t_\varepsilon \in \mathbb{N}$, such that if $t \geq t_\varepsilon$, then $\mu_{t+1} \in N(\boldsymbol{\mu}^\alpha_\infty(s^\infty), \varepsilon)$ μ_t-almost surely.

Indeed, let δ_ε be such that,[11] for every $a \in a^\infty(s^\infty)$, $m \in M$

$$||\mu_t - \boldsymbol{\mu}^\alpha_\infty(s^\infty)|| < \delta_\varepsilon \Rightarrow ||B(\mu_t, a, m) - \delta_{\boldsymbol{\mu}^\alpha_\infty(s^\infty)}|| < \varepsilon.$$

Let t_ε be such that from that period onwards updated beliefs are in $N(\boldsymbol{\mu}^\alpha_\infty(s^\infty), \delta_\varepsilon)$ and only actions in $a^\infty(s^\infty)$ are played. Therefore, since

$$||B(\mu_t, a, m) - \delta_{\boldsymbol{\mu}^\alpha_\infty(s^\infty)}|| = \max_{\mu_{t+1} \in \mathrm{Supp} B(\mu_t, a, m)} ||\mu_{t+1} - \boldsymbol{\mu}^\alpha_\infty(s^\infty)||$$

[10] Notice that u is continuous in its second argument.

[11] The existence of this δ_e is guaranteed by the continuity of the Bayes map and the invariance of limit beliefs with respect to limit actions.

we have that $\mu_{t+1} \in N\left(\boldsymbol{\mu}_{\infty}^{\alpha}\left(s^{\infty}\right), \varepsilon\right)$ μ_t-almost surely. A similar argument shows that for every $\varepsilon > 0$ and $j \in \mathbb{N}$, there exists a $t_{\varepsilon,j} \in \mathbb{N}$, such that if $t \geq t_{\varepsilon,j}$, then for every $i \in \{1, ..., j\}$, $\mu_{t+i} \in N\left(\boldsymbol{\mu}_{\infty}^{\alpha}\left(s^{\infty}\right), \varepsilon\right)$ μ_t-almost surely.

It follows that $\bar{a}$ cannot be played from period $t_{\varepsilon,\tau}$ onward, otherwise the strategy α_1 that prescribes to play always a_1 from then on would be strictly preferred to α, a contradiction. Indeed, fix a $t^* \geq t_{\varepsilon,\tau}$ and let

$$\widetilde{V}\left(\alpha_1, \mu_{t^*}\right) = \mathbb{E}_{\mu_{t^*}}\left(\sum_{t=t^*}^{\infty} \beta^{t-t^*} u\left(a_1, \boldsymbol{\mu}_t^{\alpha_1}\right)\right)$$

denote the subjective value at belief μ_{t^*} of the strategy that always plays a_1. We have

$$\begin{aligned}
& \widetilde{V}\left(\alpha_1, \mu_{t^*}\right) - V\left(\mu_{t^*}\right) \\
= \; & u\left(a_1, \mu_{t^*}\right) - u\left(\bar{a}, \mu_{t^*}\right) + \mathbb{E}_{\mu_t}\left(\sum_{t=t^*+1}^{\infty} \beta^{t-t^*}\left(u\left(a_1, \boldsymbol{\mu}_t^{\alpha_1}\right) - u\left(\mathbf{a}_t^{\alpha}, \boldsymbol{\mu}_t^{\alpha}\right)\right)\right) \\
= \; & u\left(a_1, \mu_{t^*}\right) - u\left(\bar{a}, \mu_{t^*}\right) + \mathbb{E}_{\mu_t}\left(\sum_{t=t^*+1}^{t^*+\tau} \beta^{t-t^*}\left(u\left(a_1, \boldsymbol{\mu}_t^{\alpha_1}\right) - u\left(\mathbf{a}_t^{\alpha}, \boldsymbol{\mu}_t^{\alpha}\right)\right)\right) \\
& + \mathbb{E}_{\mu_t}\left(\sum_{t=t^*+\tau+1}^{\infty} \beta^{t-t^*}\left(u\left(a_1, \boldsymbol{\mu}_t^{\alpha_1}\right) - u\left(\mathbf{a}_t^{\alpha}, \boldsymbol{\mu}_t^{\alpha}\right)\right)\right) \\
\geq \; & \frac{u\left(a_1, \boldsymbol{\mu}_{\infty}^{\alpha}\left(s^{\infty}\right)\right) - u\left(\bar{a}, \boldsymbol{\mu}_{\infty}^{\alpha}\left(s^{\infty}\right)\right)}{2} - \beta^{\tau} \frac{K}{1-\beta} \\
> \; & 0.
\end{aligned}$$

But this contradicts the optimality of strategy α. Therefore $a^{\infty}\left(s^{\infty}\right) = \{a_1\}$.

Summing up, there exists E, such that $\sigma^{\infty,*}(E) = 1$, and on E we have convergence of beliefs (Step 2), a unique action played after a finite time, the myopic best reply to limit beliefs (Step 4), and this action confirms limit beliefs (Step 3). That is, we have convergence to a self-confirming equilibrium $\sigma^{\infty,*}$-almost surely. ■

Proof of Claim 1 Fix $(a, \bar{s})$ and $n \in \mathbb{N}$. Let $E\left(a, \bar{s}, n\right) \subseteq S^{\mathbb{N}_0}$ be the set formed by the s^{∞} such that $a \in a^{\infty}\left(s^{\infty}\right)$ and such that for every $t \geq n$, $\left(\alpha\left(\boldsymbol{\mu}_{t-1}^{\alpha}\right), \mathbf{s}_t\right)\left(s^{\infty}\right) \neq (a, \bar{s})$.

