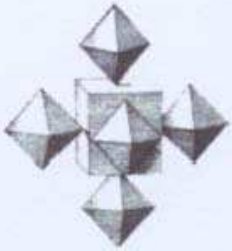

IKERLANAK

APPROXIMATE KNOWLEDGE OF RATIONALITY AND CORRELATED EQUILIBRIA

by

Fabrizio Germano and Peio Zuazo-Garin

2012

Working Paper Series: IL. 61/12

Departamento de Fundamentos del Análisis Económico I

Ekonomi Analisiaren Oinarriak I Saila

University of the Basque Country

Approximate Knowledge of Rationality and Correlated Equilibria*

Fabrizio Germano† and Peio Zuazo-Garin‡

July 16, 2012

Abstract

We extend Aumann's [3] theorem deriving correlated equilibria as a consequence of common priors and common knowledge of rationality by explicitly allowing for non-rational behavior. We replace the assumption of common knowledge of rationality with a substantially weaker notion, *joint* $\boldsymbol{p}$*-belief of rationality*, where agents believe the other agents are rational with probabilities $\boldsymbol{p} = (\boldsymbol{p_i})_{i \in I}$ or more. We show that behavior in this case constitutes a constrained correlated equilibrium of a doubled game satisfying certain $\boldsymbol{p}$-belief constraints and characterize the topological structure of the resulting set of $\boldsymbol{p}$-rational outcomes. We establish continuity in the parameters $\boldsymbol{p}$ and show that, for $\boldsymbol{p}$ sufficiently close to one, the $\boldsymbol{p}$-rational outcomes are close to the correlated equilibria and, with high probability, supported on strategies that survive the iterated elimination of strictly dominated strategies. Finally, we extend Aumann and Dreze's [4] theorem on rational expectations of interim types to the broader $\boldsymbol{p}$-rational belief systems, and also discuss the case of non-common priors.

Keywords: Correlated equilibrium, approximate common knowledge, bounded rationality, p-rational belief system, common prior, information, noncooperative game. *JEL Classification*: C72, D82, D83.

"Errare humanum est, perseverare diabolicum."

1 Introduction

Rationality, understood as consistency of behavior with stated objectives, information, and strategies available, naturally lies at the heart of game theory. In a celebrated paper, [3], Aumann takes as point of

*We are indebted to Bob Aumann, Eddie Dekel, Sergiu Hart, and Dov Samet, for valuable comments and conversations. Germano acknowledges financial support from the Spanish Ministry of Science and Technology (Grants SEJ2007-64340 and ECO2011-28965), as well as from the Barcelona GSE Research Network and the Generalitat de Catalunya. Zuazo-Garin acknowledges financial support from the Spanish Ministry of Science and Technology (Grant ECO2009-11213), and inestimable help from professors Elena Iñarra and Annick Laruelle, and from Jeon Hyang Ki. Both authors also thank the Center for the Study of Rationality at the Hebrew University of Jerusalem for their generous support and hospitality. All errors are our own.

†Universitat Pompeu Fabra, Departament d'Economia i Empresa, and Barcelona GSE, Ramon Trias Fargas 25-27, E-08005 Barcelona, Spain; fabrizio.germano@upf.edu.

‡*BRiDGE*, Universidad del País Vasco, Departamento de Fundamentos del Análisis Económico I, Avenida Lehendakari Aguirre 83, E-48015, Bilbao, Spain; peio.zuazo@ehu.es.

departure the rationality of all agents in any state of the world, knowledge thereof, as well as the existence of a common prior, to show that joint play by the players of the game will necessarily conform to a correlated equilibrium.

Given that the objectives, strategies and information structure are part of the description of the game or game situation at hand, they are relatively flexible restrictions and indeed in many cases should be adaptable to the actual underlying game. Nonetheless it is both possible and plausible that agents may deviate from a complete consistency with the assumed restrictions. This can happen in many ways as agents can make mistakes at any stage of participating in the game situation:[1] they can make a mistake in perceiving or interpreting their objective (e.g., a trader may apply a wrong set of exchange exchange rates when deciding a transaction), they may make a mistake in carrying out the selected strategy (e.g., a soccer player may trip or make a wrong step when shooting a penalty kick), they may make a mistake in processing their available information (e.g., a judge may forget to consult a crucial, requested technical report when deciding on a specific court case) and so on. It is clear that departures from full rationality not only occur, occur often, but also occur in innumerable ways.

In this paper, we build on the approach of [3] in formalizing the overall strategic interaction, but depart from it by allowing agents to entertain the possibility that other agents may *not* always be consistent. Specifically, we assume that agents believe that the other agents are rational with probability p_i and that with the remaining probability ($1 - p_i$ or less) they are *not* rational. In doing so we put as little structure as possible on what it means to be *not* rational, except that the rules of the game force agents to select *some* action in the game. This is in line with our belief that when modeling "social behavior" very broadly defined, mistakes or inconsistencies not only occur but also occur in innumerable ways. Therefore, without imposing assumptions on the kind of deviations from rational play that players may make, except that its probability of occurrence is limited by some vector $\boldsymbol{p}$, and maintaining the common prior assumption (common to all types of agents), we explore the consequences of what we call *joint $\boldsymbol{p}$-belief of rationality*. The two assumptions together, common prior and joint $\boldsymbol{p}$-belief of rationality, are what we call $\boldsymbol{p}$-rationality, for $\boldsymbol{p} \in [0,1]^I$.

