Games and Economic Behavior 120 (2020) 96–120

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Payoff information and learning in signaling games [☆]

Drew Fudenberg [a,*], Kevin He [b,c]

[a] Department of Economics, MIT, United States
[b] California Institute of Technology, United States
[c] University of Pennsylvania, United States

ARTICLE INFO

Article history:
Received 11 July 2019
Available online 30 December 2019

JEL classification:
C72
C73
D83

Keywords:
Learning
Equilibrium refinements
Bandit problems
Payoff information
Signaling games

ABSTRACT

We add the assumption that players know their opponents' payoff functions and rationality to a model of non-equilibrium learning in signaling games. Agents are born into player roles and play against random opponents every period. Inexperienced agents are uncertain about the prevailing distribution of opponents' play, but believe that opponents never choose conditionally dominated strategies. Agents engage in active learning and update beliefs based on personal observations. Payoff information can refine or expand learning predictions, since patient young senders' experimentation incentives depend on which receiver responses they deem plausible. We show that with payoff knowledge, the limiting set of long-run learning outcomes is bounded above by *rationality-compatible equilibria* (*RCE*), and bounded below by *uniform RCE*. RCE refine the Intuitive Criterion (Cho and Kreps, 1987) and include all divine equilibria (Banks and Sobel, 1987). Uniform RCE sometimes but not always exists, and implies universally divine equilibrium.

1. Introduction

Signaling games typically have many perfect Bayesian equilibria, because Bayes rule does not pin down the receiver's off-path beliefs about the sender's type. Different off-path beliefs for the receiver can justify different off-path receiver behaviors, which in turn sustain equilibria with a variety of on-path outcomes. For this reason, applied work using signaling games typically invokes some equilibrium refinement to obtain a smaller and (hopefully) more accurate subset of predictions.

However, most refinements impose restrictions on the off-path beliefs without any reference to the process that might lead to equilibrium. As in our earlier paper, Fudenberg and He (2018), this paper uses a learning model to derive restrictions on out-of-equilibrium beliefs and thus restrict the equilibrium set. The innovation here is to restrict the agents' initial beliefs about opponents' strategies to reflect knowledge of the opponents' utility functions (and that the opponents act to maximize their expected utility). This generates restrictions on long-run play that refine the Intuitive Criterion (Cho and Kreps, 1987) and include all divine equilibria (Banks and Sobel, 1987).

In our learning model, agents repeatedly play the same signaling game against random opponents each period. Agents are Bayesians who believe they face a fixed but unknown distribution of the opposing players' strategies. Importantly, the senders start with independent prior beliefs about how receivers respond to different signals, so they cannot use

[☆] We thank Laura Doval, Glenn Ellison, Lorens Imhof, Yuichiro Kamada, Robert Kleinberg, David K. Levine, Kevin K. Li, Eric Maskin, Dilip Mookherjee, Harry Pei, Matthew Rabin, Bill Sandholm, Lones Smith, Joel Sobel, Philipp Strack, Bruno Strulovici, Tomasz Strzalecki, Jean Tirole, Juuso Toikka, and two anonymous referees for helpful comments and conversations, and National Science Foundation grant SES 1643517 for financial support.

[*] Corresponding author.
E-mail addresses: drew.fudenberg@gmail.com (D. Fudenberg), hesichao@gmail.com (K. He).

https://doi.org/10.1016/j.geb.2019.11.011
0899-8256/© 2019 Elsevier Inc. All rights reserved.

the response to one signal to infer anything about the distribution of responses to a different signal. This introduces an exploration-exploitation trade-off, as each sender only observes the response to the one signal she sends each period. Long-lived and patient senders will therefore experiment with every signal that they think might yield a substantially higher payoff than the myopically optimal signal. Thus, "out of equilibrium" signals actually arise with positive probability as experiments. The key to our results is that different types of senders have different incentives for experimenting with various signals, so that some of the sender types will send certain signals more often than other types do. Consequently, even though long-lived senders only experiment for a vanishingly small fraction of their lifetimes, the play of the long-lived receivers will best respond to beliefs about the senders' types that reflects this difference in experimentation probabilities.

Of course, the senders' experimentation incentives depend on their prior beliefs about which receiver responses are plausible after each signal. The set of learning outcomes depends on the assumptions we make on the set of "allowable" priors over opponents' strategies. Fudenberg and He (2018) restricted the set of allowable priors to be *non-doctrinaire,* which implies that agents always assign a strictly positive probability to their opponents playing strictly dominated strategies. An equilibrium profile is *patiently stable* if it is a long-run outcome with patient and long-lived agents for some non-doctrinaire prior; Fudenberg and He (2018) shows that patiently stable profiles of signaling games must satisfy a condition called the *compatibility criterion* (CC).

In this paper, we instead assume that the players' prior beliefs encode knowledge of their opponents' payoff functions, so in particular the senders assign zero probability to the event that the receivers choose conditionally dominated actions after any signal.[1] Inexperienced senders with full-support beliefs about the receivers' play may experiment with a signal in the hope that the receivers respond with a certain favorable action, not knowing that this action will never be played as it is not a best response to any receiver belief. With payoff information, even very patient senders will never undertake such experiments. Conversely, receivers know that no sender type would ever want to play a signal that does not best respond to any receiver strategy, because no possible response by the receiver would make playing that signal worthwhile. For this reason, the receivers' beliefs after each signal assign probability zero to the types for whom that signal is dominated.

We introduce equilibrium refinements for signaling games that provide upper and lower bounds on the set of *rationally patiently stable* learning outcomes, the analog of patient stability when the allowable priors reflect payoff knowledge and are otherwise non-doctrinaire. Theorem 1 shows that every rationally patient learning outcome where receivers have strict best responses to the on-path signals must be a "rationality compatible equilibrium" (RCE); this is an equilibrium where off-path beliefs satisfy restrictions derived from the comparative experimentation frequencies of different sender types when they know the receiver's payoff function. Conversely, Theorem 2 shows that every equilibrium that satisfies a uniform version of rational compatibility (uRCE) and some strictness assumptions can arise as a patient learning outcome. That is, for equilibria satisfying the relevant strictness assumptions, we have

$$\text{uRCE} \subseteq \text{rationally patiently stable profiles} \subseteq \text{RCE}$$

Importantly, the set of allowable priors with payoff information are not nested with the priors we considered in Fudenberg and He (2018): Payoff information rules out priors that assign positive probability to strictly dominated actions, while the full-support priors of Fudenberg and He (2018) require that strictly dominated strategies receive positive probability. Thus the sets of patiently stable profiles and rationally patiently stable profiles are not in general nested. Through a pair of examples, we illustrate that RCE better tracks the set of rationally patiently stable outcomes than the solution concepts from Fudenberg and He (2018) do.

In Example 2, the sender can be either strong or weak. The sender has a safe option **Out** with a known payoff, and a risky option **In** whose payoff depends on the receiver's response. The receiver has three responses to **In**: **Up**, which is optimal against the strong sender; **Down**, which is optimal against the weak sender, and **X**, which is never optimal. We show that RCE refines Fudenberg and He (2018)'s CC, and identifies the unique rationally patiently stable strategy profile. Intuitively, this is because when priors encode payoff information, the strong types experiment more with **In** than the weak types do against any play of the receivers, which is not true when senders do not know the receivers' payoff functions.

In some other games, the set of rationally patiently stable profiles is larger than that of patiently stable profiles, because payoff information leads to some experiments not being taken at all. In Example 3, there is a signal that is never used as an experiment when senders have payoff information, so the receivers' beliefs and behavior after this signal are arbitrary. On the other hand, when senders are ignorant of the receivers' payoff functions, one sender type will experiment much more frequently with this signal than the other type, leading to a restriction on the receivers' off-path beliefs and off-path behavior after the signal. In this game, RCE exactly identifies the set of rationally patiently stable strategy profiles, while the *strong CC* (which is shown by Fudenberg and He (2018) to be necessary for patient stability without payoff knowledge) rules out some rationally patiently stable profiles.

Like RCE and uRCE, standard equilibrium refinements do reflect the idea that players know that their opponents will not play strictly dominated strategies. Moreover, as we explain in Section 3, the nesting relationships

$$\text{uRCE} \subsetneq \text{universally divine} \subsetneq \text{divine} \subsetneq \text{RCE} \subsetneq \text{Intuitive Criterion} \subsetneq \text{Nash}$$

[1] An action is *conditionally dominated* after a given signal if it does not best respond to any belief about the sender's type.

hold for all equilibria where the receiver has strict incentives after all on-path signals. In particular, our learning-based belief restrictions resemble those imposed by divine equilibrium (Banks and Sobel, 1987): Every divine equilibrium is also an RCE and every uRCE is path-equivalent to a universally divine equilibrium. We should point out, though, that while every game has at least one rationally patiently stable profile, a uRCE need not exist; Example 4 is one simple case where one does not.

1.1. Related literature

This paper is most closely related to the work of Fudenberg and Levine (1993), Fudenberg and Levine (2006), and Fudenberg and He (2018) on patient learning by Bayesian agents who believe they face a steady-state distribution of play. Except for the support of the agents' priors, our learning model is exactly the same as that of Fudenberg and He (2018), and the proof of Theorem 1 follows the lines of our results there. Theorem 2 is the main technical innovation. It establishes a sufficient condition for an equilibrium to be rationally patiently stable. The proof of this sufficient condition for rational patient stability constructs a suitable prior and analyzes the corresponding rationally patiently stable profiles. The only other constructive sufficient condition[2] for strategy profiles to be patiently stable is Theorem 5.5 of Fudenberg and Levine (2006), which only applies to a subclass of perfect-information games. In such games the relative probabilities of various off-path actions do not matter, because each off-path experiment is perfectly revealed when it occurs. Indeed, the central lemma leading to Theorem 2 constructs a prior belief to ensure that the receivers correctly learn the relative frequencies that different types undertake various off-path experiments. This lemma deals with an issue specific to signaling games, and is not implied by any result in Fudenberg and Levine (2006).

Like this paper, Cho and Kreps (1987) and Banks and Sobel (1987) study equilibrium refinements in signaling games. We compare our learning-based equilibrium refinements with their refinements, which implicitly assume that players are certain of the payoff functions of their opponents.

Our paper is also related to other models of Bayesian non-equilibrium learning, such as Kalai and Lehrer (1993) and Esponda and Pouzo (2016), though these papers do not study optimal experimentation and do not refine the Nash equilibrium set. Finally, Dekel et al. (1999), Fudenberg and Kamada (2015), and Fudenberg and Kamada (2018) develop equilibrium refinements that combine the idea of equilibrium arising from learning with the assumption that players know one another's payoff functions and feedback structures. These paper do no provide explicit learning models, but the implicit models they have in mind would feature impatient learners who do little or no experimentation.

2. Two equilibrium refinements for signaling games

2.1. Signaling game notation

A *signaling game* has two players, a sender ("she," player 1) and a receiver ("he," player 2). At the start of the game, the sender learns her type $\theta \in \Theta$, but the receiver only knows the sender's type distribution[3] $\lambda \in \Delta(\Theta)$. Next, the sender chooses a signal $s \in S$. The receiver observes s and chooses an action $a \in A$ in response. We assume that Θ, S, A are finite and that $\lambda(\theta) > 0$ for all θ.

The players' payoffs depend on the triple (θ, s, a). Let $u_1 : \Theta \times S \times A \to \mathbb{R}$ and $u_2 : \Theta \times S \times A \to \mathbb{R}$ denote the utility functions of the sender and the receiver, respectively.

For $P \subseteq \Delta(\Theta)$, we have

$$\text{BR}(P, s) := \bigcup_{p \in P} \left(\underset{a \in A}{\arg\max}\, \mathbb{E}_{\theta \sim p} [u_2(\theta, s, a)] \right)$$

as the set of best responses to s supported by some belief in P. Letting $P = \Delta(\Theta)$, the set $A_s^{\text{BR}} := \text{BR}(\Delta(\Theta), s) \subseteq A$ contains the receiver actions that best respond to some belief about the sender's type after s. We say that actions in A_s^{BR} are *conditionally undominated after signal s*, and that actions in $A \backslash A_s^{\text{BR}}$ are *conditionally dominated after signal s*. We denote by $\Pi_2^\bullet := \times_{s \in S} \Delta(A_s^{\text{BR}})$ the *rational receiver strategies;* these are the strategies that assign probability 0 to conditionally dominated actions.[4] The rational receiver strategies form a subset of $\Pi_2 := \times_{s \in S} \Delta(A)$, the set of all receiver strategies. A sender who knows the receiver's payoff function expects the receiver to choose a strategy in $\Pi_2^\bullet$.

A *sender strategy* $\pi_1 = (\pi_1(\cdot \mid \theta))_{\theta \in \Theta} \in \Pi_1$ specifies a distribution on S for each type, $\pi_1(\cdot \mid \theta) \in \Delta(S)$. For a given π_1, signal s is *off the path of play* if it has probability 0, i.e. $\pi_1(s \mid \theta) = 0$ for all θ. Let

[2] "Constructive," as opposed to proofs that rule out all but one equilibrium using necessary conditions and then appeal to an existence theorem for patiently stable steady states. Constructive sufficient conditions allow us to characterize learning outcomes more precisely in games where multiple equilibria satisfy the necessary conditions, such as Example 1.

[3] The notation $\Delta(X)$ means the set of all probability distributions on X.

[4] Throughout we adopt the terminology "strategies" to mean behavior strategies, not mixed strategies.

$$S_\theta := \bigcup_{\pi_2 \in \Pi_2} \left(\arg\max_{s \in S} u_1(\theta, s, \pi_2(\cdot \mid s)) \right)$$

be the set of signals that best respond to some (not necessarily rational) receiver strategy for type θ. Signals in $S \backslash S_\theta$ are *dominated* for type θ, and $\Pi_1^\bullet := \times_\theta \Delta(S_\theta)$ denotes the *rational sender strategies* where no type ever sends a dominated signal. We also write Θ_s for the types θ for whom $s \in S$ is not dominated. A receiver who knows the sender's payoff function expects the sender to choose a strategy in $\Pi_1^\bullet$ and only expects types in Θ_s to play signal s.

2.2. Rationality-compatible equilibria

We now introduce rationality-compatible equilibrium (RCE) and uniform rationality-compatible equilibrium (uRCE), two refinements of Nash equilibrium in signaling games.

In Section 4, we develop a steady-state learning model where populations of senders and receivers, initially uncertain as to the aggregate play of the opponent population, undergo random anonymous matching each period to play the signaling game. We study the steady states when agents are patient and long lived, which we term *rationally patiently stable*. Under some strictness assumptions, we show that only RCE can be rationally patiently stable (Theorem 1) and that every uRCE is "path-equivalent"[5] to a rationally patiently stable profile (Theorem 2). Thus we provide a learning foundation for these solution concepts.

Our learning foundation will assume that agents know other agents' utility functions and know that other agents are rational in the sense of playing strategies that maximize the corresponding expected utilities. We will not however iteratively assume higher orders of payoff knowledge and rationality, so that we model "rationality" as opposed to "rationalizability."[6]

In the learning model, this implies senders' uncertainty about receivers' play is always supported on $\Pi_2^\bullet$ instead of Π_2, and similarly receivers' uncertainty about senders' play is supported on $\Pi_1^\bullet$ instead of Π_1. In Section 2.3, we discuss heuristically how our solution concepts capture some of the ways in which payoff information affects learning outcomes. This discussion will later be formalized in the context of the learning model we develop in Section 4.

Definition 1. Signal s is *more rationally-compatible* with θ' than θ'', written as $\theta' \succsim_s \theta''$, if for every $\pi_2 \in \Pi_2^\bullet$ such that

$$u_1(\theta'', s, \pi_2(\cdot|s)) \geq \max_{s' \neq s} u_1(\theta'', s', \pi_2(\cdot|s')),$$

we have

$$u_1(\theta', s, \pi_2(\cdot|s)) > \max_{s' \neq s} u_1(\theta', s', \pi_2(\cdot|s')).$$

In words, $\theta' \succsim_s \theta''$ means whenever s is a weak best response for θ'' against some rational receiver behavior strategy π_2, it is a strict best response for θ' against π_2.

The next proposition shows that $\succsim_s$ is transitive and "almost" asymmetric. A signal s is *rationally strictly dominant* for θ if it is a strict best response against any rational receiver strategy, $\pi_2 \in \Pi_2^\bullet$. A signal s is *rationally strictly dominated* for θ if it is not a weak best response against any rational receiver strategy.