Clearly,

$$\begin{aligned}
& S^{\mathbb{N}_0} \backslash \left\{s^{\infty} : \forall\, (a, \bar{s}) \in a^{\infty}\left(s^{\infty}\right) \times \mathrm{Supp}\sigma^*, \left(\alpha\left(\boldsymbol{\mu}_{t-1}^{\alpha}\right), \mathbf{s}_t\right)\left(s^{\infty}\right) = (a, \bar{s}) \text{ i.o.}\right\} \\
\subseteq \; & \bigcup_{a \in A} \bigcup_{\bar{s} \in \mathrm{Supp}\sigma^*} \bigcup_{n \in \mathbb{N}} E\left(a, \bar{s}, n\right).
\end{aligned}$$

For $(a, \bar{s})$ and n

$$E\left(a, \bar{s}, n\right) \subseteq \left\{s^{\infty} : \exists t_1 \geq n, \left(\alpha\left(\boldsymbol{\mu}_{t_1}^{\alpha}\right), \mathbf{s}_{t_1+1}\right)\left(s^{\infty}\right) = \left(a, s'\right),\ s' \neq \bar{s}\right\}.$$

Therefore:

$$\begin{aligned}\mathbb{P}_{\sigma^{\infty},*}\left(E\left(a,\bar{s},n\right)\right) &\leq \mathbb{P}_{\sigma^{\infty},*}\left(\left\{s^{\infty}:\exists t_1\geq n,\left(\alpha\left(\boldsymbol{\mu}_{t_1}^{\alpha}\right),\mathbf{s}_{t_1+1}\right)\left(s^{\infty}\right)=\left(a,s'\right),\ s'\neq\bar{s}\right\}\right)\\ &\leq \sigma^{*}\left(S\backslash\left\{\bar{s}\right\}\right)<1.\end{aligned}$$

Similarly

$$E\left(a,\bar{s},n\right)\subseteq\left\{\begin{array}{c}s^{\infty}:\exists t_1,t_2\geq n,\left(\alpha\left(\boldsymbol{\mu}_{t_1}^{\alpha}\right),\mathbf{s}_{t_1+1}\right)\left(s^{\infty}\right)=\left(a,s'\right),\ s'\neq\bar{s},\\ \left(\alpha\left(\boldsymbol{\mu}_{t_2}^{\alpha}\right),\mathbf{s}_{t_2+1}\right)\left(s^{\infty}\right)=\left(a,s''\right),\ s''\neq\bar{s}\end{array}\right\}$$

Therefore:

$$\begin{aligned}&\mathbb{P}_{\sigma^{\infty},*}\left(E\left(a,\bar{s},n\right)\right)\\ \leq\ &\mathbb{P}_{\sigma^{\infty},*}\left(\left\{\begin{array}{c}s^{\infty}:\exists t_1,t_2\geq n,\left(\alpha\left(\boldsymbol{\mu}_{t_1}^{\alpha}\right),\mathbf{s}_{t_1+1}\right)\left(s^{\infty}\right)=\left(a,s'\right),\ s'\neq\bar{s},\\ \left(\alpha\left(\boldsymbol{\mu}_{t_2}^{\alpha}\right),\mathbf{s}_{t_2+1}\right)\left(s^{\infty}\right)=\left(a,s''\right),\ s''\neq\bar{s}\end{array}\right\}\right)\\ \leq\ &\left(\sigma^{*}\left(S\backslash\left\{\bar{s}\right\}\right)\right)^{2}.\end{aligned}$$

Proceeding in this way, we can show that $\mathbb{P}_{\sigma^{\infty},*}\left(E\left(a,\bar{s},n\right)\right)=0$ for every $E_{a,\bar{s},n}$, and then

$$\mathbb{P}_{\sigma^{\infty},*}\left(S^{\mathbb{N}_0}\backslash\left\{s^{\infty}:\forall\left(a,s\right)\in a^{\infty}\left(s^{\infty}\right)\times\mathrm{Supp}\sigma^{*},\left(\alpha\left(\boldsymbol{\mu}_{t-1}^{\alpha}\right),\mathbf{s}_{t}\right)\left(s^{\infty}\right)=\left(a,s\right)\ \text{i.o.}\right\}\right)=0.$$

■

References

[1] Battigalli, P., E. Catonini, and N. De Vito (2024): *Game Theory. Analysis of Strategic Thinking.* Unpublished textbook.

[2] Battigalli, P., S. Cerreia, F. Maccheroni, and M. Marinacci (2015): "Self-Confirming Equilibrium and Model Uncertainty," *American Economic Review*, 105, 646-677.

[3] Battigalli, P., A. Francetich, G. Lanzani, and M. Marinacci, (2019): "Learning and self-confirming long-run biases," *Journal of Economic Theory,* 183, 740-785.

[4] Billingsley, P. (2012): *Probability and Measure,* John Wiley & Sons.

[5] Easley, D. and N.M. Kiefer (1988): "Controlling a Stochastic Process with Unknown Parameters," *Econometrica*, 5, 1045-1064.