The main result of the paper, Theorem 1, then characterizes strategic behavior in games where players are $\boldsymbol{p}$-rational and share a common prior. This provides a generalization of [3], since our $\boldsymbol{p}$-rational outcomes collapse to the correlated equilibria when $\boldsymbol{p} = 1$. Our second result, Theorem 2, shows basic topological properties of the $\boldsymbol{p}$-rational outcomes, in particular convexity, compactness but, more importantly, also that the set varies continuously in the underlining parameters $\boldsymbol{p}$. We also show that when $p = \min \boldsymbol{p}$ is sufficiently close to 1, then strategy profiles involving strategies that do not survive iterated elimination of strictly dominated strategies get probability at most p under the common prior. Following Aumann and Dreze in [4], we also ask what payoffs agents should expect under $\boldsymbol{p}$-rationality, and thus characterize in Theorem 3, what we call $\boldsymbol{p}$-rational expectations of interim types. Finally, we also briefly discuss the case of

[1] We refer to [9] and [16] for surveys on bounded rationality, and to [8] for a survey of behavioral game theory.

non-common priors and show that in this case, joint $\boldsymbol{p}$-belief of rationality puts no restriction on behavior. The paper only considers the case of complete information games.

Overall, the paper is structured as follows. Section 2 provides the basic definitions, Section 3 contains the main results characterizing the $\boldsymbol{p}$-rational outcomes and expectations, Section 4 provides some examples, and Section 5 some concluding remarks. All proofs are relegated to the Appendix.

2 p-Rational Belief Systems

Our main object of study are what we define as $\boldsymbol{p}$-rational belief systems. These generalize the rational belief systems of [3], [4], and [10], and allow us to formalize a situation where there is no common knowledge of rationality, but where players nonetheless believe, with probabilities above certain levels ($\boldsymbol{p} = (p_i)_i$), that the other players are rational. Following [4] and [10] we start by defining belief systems, which are more basic and which play a central role throughout the paper. They encapsulate what players believe about the game and about the other players.

2.1 Belief Systems

Throughout the paper we denote by $G = \langle I, (A_i)_{i \in I}, (h_i)_{i \in I} \rangle$ a finite *game* in strategic form defined the usual way, with I a finite set of players; A_i player i's set of pure strategies, which we also assume to be finite, where $A = \Pi_{i \in I} A_i$ and $A_{-i} = \Pi_{j \in I \setminus \{i\}} A_j$; and $h_i : A \to \mathbb{R}$ player i's payoff function.

Definition 1 (Belief System) *A* belief system *for the game G is a triple $B = \langle (T_i)_{i \in I}, (s_i)_{i \in I}, (f_i)_{i \in I} \rangle$ such that, for each player $i \in I$,*

- *T_i is a finite type set,*
- *$s_i : T_i \to A_i$ is a strategy map and*
- *$f_i \in \Delta(T)$ is i's prior, such that $T = \prod_{i \in I} T_i$, has full marginal support in T_i.*

We refer to any given element of T, $t = (t_i)_{i \in I}$, as a *state of the world*, or simply, *state*. Note that for each player $i \in I$, the prior f_i induces *posteriors* on other players' types, conditional on i's type:

$$f_i(t_i)(t_{-i}) = \frac{f_i(t_{-i}; t_i)}{f_i(T_{-i} \times \{t_i\})}, \text{ for any } t_{-i} \in T_{-i}.$$

We say that B satisfies *the common prior assumption*, CP, if $f_i = f_j$ for all $i, j \in I$, will denote them by just f, will refer to the latter as the *common prior* and will represent B by $\langle (T_i)_{i \in I}, (s_i)_{i \in I}, f \rangle$. For any state $t = (t_{-i}; t_i)$, we say that *player i is rational at t*, if

$$s_i(t_i) \in \underset{a_i \in A_i}{\operatorname{argmax}} \mathbb{E}_B \left(h(a_{-i}; a_i) \,|\, f_i(t_i) \right),$$

where $\mathbb{E}_B\left(h\left(a_{-i};a_i\right)|f_i\left(t_i\right)\right)=\sum_{t_{-i}\in T_{-i}} f_i(t_i)\left(t_{-i}\right)h_i\left(s_{-i}\left(t_{-i}\right);s_i\right)$. We denote by R_i the set of states in which player i is rational. Note that at any state t, player i being rational in this state only depends on the i-th component of t, namely, player i's type t_i. This implies the following characterization,

$$R_i=T_{-i}\times\left\{t_i\in T_i\;\middle|\;s(t_i)\in\underset{a_i\in A_i}{\operatorname{argmax}}\,\mathbb{E}_B\left(h_i\left(a_{-i};a_i\right)|f_i\left(t_i\right)\right)\right\}.$$

The set of the states in which every player but i is rational (without excluding i's possible rationality), $\bigcap_{j\in I\setminus\{i\}}R_j$, will be denoted by R_{-i}. By R, we will denote the set of states in which every player is rational. We will say that a belief system B satisfies *common knowledge of rationality*, $CK\left(R\right)$, if every player is rational at every state, that is, if $T=R$.[2]

Finally, we say that a belief system is *rational* if it satisfies both CP and $CK\left(R\right)$. It is well known from [3] that the outcome distributions over the set of strategies of a finite game in strategic form G played under some rational belief system B, are precisely the correlated equilibria of game G, which we denote by $CE(G)$.