Proposition 1. *We have*

1. $\succsim_{s'}$ *is transitive.*
2. *Except when s' is either rationally strictly dominant for both θ' and θ'' or rationally strictly dominated for both θ' and θ'', $\theta' \succsim_{s'} \theta''$ implies $\theta'' \not\succsim_{s'} \theta'$.*

Appendix A provides proofs for all of our results except where otherwise noted.

We require two auxiliary definitions before defining RCE.

Definition 2. For any two types θ', θ'', let $P_{\theta' \rhd \theta''}$ be the set of beliefs where the odds ratio of θ' to θ'' exceeds their prior odds ratio, that is[7]

$$P_{\theta' \rhd \theta''} := \left\{ p \in \Delta(\Theta) : \frac{p(\theta'')}{p(\theta')} \leq \frac{\lambda(\theta'')}{\lambda(\theta')} \right\}. \tag{1}$$

[5] Roughly speaking, this means equivalent up to changing some of the receiver's off-path behavior, see Subsection 3.3.

[6] It is straightforward to extend our results about RCE to priors that reflect higher-order knowledge of the rationality and payoff functions of the other player. The resulting equilibrium refinement always exists, and like RCE is implied by universal divinity. We do not include it here both because we are unaware of any interesting examples where the additional power has bite, and because we are skeptical about the hypothesis of iterated rationality.

[7] With the convention $\frac{0}{0} := 0$.

Note that if $\pi_1(s|\theta') \geq \pi_1(s|\theta'')$, $\pi_1(s|\theta') > 0$, and the receiver updates beliefs using π_1, then the receiver's posterior belief about the sender's type after observing s falls in the set $P_{\theta' \rhd \theta''}$. In particular, in any Bayesian Nash equilibrium, the receiver's on-path belief falls in $P_{\theta' \rhd \theta''}$ after any on-path signal s with $\theta' \succsim_s \theta''$.

We now introduce some additional definitions to let us investigate the implications of the agents' knowledge of their opponent's payoff function. For a strategy profile π^*, let $\mathbb{E}_{\pi^*}[u_1 \mid \theta]$ denote type θ's expected payoff under π^*.

Definition 3. For any strategy profile π^*, let

$$\tilde{J}(s, \pi^*) := \left\{ \theta \in \Theta : \max_{a \in A_s^{\mathrm{BR}}} u_1(\theta, s, a) \geq \mathbb{E}_{\pi^*}[u_1 \mid \theta] \right\}.$$

This is the set of types for which *some* best response to signal s is at least as good as their payoff under π^*. For all other types, the signal s is equilibrium dominated in the sense of Cho and Kreps (1987).

Definition 4. The set of *rationality-compatible beliefs* for the receiver at strategy profile π^*, $\left(\tilde{P}(s, \pi^*)\right)_s$, is defined as follows:

$$\begin{cases} \tilde{P}(s, \pi^*) := \Delta(\tilde{J}(s, \pi^*)) \bigcap \left(\bigcap_{(\theta', \theta'') \text{ s.t. } \theta' \succsim_s \theta''} P_{\theta' \rhd \theta''} \right) & \text{if } \tilde{J}(s, \pi^*) \neq \varnothing \\ \tilde{P}(s, \pi^*) := \Delta(\Theta_s) & \text{if } \tilde{J}(s, \pi^*) = \varnothing. \end{cases}$$

The main idea behind the rationality-compatible beliefs is that the receiver's posterior likelihood ratio for types θ' and θ'' dominates the prior likelihood ratio whenever $\theta' \succsim_s \theta''$. A second feature involves equilibrium dominance. Note that $\tilde{P}$ assigns probability 0 to equilibrium-dominated types; this is similar to the belief restriction of the Intuitive Criterion. Note that this definition imposes no belief restrictions based on $\theta' \succsim_s \theta''$ when s is equilibrium dominated for every type. As we illustrate in Example 3, the receiver needs not learn the rational compatibility relation when equilibrium dominance leads to steady states where no type ever experiments with a certain signal.

Definition 5. Strategy profile π^* is a *rationality-compatible equilibrium (RCE)* if it is a Nash equilibrium and $\pi_2^*(\cdot \mid s) \in \Delta(\mathrm{BR}(\tilde{P}(s, \pi^*), s))$ for every s.

RCE requires that the receiver only plays best responses to rationality-compatible beliefs after each signal. This solution concept allows for the possibility that after off-path signals the receiver's strategy $\pi_2^*(\cdot \mid s)$ may not correspond to a single belief about the sender's type. We show below that rationally patiently stable profiles exist and that every rationally patiently stable profile is a RCE, so that RCE exist as well.

Theorem 1 shows that RCE is a necessary condition for a strategy profile where receivers have strict preferences after each on-path signal to be rationally patiently stable. Intuitively, this result holds because the optimal experimentation behavior of the senders respects the compatibility order, and because, since players eventually learn the equilibrium path, types will not experiment much with signals that are equilibrium dominated. As we show in Section 3, RCE rules out the implausible equilibria in a number of games, but is weaker than some past signaling game refinements in the literature. However, RCE is only a necessary condition for rational patient stability, which leaves open the question of whether patient learning has additional implications. For this reason, we now define uRCE, a subset of RCE (up to path-equivalence). As we show below, uRCE is a sufficient condition for rational patient stability.

Definition 6. The set of *uniformly rationality-compatible beliefs* for the receiver is $\left(\hat{P}(s)\right)_s$ where

$$\hat{P}(s) := \Delta(\Theta_s) \bigcap \left(\bigcap_{(\theta', \theta'') \text{ s.t. } \theta' \succsim_s \theta''} P_{\theta' \rhd \theta''} \right).$$

Note that $\left(\hat{P}(s)\right)_s$ makes no reference to a particular strategy profile, unlike $\left(\tilde{P}(s, \pi^*)\right)_s$. Since $\Delta(\Theta_s)$ contains types for whom s is undominated and $\tilde{J}(s, \pi^*)$ contains types for whom s is equilibrium-undominated (relative to the profile π^*), we have $\tilde{P}(s, \pi^*) \subseteq \hat{P}(s)$ whenever $\tilde{J}(s, \pi^*) \neq \varnothing$.

Definition 7. A Nash equilibrium strategy profile π^* is called a *uniform rationality-compatible equilibrium (uRCE)* if for all θ, all off-path signals s and all $a \in \mathrm{BR}(\hat{P}(s), s)$, we have $\mathbb{E}_{\pi^*}[u_1 \mid \theta] \geq u_1(\theta, s, a)$.

The "uniformity" in uniform RCE comes from the requirement that *every* best response to *every* belief in $\hat{P}(s)$ deters *every* type from deviating to the off-path s. By contrast, a RCE is a Nash equilibrium where *some* best response to $\tilde{P}(s, \pi^*)$ deters every type from deviating to s. Unlike with RCE, uRCE need not exist (see Example 4).

As the names imply, uRCE is a stronger solution concept than RCE, up to path-equivalence.

Proposition 2. *Every uRCE is path-equivalent to an RCE.*

2.3. Examples

In this subsection, we show how to apply RCE and uRCE in specific games, and compare them to the solution concepts from Fudenberg and He (2018), which are based on necessary conditions for patient stability with full-support priors.

The following example illustrates that uRCE is a strict subset of RCE in some games.

Example 1. Suppose a worker has either high ability (θ_H) or low ability (θ_L). She chooses between three levels of higher education: **None** (**N**), **College** (**C**), or **Ph.D.** (**D**). An employer observes the worker's education level and pays a wage, $a \in \{\textbf{low}, \textbf{med}, \textbf{high}\}$. The worker's utility function is separable between wage and (ability, education) pair, with $u_1(\theta, s, a) = z(a) + v(\theta, s)$ where $z(\textbf{l}\text{ow}) = 0$, $z(\textbf{m}\text{ed}) = 6$, $z(\textbf{h}\text{igh}) = 9$ and $v(\theta_H, \textbf{N}) = 0$, $v(\theta_L, \textbf{N}) = 0$, $v(\theta_H, \textbf{C}) = 2$, $v(\theta_L, \textbf{C}) = 1$, $v(\theta_H, \textbf{D}) = -2$, $v(\theta_L, \textbf{D}) = -4$. (With this payoff function, going to college has a consumption value while getting a Ph.D. is costly.) The employer's payoffs reflect a desire to pay a wage corresponding to the worker's ability and increased productivity with education, given in the tables below.

N	**low**	**med**	**high**
θ_H	0, −2	6, 0	9, 1
θ_L	0, 1	6, 0	9, −2

C	**low**	**med**	**high**
θ_H	2, −1	8, 1	11, 2
θ_L	1, 2	7, 1	10, −1

D	**low**	**med**	**high**
θ_H	−2, 0	4, 2	7, 3
θ_L	−4, 3	2, 2	5, 0

No education level is dominated for either type and no wage is conditionally dominated after any signal. Since $v(\theta_H, \cdot) - v(\theta_L, \cdot)$ is maximized at **D**, it is simple to verify that $\theta_H \succsim_{\mathbf{D}} \theta_L$. Similarly, $\theta_L \succsim_{\mathbf{N}} \theta_H$. There is no compatibility relation at signal **C**.

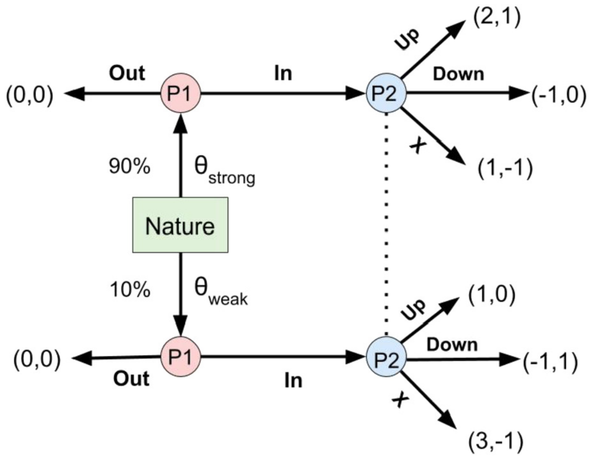
When the prior is $\lambda(\theta_H) = 0.5$, the strategy profile where the employer always pays a medium wage and both types of worker choose **C** is a uRCE. This is because $\hat{P}(\mathbf{N})$ contains only those beliefs with $p(\theta_H) \le 0.5$, so $\text{BR}(\hat{P}(\mathbf{N}), \mathbf{N}) = \{\textbf{low}, \textbf{med}\}$. Both of these wages deter every type from deviating to **N**. At the same time, no type wants to deviate to **D**, even if she gets paid the best wage.

On the other hand, the equilibrium π^* where the employer pays low wages for **N** and **C**, a medium wage for **D**, and both types choose **D** is an RCE but not a uRCE. The belief that puts probability 1 on the worker being θ_L belongs to $\tilde{P}(\mathbf{N}, \pi^*)$ and $\tilde{P}(\mathbf{C}, \pi^*)$ and induces the employer to choose low wage. However, medium salary is a best response to $\lambda \in \hat{P}(\mathbf{N})$ and medium wage would tempt type θ_L to deviate to **N**. ♦

In the learning model of Fudenberg and He (2018), agents do not know others' utility functions and have full-support prior beliefs about others' play. That paper's compatibility criterion (CC) is based on a family of binary relations on types (one for each signal s) that are less complete than the rational compatibility relations, because the CC requires the condition that "whenever s is a weak best response for θ'', it is also a strict best response for θ'" for all $\pi_2 \in \Pi_2$ instead of only for $\pi_2 \in \Pi_2^{\bullet}$. Hence, RCE is always at least as restrictive as the CC, and RCE can eliminate some equilibria that the CC allows.

Example 2. Consider a game where the sender has type distribution $\lambda(\theta_{\text{strong}}) = 0.9$, $\lambda(\theta_{\text{weak}}) = 0.1$ and chooses between two signals **In** or **Out**. The game ends with payoffs (0,0) if the sender chooses **Out**. If the sender chooses **In,** the receiver then chooses **Up**, **Down**, or **X**. **Up** is the receiver's optimal response if the sender is more likely to be θ_{strong}, **Down** is optimal when the sender is more likely to be θ_{weak}, and **X** is never optimal.[8] This game has two sequential equilibrium outcomes: one involving both types choosing **Out**, and another where both types go **In** and the receiver responds with **Up**.

[8] This is a modified version of Cho and Kreps (1987)'s "beer-quiche game," where an outside option with certain payoffs (**Out**) replaces the **Quiche** signal. The responses **Up** and **Down** correspond to **Not Fight** and **Fight** in the beer-quiche game, while **X** is a conditionally dominated response for the receiver following **In**. Also, while our definition of signaling games requires that the receiver has the same action set after every signal, this situation is clearly equivalent to one where the receiver chooses **Up**, **Down**, or **X** after **Out**, but all of these choices lead to the payoffs (0,0).



The sequential equilibrium outcome **Out** satisfies the CC, because the compatibility relation of Fudenberg and He (2018) does not rank the two types after signal **In**. (For example, if $\pi_2(\mathbf{Down} \mid \mathbf{In}) = 2/3$ and $\pi_2(\mathbf{X} \mid \mathbf{In}) = 1/3$, θ_{weak} finds **In** optimal but θ_{strong} does not.)

However, since the stronger rational compatibility relation ranks $\theta_{\text{strong}} \succsim_{\mathbf{In}} \theta_{\text{weak}}$, the unique RCE is the equilibrium where both types go **In**, which implies that this is also the unique rationally patiently stable outcome.[9] ♦

The previous example shows how RCE can exclude some equilibria that satisfy the CC. The next one cautions that RCE may allow more equilibrium profiles than the strong compatibility criterion (strong CC), which Fudenberg and He (2018) show to be another necessary condition for patient stability (with full-support priors). The strong CC requires the receiver to put zero probability after signal s on sender types for whom s is equilibrium dominated, but unlike in Definition 3, in the strong CC equilibrium dominance is computed by comparing the type's equilibrium payoff to that of any response to the unsent signal, including responses that are conditionally dominated.

Example 3. Consider a game with two sender types, θ_1 and θ_2, equally likely, and two possible signals, **L** or **R.** Payoffs are given in the tables below.

signal: **L**	action: a_1	action: a_2	action: a_3
type: θ_1	-2, 0	2, 2	2, 1
type: θ_2	-2, 1	2, 0	2, -1

signal: **R**	action: a_1	action: a_2	action: a_3
type: θ_1	5, -1	-3, 2	-4, 0
type: θ_2	-2, -1	1, 0	0, 1

Action a_1 is conditionally dominated for the receiver after signal **R**. It is easy to see that in every perfect Bayesian equilibrium π^*, we must have $\pi_1^*(\mathbf{L} \mid \theta_1) = \pi_1^*(\mathbf{L} \mid \theta_2) = 1$, $\pi_2^*(a_2 \mid \mathbf{L}) = 1$, and that $\pi_2^*(\cdot \mid \mathbf{R})$ must be supported on $A^{\text{BR}}_{\mathbf{R}} = \{a_2, a_3\}$. This means the off-path signal **R** is equilibrium dominated for every type in π^* (when they know the receiver's payoffs), i.e. $\tilde{J}(\mathbf{R}, \pi^*) = \varnothing$. So, $\tilde{P}(\mathbf{R}, \pi^*) = \Delta(\Theta_{\mathbf{R}}) = \Delta(\Theta)$ and RCE permits the receiver to play either a_2 or a_3 after **R**. (This is despite the fact that θ_2 is more rationally compatible with **R** than θ_1 is. As we discussed after Definition 4, RCE does not restrict the receiver's belief based on rational type compatibility after an off-path signal that is equilibrium dominated for every type.)

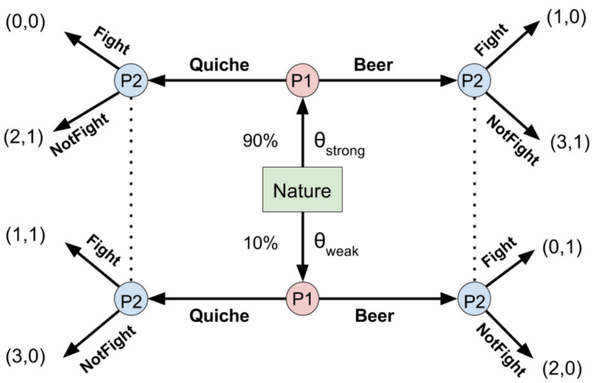
By contrast, the strong CC from Fudenberg and He (2018) requires that the receiver plays a_2 after **R**: the equilibrium payoff of both types is 2. Type θ_1 has $\max_{a \in A} u_1(\theta_1, \mathrm{R}, a) \geq 2$ but this is not true for θ_2. So the strong CC requires the receiver to put probability 1 on θ_1 after **R**, which pins down the receiver's off-path play.