2.2 p-Belief Operators and p-Rationality

By an *event* we mean any subset $E\subseteq T$. Recall that, following [14], and making the corresponding adaptations to the setup in the previous section, for any given $p_i\in[0,1]$, and any event $E\subseteq T$, we can define player i's *p_i-belief operator* as,

$$B_i^{p_i}\left(E\right)=T_{-i}\times\left\{t_i\in T_i\mid f_i\left(t_i\right)\left(E\right)\geq p_i\right\}.$$

For any $t\in T$, any event E, and any $p_i\in[0,1]$, we say that *player i p_i-believes E in t*, if $t\in B_i^{p_i}\left(E\right)$. For $\boldsymbol{p}=\left(p_i\right)_{i\in I}\in[0,1]^I$, we say that an event E is *$\boldsymbol{p}$-evident* if $E\subseteq\bigcap_{i\in I}B_i^{p_i}\left(E\right)$, and given any state t, we say that an event C is *common $\boldsymbol{p}$-belief in t* if there exists some $\boldsymbol{p}$-evident event E such that $t\in E\subseteq\bigcap_{i\in I}B_i^{p_i}\left(C\right)$. We denote the event that C is common $\boldsymbol{p}$-belief by $CB^{\boldsymbol{p}}(C)$. We say that a belief system B satisfies *common $\boldsymbol{p}$-belief of rationality* if $T=CB^{\boldsymbol{p}}\left(R\right)$. Now, it is easy to see that common $\boldsymbol{p}$-belief of rationality is satisfied if and only if $T=\bigcap_{i\in I}B_i^{p_i}\left(R\right)$, so the latter is, for convenience, the characterization of common $\boldsymbol{p}$-belief of rationality we will make use of through the paper.[3] As T is finite, it is again easy to check that a belief system satisfies $\mathbf{1}$-belief of rationality if and only if it satisfies common knowledge of rationality.

For reasons explained below, the following notion of approximate knowledge of rationality replaces the notion of $CK(R)$ assumed in [3] and elsewhere.

Definition 2 (Joint p-Belief of Rationality) *A belief system B satisfies* joint $\boldsymbol{p}$-belief of rationality*, $J\boldsymbol{p}B\left(R\right)$, if $T=\bigcap_{i\in I}B_i^{p_i}\left(R_{-i}\right)$.*

[2] As is shown in [14], p.174, this does indeed imply common knowledge of rationality. In this case, T is clearly such an *evident knowledge*, and, hence, also R.

[3] Despite not being true, in general for an event E, that $CB^{\boldsymbol{p}}(E)=\bigcap_{i\in I}B_i^{p_i}\left(E\right)$.

The following are basic results regarding relations between common knowledge of rationality, common $\boldsymbol{p}$-belief of rationality and $J\boldsymbol{p}B(R)$.

Lemma 1 *Let G be a game and B a belief system for G, then:*

(a) For any $\boldsymbol{p} \in (0,1]^I$, if B satisfies $CB^{\boldsymbol{p}}(R)$, then B satisfies $CK(R)$, and is therefore, rational.

(b) For any $\boldsymbol{p} \in (0,1]^I$, if B satisfies CP and $J_{\boldsymbol{p}}B(R)$, then $f(CB^{\boldsymbol{p}}(R)) \geq f(R) \geq p^2$, where $p = \min \boldsymbol{p}$.

(c) For $\boldsymbol{p} = 1$, we have that B satisfies $CB^{\boldsymbol{p}}(R)$ if and only if B satisfies $J_{\boldsymbol{p}}B(R)$.

Hence, if our aim is to replace the $CK(R)$ assumption by a less restrictive one involving some $\boldsymbol{p}$-belief of rationality, the lemma above suggests that $J\boldsymbol{p}B(R)$ could be a sensible choice, since:

- *common* $\boldsymbol{p}$-belief of rationality would lead to a model analogous to the one already considered before (or without) introducing $\boldsymbol{p}$-beliefs (this holds true with or without a common prior),
- although joint $\boldsymbol{p}$-belief of rationality is defined as a *joint* and not necessarily *common* belief, it nonetheless implies common $\boldsymbol{p}$-belief of rationality with probability $(\min \boldsymbol{p})^2$, and,
- as every p_i converges to 1, joint $\boldsymbol{p}$-belief of rationality converges to *common* $\mathbf{1}$-belief of rationality (or *common certainty* of rationality) and therefore, to common knowledge of rationality.

In view of this, the following generalization of rational belief systems plays a key role in our analysis.

Definition 3 ($\boldsymbol{p}$-Rational Belief System) *A belief system is* $\boldsymbol{p}$-rational *if it satisfies both CP and $J\boldsymbol{p}B(R)$.*

Our aim is to characterize the set of distributions over outcomes or strategy profiles of complete information games, played under $\boldsymbol{p}$-rational belief systems, as is done for rational belief systems in [3].[4]

3 $\boldsymbol{p}$-Rational Outcomes

The following definition formalizes the notion of distributions over outcomes that satisfy CP and $J\boldsymbol{p}B(R)$.

Definition 4 ($\boldsymbol{p}$-Rational Outcome) *Let G be a game and $\boldsymbol{p} \in [0,1]^I$, then $\pi \in \Delta(A)$ is a* $\boldsymbol{p}$-rational outcome *of G if there exists a $\boldsymbol{p}$-rational belief system B with strategy map s and common prior f such that,*

$$\pi(a) = \left(f \circ [s]^{-1}\right)(a) \text{ for any } a \in A.$$

We denote the outcome distribution over the set of strategies of G, induced by the $\boldsymbol{p}$-rational belief system B, by π_B, and the set of $\boldsymbol{p}$-rational outcomes of G, by $\boldsymbol{p}$-$RO(G)$.

[4] Another natural alternative, suggested to us by Dov Samet, would be to consider belief systems that satisfy $f(CB^{\boldsymbol{p}}(R)) \geq 1-\epsilon$ for some $\epsilon > 0$. This is a weakening of our notion of $J_{\boldsymbol{p}}B(R)$ in that it imposes less structure on the beliefs of all types, nonetheless it appears to be less tractable; we return to this later.

Our main objective in this paper is to characterize the set $\boldsymbol{p}$-$RO(G)$. In order to do so, we will first introduce the following notion of equilibrium, which generalizes that of correlated equilibrium:

Definition 5 ($(X,\boldsymbol{p})$-Correlated Equilibrium) *Let G be a game, $X=\prod_{i\in I}X_i\subseteq A$, and $\boldsymbol{p}\in[0,1]^I$. Then, $\pi\in\Delta(A)$ is a* $(X,\boldsymbol{p})$-correlated equilibrium *of G if and only if, for any $i\in I$:*

- *For any $a_i'\in X_i$, the following* incentive constraints *are satisfied:*

$$\sum_{a_{-i}\in A_{-i}}\pi(a_{-i};a_i')\left[h_i(a_{-i};a_i')-h_i(a_{-i};a_i)\right]\geq 0,\text{ for any } a_i\in A_i,$$

- *For any $a_i\in A_i$ the following p_i-*belief constraint *is satisfied:*

$$\sum_{a'_{-i}\in X_{-i}}\pi\left(a'_{-i};a_i\right)\geq p_i\sum_{a_{-i}\in A_{-i}}\pi(a_{-i};a_i).$$

We denote by $(X,\boldsymbol{p})$-$CE(G)$ the set of $(X,\boldsymbol{p})$-correlated equilibria of G.

Hence, we fix for each player, $i\in I$, pure actions $X_i\subseteq A_i$ and a probability $p_i\in[0,1]$, such that the distribution on the overall set of action profiles A satisfies, for each $i\in I$, (i) standard *incentive constraints* for all actions in X_i, and, (ii) *p_i-belief constraints*, meaning each player assigns probability at least p_i to the *other* players all choosing action profiles from $X_{-i}=\prod_{j\neq i}X_i$. In words, the $(X,\boldsymbol{p})$-correlated equilibria allow to *relax* incentive constraints on actions not in the X_i's, while restricting the probability with which this occurs. Note that if $X=A$ or $\boldsymbol{p}=\mathbf{1}$, then $(X,\boldsymbol{p})$-$CE(G)=CE(G)$. The following definition generalizes the idea of *doubled game* in [3]. For any $n\in\mathbb{N}$ we denote $N=\{0,1,\ldots,n-1\}$.

Definition 6 (n-Game) *Let $G=\left\langle I,(A_i)_{i\in I},(h_i)_{i\in I}\right\rangle$ be a game, then the* n-game *is the tuple $nG=\left\langle I,(nA_i)_{i\in I},(h_{n,i})_{i\in I}\right\rangle$, where for each player $i\in I$ we have,*

- *$nA_i=A_i\times N$ is player i's set of pure actions; we denote a generic element of $nA=\prod_{i\in I}nA_i$ by (a,α), where $\alpha\in N^I$ specifies which copy of A_i in nA_i each player i's pure action belongs to.*
- *$h_{n,i}$ is i's payoff function, where for each $(a,\alpha)\in nA$, $h_{n,i}(a,\alpha)=h_i(a)$.*

In this context when writing the action spaces of the game nG as $nA_i=A\times N$ meaning that for each player there are n *copies* of the original action space A_i, which we denote by $A_i\times\{k\}$ for $k\in N$.

Note that any distribution on the action profiles of nG, $\hat{\pi}\in\Delta(nA)$, induces a distribution on the action profiles of G in a natural way:

$$\begin{array}{rccc}\text{Proj}_A: & \Delta(nA) & \longrightarrow & \Delta(A)\\ & \hat{\pi} & \to & \pi: A\longrightarrow [0,1]\\ & & & a\to \hat{\pi}\left(\{a\}\times N^I\right)\end{array}$$

With these definitions, the $\boldsymbol{p}$-rational outcomes of G are readily characterized:

Theorem 1 *Let $G = \langle I, (A_i)_{i\in I}, (h_i)_{i\in I}\rangle$ be a game and $\boldsymbol{p} \in [0,1]^I$. Then the $\boldsymbol{p}$-rational outcomes of G are the projection on $\Delta(A)$ of the $(X, \boldsymbol{p})$-correlated equilibria of the doubled game $2G$, where $X = A \times \{0\}$ is a copy of the original action space of G. Formally,*

$$\boldsymbol{p}\text{-}RO(G) = \mathrm{Proj}_A\left[(X, \boldsymbol{p})\text{-}CE(2G)\right],$$

where $X = A \times \{0\}$.

Clearly, due to the symmetric role of the different copies of the action spaces in the game $2G$, the theorem would also hold for $X = A \times \{1\}$. Only one of the two copies of players' actions satisfies the incentive constraints (these are *rational* types) and does so with probabilities consistent with the p_i's. The next two results characterize further the structure and nature of the set of $\boldsymbol{p}$-rational outcomes.

Theorem 2 *Let G be a finite game in strategic form with set of players I, and $\boldsymbol{p} \in [0,1]^I$. Then the set of $\boldsymbol{p}$-rational outcomes of the game G is a nonempty, convex, compact set that varies continuously in $\boldsymbol{p}$.*[5]

Moreover, for $\boldsymbol{p} = \mathbf{0}$, we have $\mathbf{0}$-$RO(G) = \Delta(A)$, for $\boldsymbol{p} = \mathbf{1}$, we have $\mathbf{1}$-$RO(G) = CE(G)$, and for any $\boldsymbol{p} \in [0,1)^I$, we have $dim[\boldsymbol{p}\text{-}RO(G)] = dim[\Delta(A)]$.