We will show in Section 7 that when learners have payoff information, there is a rationally patiently stable profile where the receivers play a_2 after **R** and another rationally patiently stable profile where the receivers respond to **R** with a_3. However, we will also show that without payoff information, patient stability requires that the receivers play a_2 after **R**. ♦

[9] RCE requires that $p(\theta_{\text{strong}} \mid \mathrm{In}) \geq \lambda(\theta_{\text{strong}}) = 0.9$, which implies that $\pi_2(\mathbf{Up} \mid \mathbf{In}) = 1$ in every RCE. Therefore both types must be playing **In** in RCE.

Finally, we show that uRCE may not exist.

Example 4.



In Cho and Kreps (1987)'s "beer-quiche game," it is easy to verify that the pooling equilibrium on **Beer** is the unique RCE. This equilibrium is not a uRCE because $\hat{P}(\textbf{Quiche}) = \{p : p(\theta_{\text{weak}}) \geq 0.1\}$, so **NotFight** is a best response to a belief in $\hat{P}(\textbf{Quiche})$ that does not deter θ_{weak} from deviating. So the game does not have any uRCE. Moreover, it is easy to see that the same conclusions hold with slightly different assignments of payoffs to terminal nodes, so that the non-existence result applies to an open set of games. ♦

3. Comparison to other equilibrium refinements

This section compares RCE to other equilibrium refinement concepts in the literature.

3.1. Iterated dominance

We first relate RCE to a form of iterated dominance in the ex-ante strategic form of the game, where the sender chooses a signal π_1 as function of her type. We show that every sender strategy that specifies playing signal s as a less compatible type θ'' but not as a more compatible type θ' will be removed by iterated deletion. The idea is that such a strategy is never a weak best response to any receiver strategy in $\Pi_2^\bullet$: if the less compatible θ'' does not have a profitable deviation, then the more compatible type strictly prefers deviating to s.

Proposition 3. *Suppose $\theta' \succsim_s \theta''$. Then any ex-ante strategy of the sender π_1 with $\pi_1(s|\theta'') > 0$ but $\pi_1(s|\theta') < 1$ is removed by strict dominance once the receiver is restricted to using strategies in $\Pi_2^\bullet$.*

3.2. The Intuitive Criterion

We next relate RCE to the Intuitive Criterion.

Proposition 4. *Every RCE satisfies the Intuitive Criterion.*

The next example shows that the set of RCE is strictly smaller than the set of equilibria that pass the Intuitive Criterion.[10] The idea is that the Intuitive Criterion does not impose any restriction on the relative likelihood of two types after a signal that is not equilibrium dominated for either of them, but RCE can.

Example 5. Consider a signaling game where the prior probabilities of the two types are $\lambda(\theta') = 3/4$ and $\lambda(\theta'') = 1/4$, and the payoffs are:

signal: s'	action: a'	action: a''
type: θ'	4, 1	0, 0
type: θ''	6, 0	2, 1

signal: s''	action: a'	action: a''
type: θ'	7, 1	3, 0
type: θ''	7, 0	3, 1

[10] Fudenberg and He (2018)'s compatibility criterion is not always more stringent than the Intuitive Criterion, but the strong compatibility criterion studied in the same paper is.

Against any receiver strategy, the two types θ' and θ'' get the same payoffs from s'', but θ'' gets strictly higher payoffs than θ' from s'. So, $\theta' \succsim_{s''} \theta''$.

Consider now the Nash equilibrium in which the types pool on s', i.e. $\pi_1^*(s'|\theta') = \pi_1^*(s'|\theta'') = 1$, $\pi_2^*(a'|s') = 1$, and $\pi_2^*(a''|s'') = 1$. It passes the Intuitive Criterion since the off-path signal s'' is not equilibrium dominated for either type. On the other hand, RCE requires that every action played with positive probability in $\pi_2^*(\cdot|s'')$ best responds to some belief p about sender's type satisfying $\frac{p(\theta'')}{p(\theta')} \le \frac{\lambda(\theta'')}{\lambda(\theta')} = \frac{1}{3}$. But action a'' does not best respond to any such belief, so π^* is not an RCE. ◆

3.3. *Divine equilibrium*

Next, we compare divine equilibrium with RCE and uRCE. For a strategy profile π^*, let

$$D(\theta, s; \pi^*) := \{\alpha \in \text{MBR}(s) \text{ s.t. } \mathbb{E}_{\pi^*}[u_1 \mid \theta] < u_1(\theta, s, \alpha)\}$$

be the subset of mixed best responses[11] to s that would make type θ strictly prefer deviating from the strategy $\pi_1^*(\cdot \mid \theta)$. Similarly let

$$D^\circ(\theta, s; \pi^*) := \{\alpha \in \text{MBR}(s) \text{ s.t. } \mathbb{E}_{\pi^*}[u_1 \mid \theta] = u_1(\theta, s, \alpha)\}$$

be the set of mixed best responses that would make θ indifferent to deviating.

Intuitively, RCE and uRCE are "close" to divine equilibrium because both the definition of $\succsim_s$ and the divine equilibrium belief restriction involve a condition of the form "any receiver play that makes one type weakly prefer s must make another type strictly prefer s." Propositions 5 and 6 make this relationship precise.

Proposition 5.

1. *If π^* is a Nash equilibrium where s' is off-path, and $\theta' \succsim_{s'} \theta''$, then $D(\theta'', s'; \pi^*) \cup D^\circ(\theta'', s'; \pi^*) \subseteq D(\theta', s'; \pi^*)$.*
2. *Every divine equilibrium is a RCE.*

However, the converse is not true, as the following example illustrates.

Example 6. Consider the following signaling game with two types and three signals, with prior $\lambda(\theta_1) = 2/3$.

s'	a'	a''
θ'	0, 1	−1, 0
θ''	0, 0	−1, 1

s''	a'	a''
θ'	2, 1	−1, 0
θ''	1, 0	−1, 1

s'''	a'	a''
θ'	5, 0	−3, 1
θ''	0, 1	−2, 0

We check that the following is a pure-strategy RCE: $\pi_1(s'|\theta') = \pi_1(s'|\theta'') = 1, \pi_2(a'|s') = 1, \pi_2(a''|s'') = 1, \pi_2(a''|s''') = 1$. Evidently π is a Nash equilibrium and no type is equilibrium-dominated at any off-path signal. We now check that we do not have $\theta' \succsim_{s''} \theta''$ or $\theta'' \succsim_{s'''} \theta'$. Observe that against the receiver strategy $\tilde{\pi}_2(a'|s) = \frac{1}{2}$ for every s, s'' is strictly optimal for θ'' but s''' is strictly optimal for θ', so $\theta' \not\succsim_{s''} \theta''$. And for the receiver strategy $\hat{\pi}_2(a'|s) = 1$ for every s, s''' is strictly optimal for θ' but s'' is strictly optimal for θ'', so $\theta'' \not\succsim_{s'''} \theta'$. This shows the strategy profile is an RCE.

However, $D(\theta'', s''; \pi) \cup D^\circ(\theta'', s''; \pi)$ is the set of distributions on $\{a', a''\}$ that put at least weight 0.5 on a'. Any such distribution is in $D(\theta', s''; \pi)$. So in every divine equilibrium, the receiver plays a best response to a belief that puts weight no less than 2/3 on θ' after signal s'', which can only be a'.[12] ◆

This example illustrates one difference between divine equilibrium and RCE: Under divine equilibrium, the beliefs after signal s'' only depend on the comparison between the payoffs to s'' with those of the equilibrium signal s', while the compatibility criterion also considers the payoffs to a third signal s'''. In the learning model, this corresponds to the possibility that θ' chooses to experiment with s''' at beliefs that induce θ'' to experiment with s''.

RCE differs from divine equilibrium in another way, as divine equilibrium involves an *iterative* application of a belief restriction. The next example illustrates this difference.[13]

[11] To be precise, $\text{MBR}(p, s) := \underset{\alpha \in \Delta(A)}{\arg\max}(\mathbb{E}_{\theta \sim p}[u_2(\theta, s, \alpha)])$ and $\text{MBR}(s) := \cup_{p \in \Delta(\Theta)} \text{MBR}(p, s)$.

[12] As noted by Van Damme (1987), it may seem more natural to replace the set $\alpha \in \text{MBR}(m)$ in the definitions of D and D^0 with the larger set $\alpha \in \text{co}(\text{BR}(s))$, which leads to the weaker equilibrium refinement that Sobel et al. (1990) call "co-divinity". This example also shows that RCE need not be co-divine.

[13] We thank Joel Sobel for this example.

Example 7. There are three types, $\theta', \theta'', \theta'''$, all equally likely. The signal space is $S = \{s', s''\}$, and the set of receiver actions is $A = \{a^1, a^2, a^3, a^4\}$. When any sender type chooses the signal s', all parties get a payoff of 0 regardless of the receiver's action. When the sender chooses s'', the payoffs are determined by the following matrix.

s''	a^1	a^2	a^3	a^4
θ'	1, 0.9	−1, 0	−2, 0	−7, 0
θ''	5, 0	3, 1	−1, 0	−5, 0.8
θ'''	−3, 0	5, 0	1, 1.7	−3, 0.8

Consider the pure strategy profile $\pi_1^*(s'|\theta) = 1$ for all $\theta \in \Theta$ and $\pi_2^*(a^4|s) = 1$ for all $s \in S$. Since θ'' gains more from deviating to s'' than θ' does, applying the divine belief restriction for the off-path signal s'' eliminates the action a^1, since it is not a best response to any belief $p \in \Delta(\Theta)$ with $p(\theta'') \geq p(\theta')$. But after action a^1 is deleted for the receiver after signal s'', type θ''' now gains more from deviating to s'' than θ'' does. So, applying the divine belief restriction again eliminates actions a^2 and a^4, since it is not a best response against any $p \in \Delta(\Theta)$ with $p(\theta') = 0$ (for now s'' is equilibrium dominated for θ') and $p(\theta''') \geq p(\theta'')$. So π^* is not a divine equilibrium.

On the other hand, no type is equilibrium dominated at s'' and the only rational compatibility order is $\theta'' \succsim_{s''} \theta'$. But a^4 is a best response against the belief $p(\theta') = 0$, $p(\theta'') = 0.6$, $p(\theta''') = 0.4$, which belongs to the set $\Delta(\Theta_{s''}) \bigcap P_{\theta'' \triangleright \theta'}$. So π^* is an RCE. ♦

Finally, we show that every uRCE is path-equivalent to an equilibrium that is not ruled out by the "NWBR in signaling games" test (Banks and Sobel, 1987; Cho and Kreps, 1987),[14] which comes from iterative applications of the following pruning procedure: after signal s the receiver is required to put 0 probability on those types θ such that

$$D^\circ(\theta, s; \pi^*) \subseteq \cup_{\theta' \neq \theta} D(\theta', s; \pi^*).$$

If this would delete every type, then the procedure instead puts no restriction on receiver's beliefs and no type is deleted.

By "path-equivalent" we mean that by modifying some of the receiver's off-path responses, but without altering the sender's strategy or the receiver's on-path responses, we can change the uRCE into another uRCE that passes the NWBR test. Since every equilibrium passing the NWBR test is universally divine[15] (Cho and Kreps, 1987), this implies that every uRCE is path-equivalent to a universally divine equilibrium.

Proposition 6. *Every uRCE is path-equivalent to a uRCE that passes the NWBR test.*

Corollary 1. *Every uRCE is path-equivalent to a universally divine equilibrium.*

3.4. Summary

To summarize this subsection, we note that for strategy profiles that are on-path strict for the receiver, we have the following inclusion relationships. The first inclusion should be understood as inclusion up to path-equivalence. We use the symbol "$\subsetneq$" to mean that the former solution set is always nested within the latter one in every signaling game, and that there exist games where the nesting relationship is strict.

$$\text{uRCE} \subsetneq \text{universally divine} \subsetneq \text{divine} \subsetneq \text{RCE} \subsetneq \text{Intuitive Criterion} \subsetneq \text{Nash}.$$

In interpreting these inclusions, it is important to remember that a universally divine equilibrium generically exists. In contrast, we showed in Example 4 that uRCE can fail to exist for an open set of signaling games.[16]

4. Steady-state learning in signaling games

4.1. Random matching and aggregate play

We study the same discrete-time steady-state learning model as Fudenberg and He (2018) except for a different restriction on the players' prior beliefs over other players' strategies.

[14] This is closely related to, but not the same as, the NWBR property of Kohlberg and Mertens (1986).

[15] Universal divinity is defined as the iterative application of the following procedure: after signal s the receiver is required to put 0 probability on those types θ such that

$$D^\circ(\theta, s; \pi^*) \cup D(\theta, s; \pi^*) \subseteq \cup_{\theta' \neq \theta} D(\theta', s; \pi^*).$$

[16] That is, for an open set of payoff vectors at the terminal nodes of the game.

There is a continuum of agents in the society, with a unit mass of receivers and $\lambda(\theta)$ mass of type θ senders. Each population is further stratified by age, with a fraction $(1-\gamma)\cdot\gamma^t$ of each population age t for $t = 0, 1, 2, \ldots$ At the end of each period, each agent has probability $0 \le \gamma < 1$ of surviving into the next period, increasing their age by 1. With complementary probability, the agent dies. Each agent's survival is independent of calendar time and independent of the survival of other agents. At the start of the next period, $(1-\gamma)$ new receivers and $\lambda(\theta)(1-\gamma)$ new type θ senders are born into the society, thus preserving population sizes and the age distribution.

Agents play the signaling game every period against a randomly matched opponent. Each sender has probability $(1-\gamma)\gamma^t$ of matching with a receiver of age t, while each receiver has probability $\lambda(\theta)(1-\gamma)\gamma^t$ of matching with a type θ sender of age t.

4.2. Learning by individual agents with payoff knowledge

Each agent is born into a player role in the signaling game: either a receiver or a type θ sender. Agents know their role, which is fixed for life. The agents' payoff each period is determined by the outcome of the signaling game they played, which consists of the sender's type, the signal sent, and the action played in response. The agents observe this outcome, but the sender does not observe how her matched receiver would have played had she sent a different signal.

In addition to only surviving to the next period with probability $0 \le \gamma < 1$, agents discount[17] future utility flows by $0 \le \delta < 1$ and seek to maximize expected discounted utility. Letting u_t represent the payoff t periods from today, each agent's objective function is $\mathbb{E}[\sum_{t=0}^{\infty}(\gamma\delta)^t \cdot u_t]$. (Define $0^0 := 1$, so that a myopic agent just maximizes current period's expected payoff in every period.)

Agents believe they face a fixed but unknown distribution of opponents' aggregate play, updating their beliefs at the end of every period based on the outcome in their own game. Formally, each sender is born with a prior density function over receivers' behavior strategies, $g_1 : \Pi_2 \to \mathbb{R}_+$. Similarly, each receiver is born with a prior density over the senders' behavior strategies, $g_2 : \Pi_1 \to \mathbb{R}_+$. We denote the marginal distribution of g_1 on signal s as $g_1^{(s)} : \Delta(A) \to \mathbb{R}_+$, so that $g_1^{(s)}(\pi_2(\cdot|s))$ is the density of the new senders' prior over how receivers respond to signal s. Similarly, we denote the θ marginal of g_2 as $g_2^{(\theta)} : \Delta(S) \to \mathbb{R}_+$, so that $g_2^{(\theta)}(\pi_1(\cdot|\theta))$ is the new receivers' prior density over the signal choice of type θ.

We now state a regularity assumption on agents' priors that will be maintained throughout.