In [14], Monderer and Samet show (using common $\boldsymbol{p}$-beliefs) that the $\boldsymbol{p}$-rational outcomes for $\boldsymbol{p} = \mathbf{1}$ are the correlated equilibria. The above result strengthens this by showing that as $\boldsymbol{p}$ converges to 1 the $\boldsymbol{p}$-rational outcomes converge to the set of correlated equilibria. But more generally it also shows that the $\boldsymbol{p}$-rational outcomes *always* vary continuously in $\boldsymbol{p}$, at any $\boldsymbol{p} \in [0,1]^I$; and go from being the entire set $\Delta(A)$ when $\boldsymbol{p} = \mathbf{0}$ to being the set of correlated equilibria when $\boldsymbol{p} = \mathbf{1}$.

The very last statement further shows that *all* strategies can be in the support of $\boldsymbol{p}$-rational outcomes whenever $\boldsymbol{p} < \mathbf{1}$. The next result qualifies this by showing that if $p = \min \boldsymbol{p}$ is close enough 1, then strategy profiles involving strategies that do not survive the iterated elimination of strictly dominated strategies get a total weight of at most $1 - p$. This can be interpreted as the $\boldsymbol{p}$-rationality counterpart of the fact that strategies that do not survive the iterated elimination of strictly dominated strategies are not in the support of correlated equilibria. In what follows we denote by A^∞ the set of all strategy profiles that survive the iterated elimination of strictly dominated strategies and denote its complement in A by $(A^\infty)^c = A \setminus A^\infty$.

Proposition 1 *Let G be a finite game in strategic form with set of players I, then there exists $\bar{p} \in (0,1)$ such that, for any $\boldsymbol{p} \in [0,1]^I$ with $\min \boldsymbol{p} = p \in [\bar{p}, 1]$, we have that, if $\pi \in \boldsymbol{p}$-$RO(G)$, then $\pi((A^\infty)^c) \leq 1 - p$.*

The above results confirm in a precise sense the robustness of the correlated equilibrium benchmark when weakening the underlying assumption of common knowledge of rationality to joint $\boldsymbol{p}$-belief of rationality.

[5] A correspondence is *continuous* if it is both upper- and lower-hemicontinuous; see, e.g., Ch. 17 in [1] for further details and related definitions.

p-Rational Expectations

Following [4], we can analyze expected payoffs or *expectations in a game* from the point of view of a fixed player. We assume this player knows his or her type given any belief system.

Definition 7 (p-Rational Expectation) *Let G be a game, $\boldsymbol{p} \in [0,1]^I$, and B a $\boldsymbol{p}$-rational belief system for G, then a* $\boldsymbol{p}$-rational expectation *in G is the expected payoff of some type of some player. We denote the set of all such $\boldsymbol{p}$-rational expectations of G, by $\boldsymbol{p}$-$RE(G)$.*

It is then easy to characterize:

Theorem 3 *Let G be a game and $\boldsymbol{p} \in [0,1]^I$. Then the $\boldsymbol{p}$-rational expectations in G are the expected payoffs of the $(A \times \{0\}, \boldsymbol{p})$-correlated equilibria of the tripled game $3G$, conditional on playing an action in $3A_i$. Moreover, the $\boldsymbol{p}$-rational expectations of the rational types are the expected payoffs of the $(A \times \{0\}, \boldsymbol{p})$-correlated equilibria of the tripled game $3G$, conditional on playing an action in $A_i \times \{0\}$.*

This provides the joint $\boldsymbol{p}$-belief of rationality counterpart of Aumann and Dreze's characterization in [4].

4 Examples

The following examples illustrate the $\boldsymbol{p}$-rational outcomes for some 2×2 games.

Example 1 (Dominance Solvable Game) Consider the following game G_D solvable by strict dominance with corresponding augmented game $2G_D$,

$$G_D \equiv \begin{array}{c|cc} & L & R \\ \hline T & 2,2 & 1,1 \\ B & 1,1 & 0,0 \end{array}, \qquad 2G_D \equiv \begin{array}{c|cccc} & (L,0) & (R,0) & (L,1) & (R,1) \\ \hline (T,0) & 2,2 & 1,1 & 2,2 & 1,1 \\ (B,0) & 1,1 & 0,0 & 1,1 & 0,0 \\ (T,1) & 2,2 & 1,1 & 2,2 & 1,1 \\ (B,1) & 1,1 & 0,0 & 1,1 & 0,0 \end{array}.$$

To compute the $\boldsymbol{p}$-$RO(G_D)$ we compute the $(A \times \{0\}, \boldsymbol{p})$-$CE(2G_D)$. For this notice that the strategies $(B,0)$ and $(T,1)$ of the row player and $(R,0)$ and $(L,1)$ of the column player are strictly dominated, so that the remaining constraints that need to be satisfied are the $\boldsymbol{p}$-belief constraints, and one obtains,

$$\boldsymbol{p}\text{-}RO(G_D) = \left\{ \pi \in \Delta(A) \;\middle|\; \begin{array}{l} \pi_{TL} \geq p_1(\pi_{TL} + \pi_{TR}), \pi_{BL} \geq p_1(\pi_{BL} + \pi_{BR}) \\ \pi_{TL} \geq p_2(\pi_{TL} + \pi_{BL}), \pi_{TR} \geq p_2(\pi_{TR} + \pi_{BR}) \end{array} \right\}.$$

Figures 1 and 2 show the set $\boldsymbol{p}$-$RO(G_D)$ for $\boldsymbol{p} = (0.95, 0.95)$ together with respectively the ϵ-neighborhood of the set of correlated equilibria of G_D, $N_\epsilon(CE(G_D))$, and the set of ϵ-correlated equilibria, ϵ-$CE(G_D)$,[6] both with $\epsilon = 0.20$. Clearly the three sets are all distinct.