Definition 8. A prior $g = (g_1, g_2)$ is **regular** if

(a) [*independence*] $g_1(\pi_2) = \prod_{s\in S} g_1^{(s)}(\pi_2(\cdot|s))$ and $g_2(\pi_1) = \prod_{\theta\in\Theta} g_2^{(\theta)}(\pi_1(\cdot|\theta))$.
(b) [*payoff knowledge*] g_1 puts probability 1 on $\Pi_2^\bullet$ and g_2 puts probability 1 on $\Pi_1^\bullet$.
(c) [g_1 *non-doctrinaire*] g_1 is continuous and strictly positive on the interior of $\Pi_2^\bullet$.
(d) [g_2 *nice*] For each type θ, there are positive constants $\left(\alpha_s^{(\theta)}\right)_{s\in S}$ such that

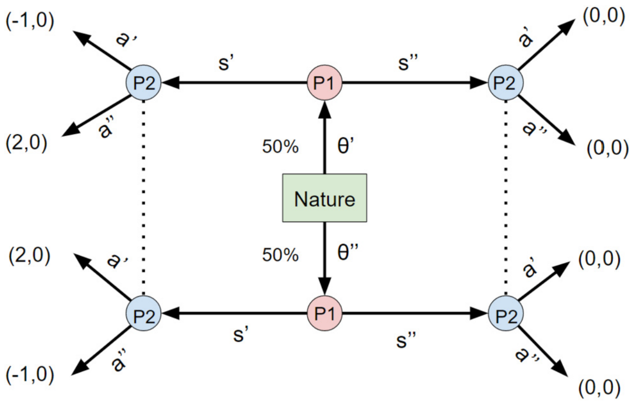
$$\pi_1(\cdot|\theta) \mapsto \frac{g_2^{(\theta)}(\pi_1(\cdot|\theta))}{\prod_{s\in S}\pi_1(s|\theta)^{\alpha_s^{(\theta)}-1}}$$

is uniformly continuous and bounded away from zero on the relative interior of $\Pi_\theta^\bullet$, the set of rational behavior strategies of type θ.

This assumption bears the same name as the regularity assumption in Fudenberg and He (2018), and is identical except that agents now know others' payoffs and others' rationality. In the learning model, this payoff knowledge translates into a restriction on the supports of the priors g_1, g_2, reflecting a dogmatic belief that senders will never play dominated signals and receivers will never play conditionally dominated actions. (These beliefs are correct in the learning model.)

Even with payoff knowledge, the receiver's prior can assign positive probability to ex-ante dominated sender strategies. For instance, in the signaling game below,

[17] We separately consider survival probability and patience so that we may consider agents who are impatient relative to their expected lifespan. Such agents experiment early in their life cycle, but spend most of their life myopically best responding to their beliefs, which makes our analysis more tractable.



the sender strategy $\pi_1(s'' \mid \theta') = \pi_1(s'' \mid \theta'') = 1$ belongs to the set $\Pi_1^\bullet$, and so must belong to the support of any regular receiver prior. But, even though $s'' \in S_{\theta'}$ and $s'' \in S_{\theta''}$, the receiver strategies to which they respectively best respond form disjoint sets, and π_1 is ex-ante dominated because it is not a best response to any single receiver strategy. It is nevertheless consistent for a receiver who knows the sender's payoff as a function of their type to assign positive density to π_1, because different types of agents can choose best responses to different beliefs about receiver play.

4.3. History and aggregate play

Let $Y_\theta[t] := (\cup_{s \in S}(s \times A_s^{\text{BR}}))^t$ represent the set of possible histories for a type θ sender with age t. Note that a valid history encodes the signal that θ sent each period and the (conditionally undominated) action that her opponent played in response. Let $Y_\theta := \bigcup_{t=0}^{\infty} Y_\theta[t]$ be the set of all histories for type θ.

Similarly, write $Y_2[t] := (\Theta \times S_\theta)^t$ for the set of possible histories for a receiver with age t. Each period, his history encodes the type of the matched sender and the (undominated) signal observed. The union $Y_2 := \bigcup_{t=0}^{\infty} Y_2[t]$ then stands for the set of all receiver histories.

The agents' dynamic optimization problems discussed in Subsection 4.2 give rise to *optimal policies*[18] $\sigma_\theta : Y_\theta \to S_\theta$ and $\sigma_2 : Y_2 \to \times_s(A_s^{\text{BR}})$. Here, $\sigma_\theta(y_\theta)$ is the signal that a type θ sender with history y_θ would send the next time she plays the signaling game. Analogously, $\sigma_2(y_2)$ is the pure extensive-form strategy that a receiver with history y_2 would commit to next time he plays the game. In the learning model, each agent solves a (single-agent) dynamic optimization problem, and chooses a deterministic optimal policy.

A *state* ψ of the learning model is a demographic description of how many agents have each possible history. It can be viewed as a distribution

$$\psi \in (\times_{\theta \in \Theta} \Delta(Y_\theta)) \times \Delta(Y_2),$$

and its components are denoted by $\psi_\theta \in \Delta(Y_\theta)$ and $\psi_2 \in \Delta(Y_2)$.

Since each state ψ is a distribution over histories and optimal policies map histories to play, ψ induces a distribution over play (i.e., a rational behavior strategy) in the signaling game $\sigma(\psi) \in \Pi^\bullet$, given by

$$\sigma_\theta(\psi_\theta)(s) := \psi_\theta \left\{ y_\theta \in Y_\theta : \sigma_\theta(y_\theta) = s \right\}$$

and

$$\sigma_2(\psi_2)(a \mid s) := \psi_2 \left\{ y_2 \in Y_2 : \sigma_2(y_2)(s) = a \right\}.$$

Here, $\sigma_\theta(\psi_\theta)$ and $\sigma_2(\psi_2)$ are the aggregate behaviors of the type θ and receiver populations in state ψ, respectively. Note that the aggregate play of a population can be stochastic even if the entire population uses the same deterministic optimal policy, because different senders will be matched with different receivers, and so different agents on the same side will observe different histories and play differently.

Of particular interest are the *steady states*, to be defined more precisely in Section 5. Loosely speaking, a steady state induces a time-invariant distribution over how the signaling game is played in the society.

5. Aggregate responses and steady state

This section defines the notion of a steady state using the "aggregate responses" of one population to the distribution of play in the other. These responses are defined using the "one-period forward" maps that describe how the agents' policies

[18] For notational simplicity, we suppress the dependence of these optimal policies on the effective discount factor $\delta\gamma$ and on the priors.

induce a map from current distributions over histories to what the distributions will be after the agents are matched and play the game using the strategies their policies prescribe.

5.1. The aggregate sender response

Fix the receivers' aggregate play at $\pi_2 \in \Pi_2^\bullet$ and fix an optimal policy σ_θ for each type θ. The *one-period-forward map for type* θ, f_θ, describes the distribution over histories that will prevail next period when the current distributions over histories in the type-θ population is ψ_θ. The next definition specifies the probability that $f_\theta[\psi_\theta, \pi_2]$ assigns to the history $(y_\theta, (s,a)) \in Y_\theta[t+1]$, that is to say a one-period concatenation of (s,a) onto the history $y_\theta \in Y_\theta[t]$.

Definition 9. The *one-period-forward map for type* θ, $f_\theta : \Delta(Y_\theta) \times \Pi_2^\bullet \to \Delta(Y_\theta)$ is

$$f_\theta[\psi_\theta, \pi_2](y_\theta, (s,a)) := \psi_\theta(y_\theta) \cdot \gamma \cdot \mathbf{1}\{\sigma_\theta(y_\theta) = s\} \cdot \pi_2(a \mid s)$$

and $f_\theta(\varnothing) := 1 - \gamma$.

To interpret, of the $\psi_\theta(y_\theta)$ fraction of the type-θ population with history y_θ, a γ fraction survives into the next period. The survivors all choose $\sigma_\theta(y_\theta)$ next period, which is met with response a with probability $\pi_2(a \mid \sigma_\theta(y_\theta))$.

Write f_θ^T for the T-fold application of f_θ on $\Delta(Y_\theta)$, holding fixed some π_2. It is easy to show that $\lim_{T\to\infty} f_\theta^T(\psi_\theta, \pi_2)$ exists and is independent of the initial ψ_θ. (This is because for any two states $\psi_\theta, \psi_\theta'$, the two distributions over histories $f_\theta^T(\psi_\theta, \pi_2)$ and $f_\theta^T(\psi_\theta', \pi_2)$ agree on all $Y_\theta[t]$ for $t < T$. As T grows large, the two resulting distributions must converge to each other since the fraction of very old agents with very long histories is rare.) Denote this limit as $\tilde{\psi}_\theta^{\pi_2}$. It is the distribution over type-θ history induced by the receivers' aggregate play π_2.

Definition 10. The *aggregate sender response* $\mathscr{R}_1 : \Pi_2^\bullet \to \Pi_1^\bullet$ is defined by

$$\mathscr{R}_1[\pi_2](s \mid \theta) := \tilde{\psi}_\theta^{\pi_2}(y_\theta : \sigma_\theta(y_\theta) = s)$$

That is, $\mathscr{R}_1[\pi_2](\cdot \mid \theta)$ describes the asymptotic aggregate play of the type-θ population when the aggregate play of the receiver population is fixed at π_2 each period. Note that $\mathscr{R}_1$ maps into $\Pi_1^\bullet$ because no type ever wants to send a dominated signal, even as an experiment, regardless of their beliefs about the receiver's response.

Technically, $\mathscr{R}_1$ depends on g_1, δ, and γ, just like σ_θ does. When relevant, we will make these dependencies clear by adding the appropriate parameters as superscripts to $\mathscr{R}_1$, but we will mostly suppress them to lighten notation.

5.2. The aggregate receiver response

We now turn to the receivers, who have a passive learning problem. They always observe the sender's type and signal at the end of each period, so their optimal policy σ_2 myopically best responds to the posterior belief at every history y_2.

Definition 11. The *one-period-forward map for the receivers* $f_2 : \Delta(Y_2) \times \Pi_1^\bullet \to \Delta(Y_2)$ is

$$f_2[\psi_2, \pi_1](y_2, (\theta, s)) := \psi_2(y_2) \cdot \gamma \cdot \lambda(\theta) \cdot \pi_1(s|\theta)$$

and $f_2(\varnothing) := 1 - \gamma$.

As with the one-period-forward maps f_θ for senders, $f_2[\psi_2, \pi_1]$ describes the distribution over receiver histories next period starting with a society where the distribution is ψ_2 and the sender population's aggregate play is π_1. We write $\tilde{\psi}_2^{\pi_1} := \lim_{T\to\infty} f_2^T(\psi_2, \pi_1)$ for the long-run distribution over Y_2 induced by fixing sender population's play at π_1. (This limit is again independent of the initial state ψ_2.)

Definition 12. The *aggregate receiver response* $\mathscr{R}_2 : \Pi_1^\bullet \to \Pi_2^\bullet$ is

$$\mathscr{R}_2[\pi_1](a \mid s) := \tilde{\psi}_2^{\pi_1}(y_2 : \sigma_2(y_2)(s) = a)$$

5.3. Steady states and rational patient stability

A *steady-state strategy profile* is a pair of mutual aggregate replies, so it is time-invariant under learning.[19]

[19] We focus on the steady states of the learning system, and do not study convergence to steady states.

Definition 13. π^* is a *steady-state strategy profile* if $\mathscr{R}_1^{g,\delta,\gamma}(\pi_2^*) = \pi_1^*$ and $\mathscr{R}_2^{g,\delta,\gamma}(\pi_1^*) = \pi_2^*$. Denote the set of all such strategy profiles as $\Pi^*(g,\delta,\gamma)$.

We now state two results about these steady states. We do not provide a proof because they follow easily from analogous results in Fudenberg and He (2018).

First, steady-state profiles always exist.

Proposition 7. *For any regular prior g and any $0 \le \delta, \gamma < 1$, $\Pi^*(g,\delta,\gamma)$ is non-empty and compact in the norm topology.*

The *rationally patiently stable strategy profiles* correspond to the set

$$\lim_{\delta\to 1}\lim_{\gamma\to 1} \Pi^*(g,\delta,\gamma).$$

This order of limits was first introduced in Fudenberg and Levine (1993). It ensures agents spend most of their lifetime playing myopically instead of experimenting, which is important for proving that rationally patiently stable profiles are Nash equilibria.

Definition 14. For each $0 \le \delta < 1$, a strategy profile π^* is *δ-stable under* g if there is a sequence $\gamma_k \to 1$ and an associated sequence of steady-state strategy profiles $\pi^{(k)} \in \Pi^*(g,\delta,\gamma_k)$, such that $\pi^{(k)} \to \pi^*$. Strategy profile π^* is *rationally patiently stable under* g if there is a sequence $\delta_k \to 1$ and an associated sequence of strategy profiles $\pi^{(k)}$ where each $\pi^{(k)}$ is δ_k-stable under g and $\pi^{(k)} \to \pi^*$. Strategy profile π^* is *rationally patiently stable* if it is rationally patiently stable under some regular prior g.

Note that δ-stable profiles always exist, since the space of strategy profiles is compact so we can always extract a convergent subsequence from a sequence of steady-state strategy profiles $\pi^{(k)} \in \Pi^*(g,\delta,\gamma_k)$ with $\gamma_k \to 1$. For the same reason, a rationally patiently stable profile always exists.

Proposition 8. *If strategy profile π^* is rationally patiently stable, then it is a Nash equilibrium.*

Note that Propositions 7 and 8 apply even if all of the Nash equilibria of the game are in mixed strategies; as noted above, the randomization here arises from the random matching process.

6. Rational patient stability, payoff knowledge, and equilibrium refinements

In this section, we relate the equilibrium refinements proposed in Section 2 to the steady-state learning model. We show that under certain strictness assumptions, RCE is necessary for rational patient stability while uRCE is sufficient for rational patient stability. We also discuss how payoff knowledge matters for learning outcomes.

6.1. RCE is necessary for rational patient stability

We show that any rationally patiently stable strategy profile satisfying a strictness assumption must be an RCE. The key lemma is analogous to Lemma 1 from Fudenberg and He (2018), so we will omit its proof.

Lemma 1. *Suppose $\theta' \succsim_s \theta''$. Then for any regular prior g_1, $0 \le \delta, \gamma < 1$, and any $\pi_2 \in \Pi_2^\bullet$, we have $\mathscr{R}_1[\pi_2](s \mid \theta') \ge \mathscr{R}_1[\pi_2](s \mid \theta'')$.*

This result says over their lifetimes, the relative frequencies with which different sender types experiment with signal s respect the rational compatible order $\succsim_s$. This follows from the fact that sender types who are more compatible with a signal will play it at least as often. The payoff knowledge embedded in g_1's support implies that senders never experiment in the hopes of seeing a response which is highly profitable for the sender but dominated for the receiver, such as the **X** action in Example 2 for θ_{weak}. This extra assumption leads to a stronger result than Lemma 1 from Fudenberg and He (2018), which is stated in terms of the less-complete compatibility order.

For a fixed strategy profile π and on-path signal s^*, let $\mathbb{E}_{\theta|\pi_1,s^*}[u_2(\theta,s^*,a)]$ denote the receiver's expected utility from responding to s^* with a, where the expectation over the sender's type θ is taken with respect to the posterior type distribution after signal s^* given the sender's strategy $\pi_1(\cdot \mid \theta)$.

Definition 15. A Nash equilibrium π^* is *on-path strict for the receiver* if for every on-path signal s^*, $\pi_2(a^* \mid s^*) = 1$ for some $a^* \in A$ and $\mathbb{E}_{\theta|\pi_1,s^*}[u_2(\theta,s^*,a^*)] > \max_{a \ne a^*} \mathbb{E}_{\theta|\pi_1,s^*}[u_2(\theta,s^*,a)]$.

We call this condition "on-path" strict for the receiver because we do not make assumptions about the receiver's incentives after off-path signals. For generic payoffs, all pure-strategy equilibria will be on-path strict for the receiver.

Theorem 1. *Every strategy profile that is rationally patiently stable and on-path strict for the receiver is an RCE.*

RCE rules out two kinds of receiver beliefs after signal s: those that assign non-zero probability to equilibrium-dominated sender types, and those that violate the rational compatibility order. The restriction on equilibrium dominated types uses the assumption that the receiver has a strict best response to each on-path signal to put a lower bound on how slowly aggregate receiver play at on-path signals converges to its limit.[20] The fact that the receiver beliefs respect the rational compatibility order comes from Lemma 1, which uses our assumptions about prior g to derive restrictions on the aggregate sender response $\mathscr{R}_1$, and show that these are reflected in the aggregate receiver response. The proof of Theorem 1 closely follows the analogous proof in Fudenberg and He (2018) and is omitted.

6.2. *Quasi-strict uRCE is sufficient for rational patient stability*

We now prove our main result: as a partial converse to Theorem 1, we show that under additional strictness conditions, every uRCE is path-equivalent to a rationally patiently stable strategy profile.

Definition 16. A *quasi-strict uRCE* π^* is a uRCE that is on-path strict for the receiver, strict for the sender (that is, there exists an equilibrium signal s^* for each type θ with $u_1(\theta, s^*, \pi_2^*(\cdot|s^*)) > \max_{s \neq s^*} u_1(\theta, s, \pi_2^*(\cdot|s))$, so every type strictly prefers its equilibrium signal to any other), and satisfies $\mathbb{E}_{\pi^*}[u_1 \mid \theta] > u_1(\theta, s', a)$ for all θ, all off-path signals s' and all $a \in \text{BR}(\hat{P}(s'), s')$.

The last condition in the definition of quasi-strictness requires that every best response to $\hat{P}(s')$ strictly deters every type from deviating to s', whenever s' is off-path. Every uRCE satisfies the weaker version of this condition where "strictly deters" is replaced with "weakly deters."

Theorem 2. *If π^* is a quasi-strict uRCE, then it is path-equivalent to a rationally patiently stable strategy profile.*

Theorem 2 provides a constructive argument for an equilibrium being rationally patiently stable in signaling games with multiple RCE, such as Example 1. It follows from three lemmas on $\mathscr{R}_1$ and $\mathscr{R}_2$ that are stated and proved in the rest of this subsection. Indeed, the theorem remains valid in any modified learning model where $\mathscr{R}_1$ and $\mathscr{R}_2$ satisfy the conclusions of these lemmas.

Recall that Example 4 showed a signaling game with a unique RCE that is not a uRCE. As noted before, a rationally patiently stable profile always exists. By Theorem 1, the unique RCE of the game must be rationally patiently stable. This shows uRCE is not a necessary condition for rational patient stability.

6.2.1. *$\mathscr{R}_1$ under a confident prior*

The first lemma shows that under a suitable prior, the aggregate sender response of the dynamic learning model approximates the sender's static best response function when applied to certain receiver strategies, namely strategies that are "close" to one inducing a unique optimal signal for each sender type. The precise meaning of "close" that we use treats on- and off-path responses differently, so it requires some auxiliary definitions.

Definition 17. Let π^* be a strategy profile where every type plays a pure strategy and the receiver plays a pure action after each on-path signal. Say π^* *induces a unique optimal signal for each sender type* if

$$\mathbb{E}_{\pi^*}[u_1 \mid \theta] > \max_{s \neq \pi_1^*(\theta)} u_1(\theta, s, \pi_2^*(\cdot|s))$$

for every type θ.

Starting with a strategy profile π^* that induces a unique optimal signal for each sender type, define for each off-path s in π^* the set of receiver actions $\tilde{A}(s) := \{a : \mathbb{E}_{\pi^*}[u_1 \mid \theta] > u_1(\theta, s, a) \ \forall \theta\}$ that strictly deter every type from deviation. Because π_2^* induces a unique optimal signal, each $\tilde{A}(s)$ must contain at least one element in the support of $\pi_2^*(\cdot|s)$, but could also contain other actions. It is clear that if π_2^* were modified off-path by changing each $\pi_2^*(\cdot|s)$ to be an arbitrary mixture over $\tilde{A}(s)$, then the resulting strategy profile would continue to induce (the same) unique optimal signal for each sender type.

For π^* that induces a unique optimal signal for each sender type, write $B_2^{\text{on}}(\pi^*, \epsilon)$ for the elements of $\Pi_2^\bullet$ no more than ϵ away from π_2^* at the on-path signals in π_1^*, that is

[20] If the receiver mixes after some equilibrium signal s for type θ, then our techniques for showing that θ does not experiment very much with equilibrium dominated signals do not go through, but we do not have a counterexample.

$$B_2^{\text{on}}(\pi^*,\epsilon) := \left\{\pi_2 \in \Pi_2^\bullet : |\pi_2(a|s) - \pi_2^*(a|s)| \le \epsilon, \forall a, \text{ on-path } s \text{ in } \pi^*\right\}.$$

Similarly, define $B_2^{\text{off}}(\pi^*,\epsilon)$ as the elements of $\Pi_2^\bullet$ putting no more than ϵ probability on actions outside of $\tilde{A}(s)$ after each off-path s, where $\tilde{A}(s)$ is the set of actions that would deter every type from deviating to s, as above.

$$B_2^{\text{off}}(\pi^*,\epsilon) := \left\{\pi_2 \in \Pi_2^\bullet : \pi_2(\tilde{A}(s)|s) \ge 1-\epsilon, \forall \text{ off-path } s \text{ in } \pi^*\right\}.$$

Lemma 2. *Suppose π^* induces a unique optimal signal for each sender type. Then there exists a regular prior g_1, some $0 < \epsilon_{off} < 1$, and a function $\gamma(\delta,\epsilon)$ valued in $(0,1)$, such that for every $0<\delta<1$, $0<\epsilon<\epsilon_{off}$, and $\gamma(\delta,\epsilon)<\gamma<1$, if $\pi_2 \in B_2^{on}(\pi^*,\epsilon) \cap B_2^{off}(\pi^*,\epsilon_{off})$, then $|\mathscr{R}_1^{g_1,\delta,\gamma}[\pi_2](s|\theta) - \pi_1^*(s|\theta)| < \epsilon$ for every θ and s.*

Note that the same ϵ appears in the hypothesis $\pi_2 \in B_2^{\text{on}}(\pi^*,\epsilon)$ as in the conclusion. That is, the aggregate sender response gets closer to π_1^* as receivers' play gets closer to π_2^*.

The idea is to specify a sender prior g_1 that is highly confident and correct about the receiver's response to on-path signals, and is also confident that the receiver responds to each off-path signal s with actions in $\tilde{A}(s)$. Take a signal s' other than the one that θ sends in π_1^*. If θ has not experimented much with s', then her belief is close to the prior and she thinks deviation does not pay. If θ has experimented a lot with s', then by the law of large numbers her belief is likely to be concentrated in $\tilde{A}(s')$, so again she thinks deviation does not pay. Since the option value for experimentation eventually goes to 0, at most histories all sender types are playing a myopic best response to their beliefs, meaning they will not deviate from π_1^*. The intuition is similar to that of Lemmas 6.1 and 6.4 from Fudenberg and Levine (2006), which says that we can construct a highly concentrated and correct prior so that in the steady state, most agents have correct beliefs about opponents' play both on and one step off the equilibrium path.

This lemma requires the assumption that π^* is strict for the sender. If s^* were only weakly optimal for θ in π^*, there could be receiver strategies arbitrarily close to π_2^* that make some other signal $s' \neq s^*$ strictly optimal for θ. In that case, we cannot rule out that a non-negligible fraction of the θ population will rationally play s' forever when the receiver population plays close to π_2^*.

6.2.2. $\mathscr{R}_2$ and learning rational compatibility

Let C be the set of sender strategies that respect the rational compatibility order, that is

$$C := \{\pi_1 \in \Pi_1^\bullet : \pi_1(s|\theta) \ge \pi_1(s|\theta') \text{ whenever } \theta \succsim_s \theta'\}.$$

The next lemma shows that there is a prior for the receivers so that when the aggregate sender play is any strategy in C, almost all receivers end up with beliefs consistent with the rational compatibility order. This lemma is the main technical contribution of the paper and enables us to provide a sufficient condition for rational patient stability when the *relative* frequencies of off-path experiments matter.

Lemma 3. *For each $\epsilon > 0$, there exists a regular receiver prior g_2 and $0 < \underline{\gamma} < 1$ so that for any $\underline{\gamma} < \gamma < 1$, $0 < \delta < 1$, and $\pi_1 \in C$,*

$$\mathscr{R}_2^{g_2,\delta,\gamma}[\pi_1](BR(\hat{P}(s),s) \mid s) \ge 1-\epsilon$$

for each signal s.

The key step in the proof is constructing a prior belief for the receivers so that when the senders' aggregate play is sufficiently close to the target equilibrium, the receiver beliefs respect the compatibility order. This step was not necessary in Fudenberg and Levine (2006), which is the only other paper that has given a sufficient condition for rational patient stability in a class of games.[21]

To prove Lemma 3, we construct a Dirichlet prior g_2 so that for any s such that $\theta' \succsim_s \theta''$, g_2 assigns much greater prior weight to θ' playing s than to θ'' playing s.[22] In the absence of data, the receiver strongly believes that the senders are using strategies π_1 such that $p(\theta''|s)/p(\theta'|s) \le \lambda(\theta'')/\lambda(\theta')$. This strong prior belief can only be overturned by a very large number of observations to the contrary. But because $\pi_1 \in C$ respects the rational compatibility order, if the receiver has a very large number of observations of senders choosing s, the law of large numbers implies this large sample is unlikely to lead the receiver to have a belief outside of $\hat{P}(s)$. So we can ensure that with high probability sufficiently long-lived receivers play a best response to $\hat{P}(s)$ after the off-path s.

[21] Their result guarantees that the receivers' beliefs about the frequency of type θ sending signal s is within ϵ of the truth. This is not sufficient for purposes, because when signal s has probability 0 under a given sender strategy, perturbing the strategy of every type by up to ϵ can generate arbitrary off-path beliefs about the sender's type.

[22] The Dirichlet prior is the conjugate prior to multinomial data, and corresponds to the updating used in fictitious play (Fudenberg and Kreps, 1993). It is readily verified that if each of $g_1^{(\theta)}$ and $g_2^{(s)}$ is Dirichlet and independent of the other components, then g is regular. In the proof, we work with Dirichlet priors since they give tractable closed-form expressions for the posterior mean belief of the opponent's strategy after a given history.

Finally, we state a lemma that says for any Dirichlet receiver prior, when lifetimes are long enough, receiver response approximates the receiver's best response function on-path when applied to a sender strategy that provides strict incentives after every on-path signal. Write $B_1^{\text{on}}(\pi^*, \epsilon)$ for the elements of $\Pi_1^\bullet$ where each type θ plays ϵ-close to $\pi_1^*(\cdot|\theta)$, that is

$$B_1^{\text{on}}(\pi^*, \epsilon) := \left\{\pi_1 \in \Pi_1^\bullet : |\pi_1(s|\theta) - \pi_1^*(s|\theta)| \le \epsilon, \forall \theta, s\right\}.$$

Lemma 4. *Fix a strategy profile π^* where the receiver has strict incentives after every on-path signal. For each regular Dirichlet receiver prior g_2, there exists $\epsilon_1 > 0$ and a function $\gamma(\epsilon)$ valued in $(0, 1)$, so that whenever $\pi_1 \in B_1^{on}(\pi^*, \epsilon_1)$, $0 < \delta < 1$, and $\gamma(\epsilon) < \gamma < 1$, we have $\mathscr{R}_2^{g_2,\delta,\gamma}[\pi_1](a|s) - \pi_2^*(a|s)| < \epsilon$ for every on-path signal s in π^* and a.*

The intuition is that when the aggregate sender strategy is close to π_1^*, the law of large numbers implies that after each signal that π_1^* gives positive probability, a receiver with enough data is likely to have a belief close to the Bayesian belief assigned by π_1^*. Coupled with the fact that π_1^* is on-path strict for the receiver, this lets us conclude that long-lived receivers play $\pi_2^*(\cdot|s)$ after every on-path s with high probability.

7. Payoff information and steady-state learning

We revisit the examples from Section 2.3 and discuss how prior beliefs reflecting knowledge or ignorance of payoff information lead to different implications for learning.

7.1. Example 2

In Example 2, it follows from Lemma 1 that for any $0 \le \delta, \gamma < 1$, any receiver play $\pi_2 \in \Pi_2^\bullet$, and any regular prior g, we have $\mathscr{R}_1^{g_1}[\pi_2](\textbf{In} \mid \theta_{\text{strong}}) \ge \mathscr{R}_1^{g_1}[\pi_2](\textbf{In} \mid \theta_{\text{weak}})$. In the absence of payoff information, we show that there exists a full-support prior g_1 so that, fixing π_2 to always play **Down**, we get $\mathscr{R}_1^{g_1}[\pi_2](\textbf{In} \mid \theta_{\text{strong}}) \le \mathscr{R}_1^{g_1}[\pi_2](\textbf{In} \mid \theta_{\text{weak}})$ for any $0 \le \delta, \gamma < 1$, with strict inequality for an open set of parameter values.

Underlying this is the fact that if the conditionally dominated response **X** is removed from the game tree, then θ_{strong} will experiment more frequently with **In** than θ_{weak} does because θ_{strong} potentially has more to gain. This story breaks down if senders do not know receivers' payoffs and thus suspect that **X** might be used after **In**. We now show that for some full-support prior beliefs, θ_{weak} experiments more with **In** than θ_{strong} does under *any* patience level when receivers always play **Down**.

Let $g_1^{(\text{In})}$ be Dirichlet with weights $(1, K, 1)$ on (**Up**, **Down**, **X**) for arbitrary $K \ge 4$. After observing $k \ge 0$ instances of receivers responding to **In** with **Down**, a sender would have the posterior Dirichlet$(1, K + k, 1)$. The θ_{weak} type's Gittins index for **In** would be unchanged if her payoffs to (**Up**, **Down**, **X**) were $(3, -1, 1)$ instead of $(1, -1, 3)$, by symmetry of her beliefs about **Up** and **X**. This observation shows her Gittins index for **In** is at least as large as θ_{strong}'s, whose payoffs to (**Up**, **Down**, **X**) are $(2, -1, 1)$. So the strong type switches away from **In** after fewer observations of **Down** than the weak type does (this includes the case of "switching away" after 0 observations of **Down**, i.e. the strong type never experimenting with **In**.) We have proven $\mathscr{R}_1^{g_1}[\pi_2](\text{In} \mid \theta_{\text{strong}}) \le \mathscr{R}_1^{g_1}[\pi_2](\text{In} \mid \theta_{\text{weak}})$ for any $0 \le \delta, \gamma < 1$.

Signal **In** is myopically suboptimal for both types, and by the previous argument, the minimum effective discount factor $\delta\gamma$ that would induce at least one period of experimentation with **In** is strictly higher for the strong type than the weak type. This shows for an open set of δ, γ parameters, $\mathscr{R}_1^{g_1}[\pi_2](\textbf{In} \mid \theta_{\text{strong}}) = 0$ but $\mathscr{R}_1^{g_1}[\pi_2](\textbf{In} \mid \theta_{\text{weak}}) > 0$.

7.2. Example 3

In Example 3 we showed that there is an RCE in which the receivers play a_3 after **R**. Because RCE is not a sufficient condition for rational patient stability, this leaves open the question of whether this strategy can arise in our learning model. Here we verify that it can, and also show that "a_3 after **R**" cannot be part of a patiently stable outcome in the absence of payoff information. This is because patient but inexperienced θ_1's without payoff information find it plausible that receivers choose a_1 after **R**, so they will experiment much more frequently with the off-path signal **R** than θ_2's, for whom every possible response to **R** leads to worse payoffs than their equilibrium payoff of 2. As a result, receivers learn that **R**-senders have type θ_1 so they respond with a_2. On the other hand, when senders know ex-ante that receivers will never choose a_1 after **R**, for some priors there are steady states where no one ever experiments with **R**. When this happens, the receivers' belief about the likelihood ratio of the types following the off-path **R** is governed by their initial beliefs about the distribution of sender play at the start of learning, which may be arbitrary and thus support a richer class of equilibrium profiles.

Specifically, in Example 3, suppose $g_1^{(\mathbf{L})}$ is Dirichlet $(1, 10, 1)$ over all three responses to **L**, while $g_1^{(\mathbf{R})}$ is Dirichlet$(1, 1)$ on $A_{\mathbf{R}}^{\text{BR}} = \{a_2, a_3\}$, which reflects the sender's knowledge that a_1 is a conditionally dominated response to **R**. And suppose that $g_2^{\theta_1}$ is the Dirichlet$(2, 1)$ distribution on {**L**, **R**} and $g_2^{\theta_2}$ is the Dirichlet$(2, x)$ distribution, where $x > 0$ is a free parameter. For any $0 \le \delta, \gamma < 1$, there exists a steady state where senders always choose **L** and receivers always respond to **L** with a_2. This is because the Gittins index for **R** is no larger than -3 for θ_1 and no larger than 1 for θ_2 after any history, while the myopic expected payoff of **L** already exceeds these values in the first period. The expected payoff of **L** only increases with additional observations of a_2 after **L**. On the receiver side, every positive-probability history y_2 must involve the senders playing **L** every period. Following such a history, the receiver believes θ_1 plays **L** with probability at least $\frac{2}{3}$, hence an **L**-sender is the θ_1 type with probability at least $\frac{2/3}{1+2/3} = \frac{2}{5}$. We have $a_2 \in \text{BR}(\{p\}, \mathbf{L})$ whenever $p(\theta_1) \geq \frac{2}{5}$, so we have shown that in the steady state receivers always play a_2 after **L**.