[6] In general, this is the set of probability distributions $\pi \in \Delta(A)$ that satisfy the incentive constraints for correlated equilibria

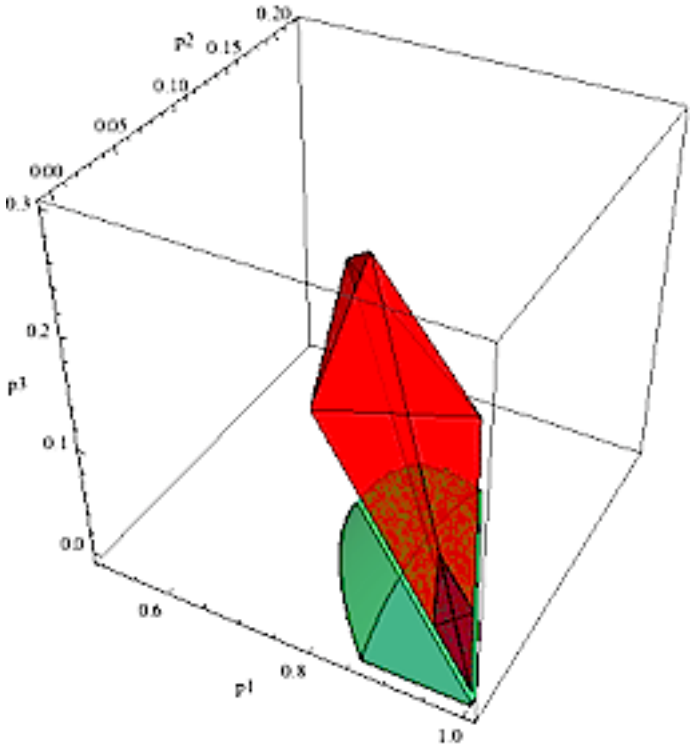

Figure 1: 0.95-$RO(\Gamma_D)$ (blue), 0.80-$RO(\Gamma_D)$ (red), $N_{0.10}CE(\Gamma_D)$ (green)

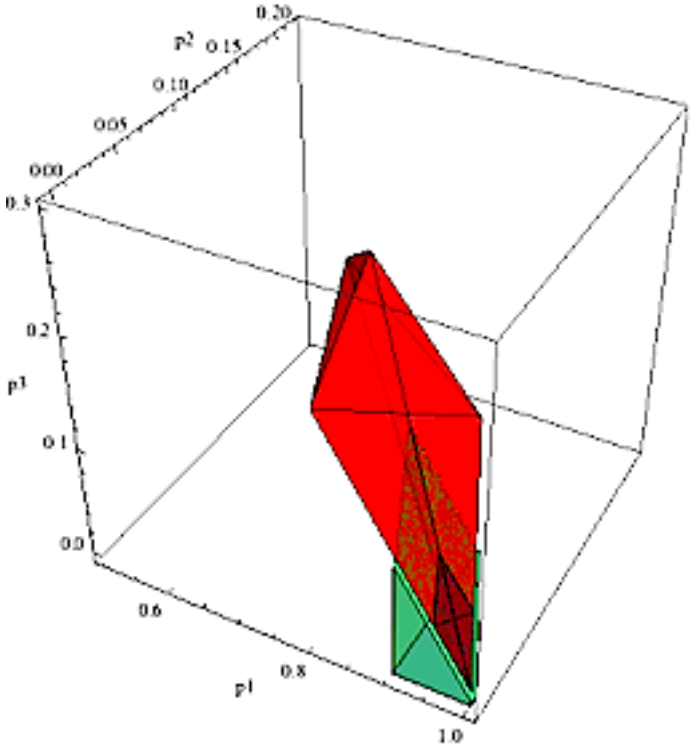

Figure 2: 0.95-$RO(G_D)$ (blue), 0.80-$RO(G_D)$ (red), 0.10-$CE(G_D)$ (green)

Example 2 (Matching Pennies Game) Consider the following version G_{MP} of matching pennies, with corresponding doubled game $2G_{MP}$,

$$G_{MP} \equiv \begin{array}{c|cc} & L & R \\ \hline T & 1,0 & 0,1 \\ B & 0,1 & 1,0 \end{array}, \qquad 2G_{MP} \equiv \begin{array}{c|cccc} & (L,0) & (R,0) & (L,1) & (R,1) \\ \hline (T,0) & 1,0 & 0,1 & 1,0 & 0,1 \\ (B,0) & 0,1 & 1,0 & 0,1 & 1,0 \\ (T,1) & 1,0 & 0,1 & 1,0 & 0,1 \\ (B,1) & 0,1 & 1,0 & 0,1 & 1,0 \end{array}.$$

The set $\boldsymbol{p}$-$RO(G_{MP})$ is now somewhat more tedious to characterize, nonetheless we know it is a compact, convex polyhedron around $\bar{\pi} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, which converges to $\bar{\pi}$ as $\boldsymbol{p}$ converges to 1. In particular it contains profiles that do not yield the agents their value of the game, but rather something in a neighborhood thereof.

Figures 3 and 4 show the set $\boldsymbol{p}$-$RO(G_{MP})$ for $\boldsymbol{p} = (0.95, 0.95)$ together with respectively the ϵ-neighbourhood of the set of correlated equilibria, $N_\epsilon(CE(G_{MP}))$, and the set of ϵ-correlated equilibria, ϵ-$CE(G_{MP})$, both with $\epsilon = 0.10$. Again, the sets $\boldsymbol{p}$-$RO(G_{MP})$ and $N_\epsilon(CE(G_{MP}))$ and ϵ-$CE(\Gamma_{MP})$ are visibly distinct.

The following example further illustrates the relationship with the ϵ equilibria.