In this steady state, signal **R** is never sent, so by choosing different values of $x > 0$, we can sustain either a_2 or a_3 after **R** as part of a rationally patiently stable profile. To be more precise, let n_1 and n_2 count the number of times the two types of senders appear in a positive-probability history y_2. The receiver's posterior assigns the following likelihood ratio to the type of an **R**-sender:

$$\frac{1}{3+n_1} \Big/ \frac{x}{2+x+n_2} = \frac{1}{x} \cdot \left(\frac{2+x+n_2}{3+n_1} \right).$$

Since the two types are equally likely, the fraction of receivers with histories y_2 so that $0.9 \leq \left(\frac{2+x+n_2}{3+n_1} \right) \leq 1.1$ approaches 1 as $\gamma \to 1$. Depending on whether $x = 1/4$ or $x = 4$, these receivers will play a_2 or a_3 after **R**, so $\pi_2(a_2 \mid \mathbf{R}) = 1$ and $\pi_2(a_3 \mid \mathbf{R}) = 1$ are both δ-stable for any $\delta \geq 0$, under two different regular priors reflecting payoff knowledge.

By contrast, Theorem 3 of Fudenberg and He (2018) implies that if priors g_1, g_2 have full support on Π_2 and Π_1 respectively, then we must have a_2 after **R** in every patiently stable profile. The idea is that when senders are patient and long-lived, new θ_1 start off by trying **R** but new θ_2 start off by trying **L**. When receivers play a_2 after **L** with high probability, it is very unlikely that θ_2 ever switches away from **L**, providing a bound on their frequency of playing **R**. On the other hand, as their effective discount factor increases, θ_1 will spend arbitrarily many periods of its early life playing **R** in hopes of getting the best payoff of 5, lacking the payoff knowledge that a_1 is conditionally dominated for the receivers after **R**. Receivers therefore end up learning that **R**-senders have type θ_1.

8. Conclusion

This paper studies non-equilibrium learning about other players' strategies in the setting of signaling games. When the agents' prior beliefs about their opponents' play reflect prior knowledge of others' payoff functions, the steady states of societies of Bayesian learners can be bounded by two equilibrium refinements, RCE and uRCE, that is we get

$$\text{uRCE} \subseteq \text{rationally patiently stable profiles} \subseteq \text{RCE}.$$

Furthermore, every divine equilibrium is an RCE. This is not true for all of the equilibria that satisfy the Intuitive Criterion, which suggests that divine equilibrium does a better job of capturing the implications of learning. That said, we do not know the exact relationship between rational patient stability and divine equilibrium, and there is scope for sharpening our conclusions.

Divine equilibrium and RCE are only defined for signaling games. In general extensive-form games, agents may find it optimal to play strictly dominated strategies as experiments to learn about the consequences of their other strategies, so requiring prior beliefs to be supported on opponents' undominated strategies can lead to situations where agents observe play that they had assigned zero prior probability. We leave the associated complications for future work.

Appendix A

A.1. Proof of Proposition 1

Proof. To show (1), suppose $\theta' \succsim_{s'} \theta''$ and $\theta'' \succsim_{s'} \theta'''$. For any $\pi_2 \in \Pi_2^\bullet$ where s' is weakly optimal for θ''', it must be strictly optimal for θ'', hence also strictly optimal for θ'. This shows $\theta' \succsim_{s'} \theta'''$.

To establish (2), partition the set of rational receiver strategies as $\Pi_2^\bullet = \Pi_2^+ \cup \Pi_2^0 \cup \Pi_2^-$, where the three subsets refer to receiver strategies that make s' strictly better, indifferent, or strictly worse than the best alternative signal for θ''. If the set Π_2^0 is nonempty, then $\theta' \succsim_{s'} \theta''$ implies $\theta'' \not\succsim_{s'} \theta'$. This is because against any $\pi_2 \in \Pi_2^0$, signal s' is strictly optimal for θ' but only weakly optimal for θ''. At the same time, if both Π_2^+ and Π_2^- are nonempty, then Π_2^0 is nonempty. This is because both $\pi_2 \mapsto u_1(\theta'', s', \pi_2(\cdot|s'))$ and $\pi_2 \mapsto \max_{s'' \neq s'} u_1(\theta'', s'', \pi_2(\cdot|s''))$ are continuous functions, so for any $\pi_2^+ \in \Pi_2^+$ and $\pi_2^- \in \Pi_2^-$, there exists $\alpha \in (0, 1)$ so that $\alpha \pi_2^+ + (1 - \alpha)\pi_2^- \in \Pi_2^0$. (Note that π_2^+ and π_2^- must be supported on A_s^{BR} after every signal s, so the same must hold for the mixture $\alpha \pi_2^+ + (1 - \alpha)\pi_2^-$. Thus, this mixture also belongs to $\Pi_2^\bullet$.) If only Π_2^+ is nonempty and $\theta' \succsim_{s'} \theta''$, then s' is rationally strictly dominant for both θ' and θ''. If only Π_2^- is nonempty, then we can have $\theta'' \succsim_{s'} \theta'$ only when s' is never a weak best response for θ' against any $\pi_2 \in \Pi_2^\bullet$. □

A.2. Proof of Proposition 2

Proof. Let π^* be a uRCE. We construct a path-equivalent RCE, π° as follows. Set $\pi_1^\circ = \pi_1^*$ and set $\pi_2^\circ(\cdot \mid s) = \pi_2^*(\cdot \mid s)$ for every on-path signal s. At each off-path signal s where $\tilde{J}(s, \pi^*) \neq \varnothing$, let $\pi_2^\circ(\cdot \mid s)$ prescribe some best response to a belief in $\tilde{P}(s, \pi^*)$. At each off-path signal s where $\tilde{J}(s, \pi^*) = \varnothing$, let $\pi_2^\circ(\cdot \mid s)$ prescribe some best response to a belief in $\Delta(\Theta_s)$.

In this strategy profile, the receiver's play is a best response to rationality-compatible beliefs after every off-path s by construction, and because the sender's play is the same as before the receiver is still playing best responses to on-path signals.

Because the on-path play of the receivers did not change, no sender type wishes to deviate to any on-path signal. Now we check that no sender type wishes to deviate to any off-path signal. Consider first off-path s where $\tilde{J}(s, \pi^*) \neq \varnothing$. Here we have $\tilde{J}(s, \pi^*) \subseteq \Theta_s$, which implies that $\tilde{P}(s, \pi^*) \subseteq \hat{P}(s)$. By the definition of uRCE, $\pi_2^\circ(\cdot \mid s)$ must deter every type from deviating to such s. Finally, no sender type wishes to deviate to any s where $\tilde{J}(s, \pi^*) = \varnothing$, by the definition of equilibrium dominance. □

A.3. Proof of Proposition 3

Proof. Fix a π_1 with $\pi_1(s|\theta'') > 0$ but $\pi_1(s|\theta') < 1$. Because the space of rational receiver strategies $\Pi_2^\bullet$ is convex, it suffices to show there is no receiver strategy $\pi_2 \in \Pi_2^\bullet$ such that π_1 is a best response to π_2 in the ex-ante strategic form. If π_1 is an ex-ante best response, then it needs to be at least weakly optimal for type θ'' to play s against π_2. By $\theta' \succsim_s \theta''$, this implies s is strictly optimal for type θ'. This shows π_1 is not a best response to π_2, as the sender can increase her ex-ante expected payoffs by playing s with probability 1 when her type is θ'. □

A.4. Proof of Proposition 4

Proof. Suppose π^* does not pass the Intuitive Criterion. Then there exists a type θ and a signal s' such that

$$u_1(\theta; \pi^*) < \min_{a \in \mathrm{BR}(\Delta(\tilde{J}(s', \pi^*)), s)} u_1(\theta, s', a).$$

If π^* were an RCE, then we would have $\pi_2^*(\cdot|s') \in \Delta(\mathrm{BR}(\tilde{P}(s, \pi^*), s))$. Since $\tilde{P}(s, \pi^*) \subseteq \Delta(\tilde{J}(s', \pi^*))$, we have

$$u_1(\theta; \pi^*) < u_1(\theta, s', \pi_2^*(\cdot|s')).$$

This means π^* is not a Nash equilibrium, contradiction. □

A.5. Proof of Proposition 5

Proof. To show (a), note first that if $D(\theta'', s'; \pi^*) \cup D^\circ(\theta'', s'; \pi^*) = \varnothing$ the conclusion holds vacuously. If $D(\theta'', s'; \pi^*) \cup D^\circ(\theta'', s'; \pi^*)$is not empty, take any $\alpha' \in D(\theta'', s'; \pi^*) \cup D^\circ(\theta'', s'; \pi^*)$ and define $\pi_2' \in \Pi_2^\bullet$ by $\pi_2'(\cdot|s') = \alpha'$, $\pi_2'(\cdot|s) = \pi_2^*(\cdot|s)$ for $s \neq s'$. Then

$$u_1(\theta''; \pi^*) = \max_{s \neq s'} u_1(\theta'', s, \pi_2'(\cdot|s)) \leq u_1(\theta'', s', \pi_2'(\cdot|s')) = u_1(\theta'', s', \alpha'),$$

and when $\theta' \succsim_{s'} \theta''$, this implies that

$$u_1(\theta'; \pi^*) = \max_{s \neq s'} u_1(\theta', s, \pi_2'(\cdot|s)) < u_1(\theta', s', \pi_2'(\cdot|s')) = u_1(\theta', s, \alpha').$$

Hence $\alpha' \in D(\theta', s'; \pi^*)$.

To show (b), suppose π^* is a divine equilibrium. Then it is a Nash equilibrium, and furthermore for any off-path signal s' where $\theta' \succsim_{s'} \theta''$, Proposition 5(a) implies that

$$D(\theta'', s'; \pi^*) \cup D^\circ(\theta'', s'; \pi^*) \subseteq D(\theta', s'; \pi^*).$$

Since π^* is a divine equilibrium, $\pi_2^*(\cdot|s')$ must then best respond to some belief $p \in \Delta(\Theta)$ with $\dfrac{p(\theta'')}{p(\theta')} \leq \dfrac{\lambda(\theta'')}{\lambda(\theta')}$. Considering all (θ', θ'') pairs, we see that in a divine equilibrium $\pi_2^*(\cdot|s')$ best responds to some belief in

$$\bigcap_{(\theta', \theta'') \text{ s.t. } \theta' \succsim_{s'} \theta''} P_{\theta' \rhd \theta''}.$$

At the same time, in every divine equilibrium, belief after off-path s' puts zero probability on equilibrium-dominated types, meaning $\pi_2^*(\cdot \mid s')$ best responds $\Delta(\tilde{J}(s', \pi^*))$. This shows π^* is an RCE. □

A.6. Proof of Proposition 6

Proof. Consider a uRCE π^*. For every off-path s, perform the following modifications on $\pi_2^*(\cdot|s)$: if the first-round application of the NWBR procedure would have deleted every type, then do not modify $\pi_2^*(\cdot|s)$. Otherwise, find some θ_s not deleted by the iterated NWBR procedure, then change $\pi_2^*(\cdot|s)$ to some action in $\mathrm{BR}(\{\theta_s\}, s)$, i.e. a best response to the belief putting probability 1 on θ_s.

This modified strategy profile passes the NWBR test. We now establish that it remains a uRCE by checking that for those off-path s where $\pi_2^*(\cdot|s)$ was modified, the modified version is still a best response to $\hat{P}(s)$. (By uniformity, this would ensure that the modified receiver play continues to deter every type from deviating to s.)

Type θ_s satisfies $\theta_s \in \Theta_s$. Otherwise, $D^{\circ}(\theta_s, s; \pi^*) = \varnothing$ and θ_s would have been deleted by NWBR in the first round. Now it suffices to argue there is no θ' such that $\theta' \succsim_s \theta_s$, which implies the belief putting probability 1 on θ_s is in $\hat{P}(s)$. If there were such θ', by Proposition 5(a) we would have $D^{\circ}(\theta_s, s; \pi^*) \subseteq D(\theta', s; \pi^*)$, so θ_s should have been deleted by NWBR in the first round, contradicting the fact that θ_s survives all iterations of the NWBR procedure. □

A.7. Proof of Corollary 1

Proof. This follows from Proposition 6 because every NWBR equilibrium is a universally divine equilibrium. □

A.8. Proof of Lemma 2

Proof. Here are three lemmas from Fudenberg and Levine (2006):

FL06 Lemma A.1: Suppose $\{X_k\}$ is a sequence of i.i.d. Bernoulli random variables with $\mathbb{E}[X_k] = \mu$, and define for each n the random variable

$$S_n := \frac{|\sum_{k=1}^{n}(X_k - \mu)|}{n}.$$

Then for any $\underline{n}, \bar{n} \in \mathbb{N}$,

$$\mathbb{P}\left[\max_{\underline{n} \le n \le \bar{n}} S_n > \epsilon\right] \le \frac{2^7}{3} \cdot \frac{1}{\underline{n}} \cdot \frac{\mu}{\epsilon^4}.$$

FL06 Lemma A.2: For all $\epsilon, \epsilon' > 0$, there is an $N > 0$ so that for all δ, γ, g, π, signal s and action $a \in A$,

$$\psi_{\theta}^{\pi_2;(g,\delta,\gamma)}\left\{y_\theta : |\hat{\pi}_2(a|s; y_\theta) - \pi_2(a|s)| > \epsilon, \#(s|y_\theta) > N\right\} < \epsilon'.$$

(Here, $\hat{\pi}_2(a|s; y_\theta)$ is the empirical frequency of receiver playing a after signal m in history y_θ, that is to say $\hat{\pi}_2(a|s; y_\theta) = \#((a, s), y_\theta)/\#(s, y_\theta)$.)

FL06 Lemma A.4: For all $\epsilon, \epsilon' > 0$ and $\delta < 1$, there exists N such that for all π, g, and γ, we get

$$\psi_{\theta}^{\pi_2;(g,\delta,\gamma)}\{y_\theta \notin Y_\theta(\epsilon), \#(\sigma_\theta(y_\theta), y_\theta) > N\} \le \epsilon'$$

where $Y_\theta(\epsilon) \subseteq Y_\theta$ are those histories y_θ where

$$\max_{s \in S} u_1(\theta, s|y_\theta) \le u_1(\sigma_\theta(y_\theta)|y_\theta) + \epsilon,$$

that is, type θ is playing a myopic ϵ best response according to posterior belief after history y_θ.

Now we proceed with our argument.

Since π^* is strict on-path, there exist $\xi_1, \xi_2 > 0$ such that whenever π_2 satisfies $|\pi_2(a|s) - \pi_2^*(a|s)| \le \xi_1$ for every on-path s and action a, while for every off-path s we have $\pi_2(\tilde{A}(s)|s) \ge 1 - \xi_1$, then for each type θ we get

$$u_1(\theta, \pi_1^*(\theta), \pi_2) > \xi_2 + \max_{s \ne \pi_1^*(\theta)} u_1(\theta, s, \pi_R).$$

That is, if receiver plays ξ_1-close to π^* on-path and ξ_1-close to $\tilde{A}(s)$ off-path, then for every type of sender, playing the prescribed equilibrium signal is strictly better than any other signal by at least $\xi_2 > 0$.

Following Fudenberg and Levine (2006), consider a prior g_1 such that whenever sender has fewer than $\underline{n} := 2^{11}/\xi_1^4$ observations of playing signal s, her belief as to receiver's probability of taking action a after signal s differs from $\pi_1^*(a|s)$ by no more than ξ_1 if s is on-path, while her belief as to the probability that receiver strategy assigns to $\tilde{A}(s)$ is at least $1 - \xi$ if s is off-path. Also, let $\epsilon_{\text{off}} := \xi_1/2$.

Now let $\delta \in (0, 1)$ and $0 < \epsilon < \epsilon_{\text{off}}$ be given. We construct $\gamma(\delta, \epsilon)$ satisfying the conclusion of the lemma.

To do this, in FL06 Lemma A.4 put $\epsilon = \xi_2$ and $\epsilon' = \epsilon/6$, to obtain a $N_1(\epsilon)$. Next, in FL06 Lemma A.2 put $\epsilon = \xi_1/2$, $\epsilon' = \epsilon/6$, to obtain $N_2(\epsilon)$. Let $N(\epsilon) := N_1(\epsilon) \vee N_2(\epsilon)$. There are 5 classes of exceptional histories for type θ that can lead to playing some signal $\hat{s}$ other than the one prescribed by the equilibrium strategy, $s^* := \pi_1^*(\theta)$.

Exception 1: θ has played $\hat{s}$ fewer than $N(\epsilon)$ times before, that is $\sigma_\theta(y_\theta) = \hat{s}$ but $\#(\hat{s}, y_\theta) < N(\epsilon)$. Such histories can be made to have mass no larger than $\epsilon/6$ by taking $\gamma(\delta, \epsilon)$ large enough.

Exception 2: y_θ is in the exceptional set described in FL06 Lemma A.4. But by choice of $N(\epsilon) \geq N_1(\epsilon)$, we know that

$$\psi_\theta^{\pi_2;(g,\delta,\gamma)}\{y_\theta \notin Y_\theta(\xi_2), \#(\sigma_\theta(y_\theta), y_\theta) > N(\epsilon)\} \leq \epsilon/6.$$

Exception 3: θ has played $\hat{s}$ more than $N(\epsilon)$ times, but has a misleading sample. By FL93 Lemma A.2,

$$\psi_\theta^{\pi_2;(g,\delta,\gamma)}\left\{y_\theta : |\hat{\pi}_2(a|\hat{s}; y_\theta) - \pi_2(a|\hat{s})| > \xi_1/2, \#(\hat{s}|y_\theta) > N(\epsilon)\right\} < \epsilon/6.$$

Since we have chosen $\pi \in B_2^{\text{on}}(\pi^*, \epsilon) \cap B_2^{\text{off}}(\pi^*, \epsilon_{\text{off}})$, we know π_2 differs from π_2^* by no more than $\epsilon_{\text{off}} = \xi_1/2$ after every on-path signal, and puts no more weight than $\xi_1/2$ on actions not in $\tilde{A}(s)$ after off-path signal s. So in particular,

$$\psi_\theta^{\pi_2;(g,\delta,\gamma)}\left\{ y_\theta : \begin{array}{c} |\hat{\pi}_2(a|\hat{s}; y_\theta) - \pi_2^*(a|\hat{s})| > \xi_1 \text{ if } \hat{s} \text{ on-path, or} \\ \hat{\pi}_2(\tilde{A}(\hat{s})|\hat{s}) < 1 - \xi_1 \text{ if } \hat{s} \text{ off-path} \\ \#(\hat{s}|y_\theta) > N(\epsilon) \end{array} \right\} < \epsilon/6.$$

Exception 4: θ has played the equilibrium signal s^* more than $N(\epsilon)$ times, but has a misleading sample. As before, we get

$$\psi_\theta^{\pi_2;(g,\delta,\gamma)}\left\{y_\theta : |\hat{\pi}_2(a|s^*; y_\theta) - \pi_2^*(a|s^*)| > \xi_1, \#(s^*|y_\theta) > N(\epsilon)\right\} < \epsilon/6.$$

Exception 5: θ has played the equilibrium signal s^* between $\underline{n}$ and $N(\epsilon)$ times, but has a misleading sample. Let $X_k \in \{0, 1\}$ denote whether θ sees the equilibrium response $\pi_2^*(s^*)$ the k-th time she plays s^* ($X_k = 0$) or whether she sees instead a different response ($X_k = 1$). As in FL06 Lemma A.1, define

$$S_n := \frac{|\sum_{k=1}^n (X_k - \mu)|}{n}$$

where $\mu = 1 - \pi_2(\pi_2^*(s^*)|s^*) < \epsilon$ since s^* is an on-path signal in π^*.

The probability that the fraction of responses other than $\pi_1^*(s^*)$ exceeds ξ_1 between the$\underline{n}$-th time and $N(\epsilon)$-th time that θ plays s^* is bounded above by FL06 Lemma A.1,

$$\begin{aligned} \mathbb{P}\left[\max_{\underline{n} \leq n \leq N(\epsilon)} S_n > \xi_1/2\right] &\leq \frac{2^7}{3} \cdot \frac{1}{\underline{n}} \cdot \frac{\mu}{(\xi_1/2)^4} \\ &\leq \frac{1}{3} \cdot \mu \text{ (by choice of } \underline{n}) \\ &\leq \epsilon_1/3. \end{aligned}$$

Finally, at a history y_θ that does not belong to those exceptions, we must have $\sigma_\theta(y_\theta) = m^*$. This is because y_θ is not in exception 1, so θ has played $\sigma_\theta(y_\theta)$ at least $N(\epsilon)$ times before, and it is not in exception 2, so $\sigma_\theta(y_\theta)$ is a ξ_2 myopic best response to current beliefs. Yet the empirical frequency for response after signal $\sigma_\theta(y_\theta)$ is no more than ξ_1 away from $\pi_2^*(\sigma_\theta(y_\theta))$ as y_θ is not in exception 3. Since the prior is Dirichlet and also has this property, this means the current posterior belief about response after signal $\sigma_\theta(y_\theta)$ also has this property. If $\#(s^*, y_\theta) > \underline{n}$, then y_θ not being in exceptions 4 or 5 implies belief as to response after signal s^* is also no more than ξ_1 away from $\pi_2^*(s^*)$, while if $\#(s^*, y_\theta) < \underline{n}$ then choice of prior implies the same. In short, beliefs on both responses after s^* and responses after $\sigma_\theta(y_\theta)$ are no more than ξ_1 away from their π_2^* counterparts. But in that case, no signal other than s^* can be an ξ_2 best response. □

A.9. Proof of Lemma 3

Proof. For each $\xi > 0$, consider the approximation to $P_{\theta' \rhd \theta''}$,

$$P_{\theta' \rhd \theta''}^\xi := \left\{p \in \Delta(\Theta) : \frac{p(\theta'')}{p(\theta')} \leq (1 + \xi)\frac{\lambda(\theta'')}{\lambda(\theta')}\right\}$$

and hence the approximation to $\hat{P}(s)$,

$$\hat{P}_\xi(s) := \Delta(\Theta_{s'}) \bigcap \left\{P_{\theta' \rhd \theta''}^\xi : \theta' \succsim_{s'} \theta''\right\}.$$

Since the BR correspondence has a closed graph, there is an $\xi > 0$ such that $\text{BR}(\hat{P}_\xi(s), s) = \text{BR}(\hat{P}(s), s)$.

Take some such ξ. Next we will choose a series of constants.

- Pick $0 < h < 1$ such that $\frac{1-h}{1+h} > (1-\xi)^{1/3}$.
- Pick $G > 0$ such that for every $\theta \in \Theta$, $1/(h^2 \cdot G \cdot (1-h) \cdot \lambda(\theta)) < \epsilon/(4 \cdot |S| \cdot |\Theta|^2)$.
- For each θ, construct a Dirichlet prior on S_θ with parameters $\alpha(\theta, s) \geq 0$. Pick Dirichlet prior parameters $\alpha(\theta, s) \geq 0$ so that whenever $\theta \succsim_s \theta'$, we have
$$\alpha(\theta, s) - \alpha(\theta', s) > (\sqrt{(4 \cdot |S| \cdot |\Theta|^2)/\epsilon} + 1) \cdot G. \tag{2}$$
In the event that $\theta \succsim_s \theta'$ and $\theta' \succsim_s \theta$, put $\alpha(\theta, s) = \alpha(\theta', s)$.
- Pick $\underline{N} \in \mathbb{N}$ so that for any $N > \underline{N}$, $\theta, \theta' \in \Theta$, we have
$$\mathbb{P}[(1-h) \cdot N \cdot \lambda(\theta) \leq \text{Binom}(N, \lambda(\theta)) \leq (1+h) \cdot N \cdot \lambda(\theta)] > 1 - \frac{\epsilon}{4 \cdot |\Theta|}$$
and
$$\frac{(1-h) \cdot N \cdot \lambda(\theta')}{(1+h) \cdot N \cdot \lambda(\theta) + \max_\theta \sum_{s \in S} \alpha(\theta, s)} > (1-\xi)^{1/3} \frac{\lambda(\theta')}{\lambda(\theta)}.$$
- Pick $\underline{\gamma} \in (0, 1)$ such that $1 - (\underline{\gamma})^{\underline{N}+1} < \epsilon/4$.

Suppose the receiver's prior over the strategy of typeθ is Dirichlet with parameters $(\alpha(\theta, s))_{s \in S}$. We claim that the conclusion of the lemma holds.

Fix some strategy $\pi_1 \in C$. Write $\#(\theta|y_2)$ for the number of times the sender has been of θ type in history y_2, while $\#(\theta, s|y_2)$ counts the number of times type θ has sent signal s in history y_2. Put $\psi_2 = \psi_2^{\pi_1;(g,\delta,\gamma)}$ and write $E \subseteq Y_2$ for those receiver histories with length at least $\underline{N}$ satisfying
$$(1-h) \cdot N \cdot \lambda(\theta) \leq \#(\theta|y_2) \leq (1+h) \cdot N \cdot \lambda(\theta)$$
for every $\theta \in \Theta$. By the choice of $\underline{N}$ and $\underline{\gamma}$, whenever $\gamma > \underline{\gamma}$ we have $\psi(E) \geq 1 - \epsilon/2$. We now show that given E, the conditional probability that the receiver's posterior belief after every off-equilibrium signal s lies in $\hat{P}_\xi(s)$ is at least $1 - \epsilon/2$. To do this, fix signal s and two types with $\theta \succsim_s \theta'$.

If s is strictly dominated for both θ and θ', then according to the receivers' Dirichlet prior, θ and θ' each sends s with zero probability. Since $\pi \in \Pi_1^\bullet$, we have $\pi_1(s|\theta) = \pi_1(s|\theta') = 0$. So after every positive-probability history, receiver's belief falls in $\hat{P}_\xi(s)$ as it puts zero probability on the s-sender being θ or θ'. Henceforth we only consider the case where s is not strictly dominated for both.

After history y_2, the receiver's updated posterior likelihood ratio for types θ and θ' upon seeing signal s is
$$\begin{aligned}&\frac{\lambda(\theta)}{\lambda(\theta')} \cdot \left(\frac{\alpha(\theta, s) + \#(\theta, s|y_2)}{\#(\theta|y_2) + \sum_{s \in S} \alpha(\theta, s)} \Big/ \frac{\alpha(\theta', s) + \#(\theta', s|y_2)}{\#(\theta'|y_2) + \sum_{s \in S} \alpha(\theta', s)} \right) \\ &= \frac{\lambda(\theta)}{\lambda(\theta')} \cdot \frac{\alpha(\theta, s) + \#(\theta, s|y_2)}{\alpha(\theta', s) + \#(\theta', s|y_2)} \cdot \frac{\#(\theta'|y_2) + \sum_{s \in S} \alpha(\theta', s)}{\#(\theta|y_2) + \sum_{s \in S} \alpha(\theta, s)}.\end{aligned}$$

Since we have $\#(\theta'|y_2) \geq (1-h) \cdot N \cdot \lambda(\theta')$ while $\#(\theta|y_2) \leq (1+h) \cdot N \cdot \lambda(\theta)$, we get
$$\frac{\#(\theta'|y_2) + \sum_{s \in S} \alpha(\theta', s)}{\#(\theta|y_2) + \sum_{s \in S} \alpha(\theta, s)} \geq \frac{(1-h) \cdot N \cdot \lambda(\theta')}{(1+h) \cdot N \cdot \lambda(\theta) + \sum_{s \in S} \alpha(\theta, s)} > (1-\xi)^{1/3} \cdot \frac{\lambda(\theta')}{\lambda(\theta)}.$$

If s is strictly dominant for both θ and θ', then $\pi_1 \in \Pi_1^\bullet$ means that $\pi_1(s|\theta) = \pi_1(s|\theta') = 1$. In this case, $\#(\theta, s|y_2) = \#(\theta|y_2)$ and $\#(\theta', s|y_2) = \#(\theta'|y_2)$. Since $\#(\theta|y_2) \geq (1-h) \cdot N \cdot \lambda(\theta)$, $\#(\theta'|y_2) \leq (1+h) \cdot N \cdot \lambda(\theta')$, we have:
$$\frac{\alpha(\theta, s) + \#(\theta, s|y_2)}{\alpha(\theta', s) + \#(\theta', s|y_2)} \geq \frac{(1-h) \cdot N \cdot \lambda(\theta)}{\sum_{s \in S} \alpha(\theta', s) + (1+h) \cdot N \cdot \lambda(\theta')} \geq (1-\xi)^{1/3} \frac{\lambda(\theta)}{\lambda(\theta')}.$$

This shows the product is no smaller than $(1-\xi)^{2/3} \frac{\lambda(\theta)}{\lambda(\theta')}$, so receiver believes in $P^\xi_{\theta \rhd \theta'}$ after every history in E.

Now we analyze the term $\frac{\alpha(\theta,s)+\#(\theta,s|y_2)}{\alpha(\theta',s)+\#(\theta',s|y_2)}$ for the case where s is not strictly dominant for both θ and θ'. We consider two cases, depending on whether N is "large enough" so that the compatible type θ experiments enough on average in a receiver history of length N under sender strategy π_1.

Case A: $\pi_1(s|\theta) \cdot N < G$. In this case, since $\pi \in C$ and $\theta \succsim_s \theta'$, we must also have $\pi_1(s|\theta') \cdot N < G$. Then $\#(\theta', s|y_2)$ is distributed as a binomial random variable with mean smaller than G, hence standard deviation smaller than $\sqrt{G}$. By Chebyshev's inequality, the probability that it exceeds $(\sqrt{(4 \cdot |S| \cdot |\Theta|^2)/\epsilon} + 1) \cdot G$ is no larger than
$$\frac{1}{G \cdot (4 \cdot |S| \cdot |\Theta|^2)/\epsilon} < \frac{\epsilon}{4|S| \cdot |\Theta|^2}.$$

But in any history y_2 where $\#(\theta', s|y_R)$ does not exceed this number, we would have

$$\alpha(\theta', s) + \#(\theta', s|y_2) \leq \alpha(\theta, s) \leq \alpha(\theta, s) + \#(\theta, s|y_2)$$

by choice of the difference between prior parameters $\alpha(\theta', s)$ and $\alpha(\theta, s)$. Therefore $\frac{\alpha(\theta,s)+\#(\theta,s|y_2)}{\alpha(\theta',s)+\#(\theta',s|y_2)} \geq 1$. In summary, under Case A, there is probability no smaller than $1 - \frac{\epsilon}{4|S|\cdot|\Theta|^2}$ that $\frac{\alpha(\theta,s)+\#(\theta,s|y_2)}{\alpha(\theta',s)+\#(\theta',s|y_2)} \geq 1$.

Case B: $\pi_1(s|\theta) \cdot N \geq G$. In this case, we can bound the probability that

$$\#(\theta, s|y_2)/\#(\theta', s|y_2) \leq \frac{\lambda(\theta)}{\lambda(\theta')} \cdot (\frac{1-h}{1+h})^2.$$

Let $p := \pi_1(s|\theta)$. Given that $\#(\theta|y_2) \geq (1-h) \cdot N \cdot \lambda(\theta)$, the distribution of $\#(\theta, s|y_2)$ first order stochastically dominates $\text{Binom}((1-h) \cdot N \cdot \lambda(\theta), p)$.

On the other hand, given that $\#(\theta|y_2) \leq (1+h) \cdot N \cdot \lambda(\theta')$ and furthermore $\pi_1(s|\theta') \leq \pi_1(s|\theta) = p$, the distribution of $\#(\theta', s|y_1)$ is first order stochastically dominated by $\text{Binom}((1+h) \cdot N \cdot \lambda(\theta'), p)$.