Example 3 (Prisoner's Dilemma Game) Consider the following game G_{PD} with corresponding probability of play π_{PD},

$$G_{PD} \equiv \begin{array}{c|cc} & L & R \\ \hline T & 1,\ 1 & 4,\ 0.99 \\ B & 0.99,\ 4 & 3.99, 3.99 \end{array}, \qquad \pi_{PD} \equiv \begin{array}{c|cc} & L & R \\ \hline T & 0 & 0 \\ B & 0 & 1 \end{array}.$$

It is clear that π_{PD} constitutes both an ϵ-Nash and an ϵ-correlated equilibrium, for any $\epsilon > 0.01$, yet $\pi_{PD} \in \boldsymbol{p}$-$RO(G_{PD})$ if and only if $\boldsymbol{p} = 0$.

5 Some Remarks

We conclude with a few remarks.

Remark 1 (*P*-Rational Outcomes and Expectations) An important objective of the paper was to put as few restrictions on the non-rational types as possible, the idea being that these should cover all sorts of departures from rationality such as making mistakes in choosing actions, mistakes in reading payoffs, presentation effects, etc. However, one important assumption made implicitly throughout the paper concerns the *beliefs* of non-rational types about other agents. Indeed it was assumed as part of the notion of $J\boldsymbol{p}R(G)$ that both rational and non-rational types all believe that the other players are rational with probability

with a slack of ϵ, analagous to Radner's ϵ-Nash equilibria, formally, π is an *ϵ-correlated equilibrium* (ϵ-CE) if for any $i \in I$,

$$\sum_{a_i \in A_i} \max_{a_i' \in A_i} \sum_{a_{-i} \in A_{-i}} \pi(a_i, a_{-i}) \left(h_i(a_i', a_{-i}) - h_i(a_i, a_{-i}) \right) \leq \epsilon.$$

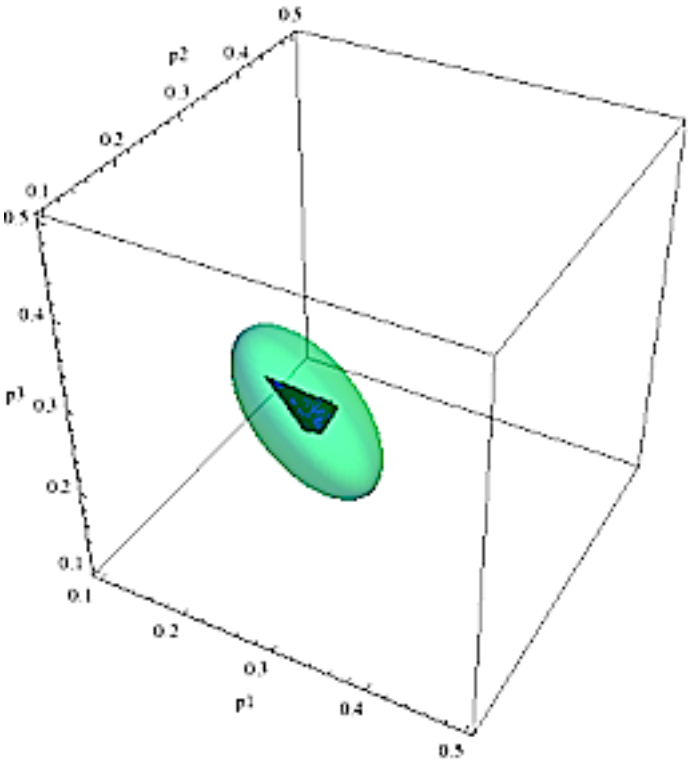

Figure 3: 0.95-$RO(G_{MP})$ (blue), $N_{0.10}CE(G_{MP})$ (green)

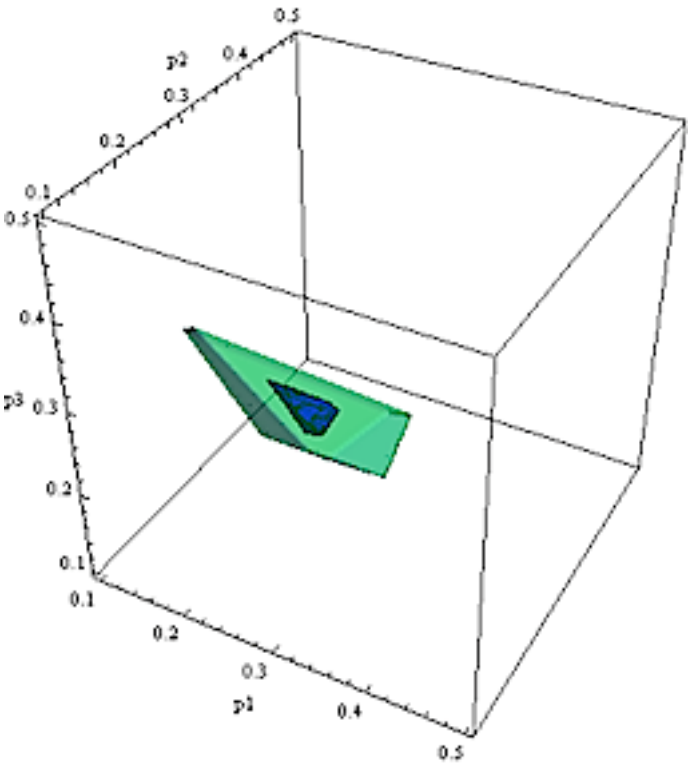

Figure 4: 0.95-$RO(G_{MP})$ (blue), 0.10-$CE(G_{MP})$ (green)

p_i or more. Another benchmark in line with our motivation would be to drop any restriction on the non-rational types and allow them to have any kind of beliefs about others. This can be formalized by assuming a larger vector $\boldsymbol{P} = (\boldsymbol{p}, \mathbf{0}) \in [0,1]^{2I}$ where the components associated to the rational types are the usual probabilities $\boldsymbol{p}$, while the components associated to the non-rational types are all zero. This leads to $\boldsymbol{P}$-*rational outcomes* of G. These are projections of $(A \times \{0\}, \boldsymbol{P})$-correlated equilibria of $2G$, in that they are distributions satisfying the same conditions as the $(A \times \{0\}, \boldsymbol{p})$-$CE(2G)$ except that the $\boldsymbol{p}$-belief constraints for the non-rational types are dropped. In other words, $(A \times \{0\}, \boldsymbol{p})$-$CE(2G) \subset (A \times \{0\}, \boldsymbol{P})$-$CE(2G)$ and therefore also $\boldsymbol{p}$-$RO(G) \subset \boldsymbol{P}$-$RO(G)$.