The first distribution has mean $(1-h) \cdot N \cdot \lambda(\theta) \cdot p$ with standard deviation no larger than $\sqrt{(1-h) \cdot N \cdot \lambda(\theta) \cdot p}$. Thus

$$\begin{aligned}&\mathbb{P}\left[\text{Binom}((1-h) \cdot N \cdot \lambda(\theta), p) < (1-h) \cdot (1-h) \cdot N \cdot \lambda(\theta) \cdot p\right] \\ &< 1/(h \cdot \sqrt{p(1-h)N\lambda(\theta)})^2 \leq 1/(h \cdot \sqrt{G \cdot (1-h) \cdot \lambda(\theta)})^2 < \epsilon/(4 \cdot |S| \cdot |\Theta|^2)\end{aligned}$$

where we used the fact that $pN \geq G$ in the second-to-last inequality, while the choice of G ensured the final inequality.

At the same time, the second distribution has mean $(1+h) \cdot N \cdot \lambda(\theta') \cdot p$ with standard deviation no larger than $\sqrt{(1+h) \cdot N \cdot \lambda(\theta') \cdot p}$, so

$$\begin{aligned}&\mathbb{P}\left[\text{Binom}((1+h) \cdot N \cdot \lambda(\theta'), p) > (1+h) \cdot (1+h) \cdot N \cdot \lambda(\theta') \cdot p\right] \\ &< 1/(h \cdot \sqrt{p(1+h)N\lambda(\theta')})^2 \leq 1/(h \cdot \sqrt{G \cdot (1+h) \cdot \lambda(\theta')})^2 < \epsilon/(4 \cdot |S| \cdot |\Theta|^2)\end{aligned}$$

by the same arguments. Combining the bounds on these two binomial random variables,

$$\mathbb{P}\left[\frac{\text{Binom}((1-h) \cdot N \cdot \lambda(\theta), p)}{\text{Binom}((1+h) \cdot N \cdot \lambda(\theta'), p)} \leq \frac{\lambda(\theta)}{\lambda(\theta')} \cdot (\frac{1-h}{1+h})^2\right] < \epsilon/(2 \cdot |S| \cdot |\Theta|^2).$$

Via stochastic dominance, this shows *a fortiori*

$$\mathbb{P}\left[\#(\theta, s|y_2)/\#(\theta', s|y_2) \leq \frac{\lambda(\theta)}{\lambda(\theta')} \cdot (\frac{1-h}{1+h})^2\right] < \epsilon/(2 \cdot |S| \cdot |\Theta|^2).$$

Therefore, for any s, θ, θ' such that $\theta \succsim_s \theta'$,

$$\psi\left(y_2 : \frac{\alpha(\theta, s) + \#(\theta, s|y_2)}{\alpha(\theta', s) + \#(\theta', s|y_2)} \geq \frac{\lambda(\theta)}{\lambda(\theta')} \cdot (\frac{1-h}{1+h})^2 \mid E\right) \geq 1 - \epsilon/(2 \cdot |S| \cdot |\Theta|^2).$$

This concludes case B.

In either case, at a history y_2 with $(1-h) \cdot N \cdot \lambda(\theta) \leq \#(\theta|y_2) \leq (1+h) \cdot N \cdot \lambda(\theta)$ for every θ, for every pair θ, θ' such that $\theta \succsim_s \theta'$, we get $\frac{\alpha(\theta,s)+\#(\theta,s|y_2)}{\alpha(\theta',s)+\#(\theta',s|y_2)} \geq \frac{\lambda(\theta)}{\lambda(\theta')} \cdot (\frac{1-h}{1+h})^2$ with probability at least $1 - \epsilon/(2 \cdot |S| \cdot |\Theta|^2)$.

But at any history y_2 where this happens, the receiver's posterior likelihood ratio for types θ and θ' after signal s satisfies

$$\begin{aligned}&\frac{\lambda(\theta)}{\lambda(\theta')} \cdot \frac{\alpha(\theta, s) + \#(\theta, s|y_2)}{\alpha(\theta', s) + \#(\theta', s|y_2)} \cdot \frac{\#(\theta'|y_2) + \sum_{s \in S} \alpha(\theta', s)}{\#(\theta|y_2) + \sum_{s \in S} \alpha(\theta, s)} \\ &\geq \frac{\lambda(\theta)}{\lambda(\theta')} \cdot \frac{\lambda(\theta)}{\lambda(\theta')} \cdot \left(\frac{1-h}{1+h}\right)^2 \cdot (1-\xi)^{1/3} \cdot \frac{\lambda(\theta')}{\lambda(\theta)} \\ &\geq \frac{\lambda(\theta)}{\lambda(\theta')} \cdot (1-\xi)^{2/3} \cdot (1-\xi)^{1/3} \geq \frac{\lambda(\theta)}{\lambda(\theta')} \cdot (1-\xi).\end{aligned}$$

As there are at most $|\Theta|^2$ such pairs for each signal s and $|S|$ total signals,

$$\psi\left(y_2 : \begin{array}{c}\frac{\lambda(\theta)}{\lambda(\theta')} \cdot \frac{\alpha(\theta,s)+\#(\theta,s|y_2)}{\alpha(\theta',s)+\#(\theta',s|y_2)} \cdot \frac{\#(\theta'|y_2)+\sum_{s \in S} \alpha(\theta',s)}{\#(\theta|y_2)+\sum_{s \in S} \alpha(\theta,s)} \\ \geq \frac{\lambda(\theta)}{\lambda(\theta')} \cdot (1-\xi)\end{array} \ \forall s, \theta \succsim_s \theta' \, |E\right) \geq 1 - \epsilon/2$$

as claimed. As the event E has ψ-probability no smaller than $1 - \epsilon/2$, there is ψ probability at least $1 - \epsilon$ that receiver's posterior belief is in $\hat{P}_\xi(s)$ after every off-path s. □

A.10. Proof of Lemma 4

Proof. Since π^* is on-path strict for the receiver, there exists some $\xi > 0$ such that for every on-path signal s and every belief $p \in \Delta(\Theta)$ with

$$|p(\theta) - p(\theta; s, \pi^*)| < \xi, \ \forall \theta \in \Theta \tag{3}$$

(where $p(\cdot; s, \pi^*)$ is the Bayesian belief after on-path signal s induced by the equilibrium π^*), we have $\text{BR}(p, s) = \{\pi_2^*(s)\}$. For each s, we show that there is a large enough $N(s, \epsilon)$ and small enough $\zeta(s)$ so that when receiver observes history y_2 generated by any $\pi \in B_{\text{on}}(\pi^*, \epsilon')$ with $\epsilon' < \zeta(s)/4$ and length at least $N(s, \epsilon)$, there is probability at least $1 - \frac{\epsilon}{2|S|}$ that receiver's posterior belief satisfies (3). Hence, conditional on having a history length of at least $N(s, \epsilon)$, there is $1 - \frac{\epsilon}{2|S|}$ chance that receiver will play as in π_2^* after s. By taking the maximum $N^*(\epsilon) := \max_s(N(s, \epsilon_1))$ and minimum $\epsilon_1 := \min_s \zeta(s)$, we see that whenever history is length $N^*(\epsilon)$ or more, and $\pi \in B_{\text{on}}(\pi^*, \epsilon')$ with $\epsilon' < \epsilon_1$, there is at least $1 - \epsilon/2$ chance that the receiver's strategy matches π_2^* after every on-path signal. Since we can pick $\gamma(\epsilon)$ large enough that $1 - \epsilon/2$ measure of the receiver population is age $N^*(\epsilon)$ or older, we are done.

To construct $N(s, \epsilon)$ and $\zeta(s)$, let $\Lambda(s) := \lambda\{\theta : \pi_1^*(s|\theta) = 1\}$. Find small enough $\zeta(s) \in (0, 1)$ so that:

- $|\frac{\lambda(\theta)}{\Lambda(s)\cdot(1-\zeta(s))} - \frac{\lambda(\theta)}{\Lambda(s)}| < \xi$
- $|\frac{\lambda(\theta)\cdot(1-\zeta(s))}{\Lambda(s)+(1-\Lambda(s))\cdot\zeta(s)} - \frac{\lambda(\theta)}{\Lambda(s)}| < \xi$
- $\frac{\zeta(s)}{1-\zeta(s)} \cdot \frac{\lambda(\theta)}{\Lambda(s)} < \xi$

for every $\theta \in \Theta$. After a history y_2, the receiver's posterior belief as to the type of sender who sends signal s satisfies

$$p(\theta|s; y_2) \propto \lambda(\theta) \cdot \frac{\#(\theta, s|y_2) + \alpha(\theta, s)}{\#(\theta|y_2) + A(\theta)},$$

where $\alpha(\theta, s)$ is the Dirichlet prior parameter on signal s for type θ and $A(\theta) := \sum_{s\in S} \alpha(\theta, s)$. By the law of large numbers, for long enough history length, we can ensure that if $\pi_1(s|\theta) > 1 - \frac{\zeta(s)}{4}$, then

$$\frac{\#(\theta, s|y_2) + \alpha(\theta, s)}{\#(\theta|y_2) + A(\theta)} \geq 1 - \zeta(s)$$

with probability at least $1 - \frac{\epsilon}{2|S|^2}$, while if $\pi_1(s|\theta) < \zeta(s)/4$, then

$$\frac{\#(\theta, s|y_2) + \alpha(\theta, s)}{\#(\theta|y_2) + A(\theta)} < \zeta(s)$$

with probability at least $1 - \frac{\epsilon}{2|S|^2}$. Moreover there is some $N(s, \epsilon)$ so that there is probability at least $1 - \frac{\epsilon}{2|S|}$ that a history y_2 with length at least $N(s, \epsilon)$ satisfies above for all θ. But at such a history, for any θ such that $\pi_1^*(s|\theta) = 1$,

$$p(\theta|s; y_2) \geq \frac{\lambda(\theta) \cdot (1 - \zeta(s))}{\Lambda(s) + (1 - \Lambda(s)) \cdot \zeta(s)}$$

and

$$p(\theta|s; y_2) \leq \frac{\lambda(\theta)}{\Lambda(s) \cdot (1 - \zeta(s))},$$

while for some θ such that $\pi_1^*(s|\theta) = 0$,

$$p(\theta|s; y_2) \leq \frac{\zeta(s)}{1 - \zeta(s)} \cdot \frac{\lambda(\theta)}{\Lambda(s)}.$$

Therefore the belief $p(\cdot|s; y_R)$ is no more than ξ away from $p(\theta; s, \pi^*)$, as desired. □

A.11. Proof of Theorem 2

Proof. We will construct a regular prior g. We will then show that for every $0 < \delta < 1$, there exists convex and compact sets of strategy profiles $E_j \subseteq \Pi^\bullet$ with $E_j \downarrow E_* \subseteq B_1^{\text{on}}(\pi^*, 0) \cap B_2^{\text{on}}(\pi^*, 0)$ and a corresponding sequence of survival probabilities $\gamma_j \to 1$ so that $(\mathscr{R}_1^{g,\delta,\gamma_j}[\pi_2], \mathscr{R}_2^{g,\delta,\gamma_j}[\pi_1]) \in E_j$ whenever $\pi \in E_j$. We proved in Fudenberg and He (2018) that $\mathscr{R}_1$ and $\mathscr{R}_2$ are continuous maps, so a fixed point theorem implies that for each j, some strategy profile in E_j is a steady state profile under parameters (g, δ, γ_j). Any convergent subsequence of these j-indexed steady state profiles has a limit in E_*, so this limit agrees with π^* on path. This shows that for every δ there is a δ-stable strategy profile path-equivalent to π^*, so there is a rationally patiently stable strategy profile with the same property.

Step 1: Constructing g and some thresholds.

Since π^* induces a unique optimal signal for each sender type, by Lemma 2 find a regular sender prior g_1, $0 < \epsilon_{\text{off}} < 0$, and a function $\gamma_{\text{LM1}}(\delta, \epsilon)$.

In Lemma 3, substitute $\epsilon = \epsilon_{\text{off}}$ to find a regular receiver prior g_2 and $0 < \underline{\gamma}_{\text{LM2}} < 1$.

Finally, in Lemma 4 let g_2 be as constructed above to find $\epsilon_{\text{LM3}} > 0$ and a function $\gamma_{\text{LM3}}(\epsilon)$.

Step 2: Constructing the sets E_j.

For each j, let

$$E_j := C \cap B_1^{\text{on}}(\pi^*, \frac{\epsilon_{\text{off}} \wedge \epsilon_{\text{LM3}}}{j}) \cap B_2^{\text{on}}(\pi^*, \frac{\epsilon_{\text{off}} \wedge \epsilon_{\text{LM3}}}{j}) \cap B_2^{\text{off}}(\pi^*, \epsilon_{\text{off}}).$$

That is, E_j is the set of strategy profiles that respect rational compatibility, differ by no more than ϵ_{off}/j from π^* on path, and differ by no more than ϵ_{off} from π^* off path. It is clear that each E_j is convex and compact, and that $\lim_{j\to\infty} E_j \subseteq B_1^{\text{on}}(\pi^*, 0) \cap B_2^{\text{on}}(\pi^*, 0)$ as claimed.

We may find an accompanying sequence of survival probabilities satisfying

$$\gamma_j > \gamma_{\text{LM1}}(\delta, \frac{\epsilon_{\text{off}} \wedge \epsilon_{\text{LM3}}}{j}) \vee \underline{\gamma}_{\text{LM2}} \vee \gamma_{\text{LM3}}(\frac{\epsilon_{\text{off}} \wedge \epsilon_{\text{LM3}}}{j})$$

with $\gamma_j \uparrow 1$.

Step 3: $\mathscr{R}^{g,\delta,\gamma_j}$ maps E_j into itself.

Let some $\pi \in E_j$ be given.

By Lemma 1, $\mathscr{R}_1^{g,\delta,\gamma_j}[\pi_2] \in C$.

By Lemma 3, $\mathscr{R}_2^{g,\delta,\gamma_j}[\pi_1] \in B_2^{\text{off}}(\pi^*, \epsilon_{\text{off}})$, because uniformity of π^* means $\text{BR}(\hat{P}(s), s) \subseteq \tilde{A}(s)$ for each off-path s.

By Lemma 4, $\mathscr{R}_2^{g,\delta,\gamma_j}[\pi_1] \in B_2^{\text{on}}(\pi^*, \frac{\epsilon_{\text{off}} \wedge \epsilon_{\text{LM3}}}{j})$.

Finally, from Lemma 2 and the fact that $\pi_2 \in B_2^{\text{on}}(\pi^*, \frac{\epsilon_{\text{off}} \wedge \epsilon_{\text{LM3}}}{j}) \cap B_2^{\text{off}}(\pi^*, \epsilon_{\text{off}})$, we have $\mathscr{R}_1^{g,\delta,\gamma_j}[\pi_2] \in B_1^{\text{on}}(\pi^*, \frac{\epsilon_{\text{off}} \wedge \epsilon_{\text{LM3}}}{j})$. □

References

Banks, J.S., Sobel, J., 1987. Equilibrium selection in signaling games. Econometrica 55, 647–661.
Cho, I.-K., Kreps, D.M., 1987. Signaling games and stable equilibria. Q. J. Econ. 102, 179–221.
Dekel, E., Fudenberg, D., Levine, D.K., 1999. Payoff information and self-confirming equilibrium. J. Econ. Theory 89, 165–185.
Esponda, I., Pouzo, D., 2016. Berk-Nash equilibrium: a framework for modeling agents with misspecified models. Econometrica 84, 1093–1130.
Fudenberg, D., He, K., 2018. Learning and type compatibility in signaling games. Econometrica 86, 1215–1255.
Fudenberg, D., Kamada, Y., 2015. Rationalizable partition-confirmed equilibrium. Theor. Econ. 10, 775–806.
Fudenberg, D., Kamada, Y., 2018. Rationalizable partition-confirmed equilibrium with heterogeneous beliefs. Games Econ. Behav. 109, 364–381.
Fudenberg, D., Kreps, D.M., 1993. Learning mixed equilibria. Games Econ. Behav. 5, 320–367.
Fudenberg, D., Levine, D.K., 1993. Steady state learning and Nash equilibrium. Econometrica 61, 547–573.
Fudenberg, D., Levine, D.K., 2006. Superstition and rational learning. Am. Econ. Rev. 96, 630–651.
Kalai, E., Lehrer, E., 1993. Rational learning leads to Nash equilibrium. Econometrica 61, 1019–1045.
Kohlberg, E., Mertens, J.-F., 1986. On the strategic stability of equilibria. Econometrica 54, 1003–1037.
Sobel, J., Stole, L., Zapater, I., 1990. Fixed-equilibrium rationalizability in signaling games. J. Econ. Theory 52, 304–331.
Van Damme, E., 1987. Stability and Perfection of Nash Equilibria. Springer-Verlag.