While the correspondence $\boldsymbol{P}$-$RO(G)$ maintains the basic topological properties of the correspondence $\boldsymbol{p}$-$RO(G)$, it need not converge to the set of correlated equilibria of G as $\boldsymbol{P} \to (\mathbf{1}, \mathbf{0})$, but does so if one also requires $\boldsymbol{P} \to (\mathbf{1}, \mathbf{1})$. This can be seen already in Example 1. A $(\mathbf{1}, \mathbf{0})$-rational belief system can be very far from a $(\mathbf{1}, \mathbf{1})$-rational belief system in that the former need not put any restriction on the total mass of rational types $f(R)$.[7]

The alternative notion of approximate knowledge of rationality requiring $f(CB^{\boldsymbol{p}}) > 1 - \epsilon$, for $\epsilon > 0$, (instead of $J_{\boldsymbol{p}}B(R)$), is more flexible with respect to the players' beliefs in that it only restricts the total mass of common $\boldsymbol{p}$-belief and hence does not specify directly what beliefs individual players and types have. A characterization of $\boldsymbol{p}$-rational outcomes with this definition is possible along the lines of our Theorem 1, but involves more complicated incentive and $\boldsymbol{p}$-belief constraints that are imposed over all possible subsets and permutations of players.

Remark 2 (Non-Common Priors) Throughout the paper we assumed the existence of a common prior (CP). This together with the notion of joint $\boldsymbol{p}$-belief of rationality allowed us to derive relatively stringent restrictions on behavior. At the same time it is natural to ask, what happens if the common prior assumption is relaxed. As it turns out, under *subjective* or *non-common* priors, joint $\boldsymbol{p}$-belief of rationality puts *no* restrictions on possible behavior – even when $\boldsymbol{p} = \mathbf{1}$. This provides a stark contrast with the cases of common knowledge of rationality and also common p-belief of rationality as studied respectively in [2, 5, 7, 15, 17] and [6, 12], and in a sense further highlights the stringency of the common prior assumption.[8]

To see the non-common prior case, define belief system $B = \left\langle (T_i)_{i \in I}, (s_i)_{i \in I}, (f_i)_{i \in I} \right\rangle$ to be *subjectively* $\boldsymbol{p}$-*rational* if $T = \bigcap_{i \in I} B_i^{p_i}(R_{-i})$. Given a finite game in strategic form G with set of players I and set of action profiles A, and given $\boldsymbol{p} \in [0,1]^I$, we say that a family of distributions $(\pi_i)_{i \in I} \in (\Delta(A))^I$ is a

[7] To see this, let $\boldsymbol{P} = (\boldsymbol{p}, \boldsymbol{q}) \in [0,1]^{2I}$, where $\boldsymbol{p}$, $\boldsymbol{q}$ are the probabilities for the rational and non-rational types respectively. To see that in a $(\mathbf{1}, \mathbf{0})$-rational belief system the total mass of non-rational types is unrestricted, take the game in Example 1 and consider the belief system $B = \left\langle (T_i)_{i \in I}, (s_i)_{i \in I}, (f_i)_{i \in I} \right\rangle$, where $T_i = A_i$, $s_i(a_i) = a_i$, for all $a_i \in A_i$, $i \in I$, and where $f \in \Delta(A)$ is given by $f_{TL} = f_{TR} = f_{BL} = 0$ and $f_{BR} = 1$. It can be checked that it is $(\mathbf{1}, \mathbf{0})$-rational and clearly $f(R) = 0$. At the same time, in a $(\boldsymbol{p}, \boldsymbol{q})$-rational belief system it is always the case that, for any $i \in I, t_i \in T_i$, $f_i(t_i)(R_{-i}) \geq q_i$, hence

$$f(R_{-i} \cap (T_{-i} \times \{t_i\}) \geq q_i f(T_{-i} \times \{t_i\}) \Longrightarrow \sum_{t_i \in T_i} f(R_{-i} \cap (T_{-i} \times \{t_i\}) = q_i \sum_{t_i \in T_i} f(T_{-i} \times \{t_i\}) \Longrightarrow f(R_{-i}) \geq q_i,$$

which besides confirming the expected convergence to the correlated equilibria as $(\boldsymbol{p}, \boldsymbol{q}) \to (\mathbf{1}, \mathbf{1})$, also shows that positive q_i's do put restrictions on the total mass of rational types $f(R)$.

[8] Recall that the result of Lemma 1(a) also holds with non-common priors.

p*-*subjectively rational outcome of G (***p*-$SRO(G)$**) if there exists some subjectively $\boldsymbol{p}$-rational belief system $B = \left\langle (T_i)_{i\in I}, (s_i)_{i\in I}, (f_i)_{i\in I} \right\rangle$ for G such that for any $i \in I$, we have $\pi_i = f_i \circ [s]^{-1}$. As shown in the Appendix, it is easy to see that, for any $\boldsymbol{p} \in [0,1]^I$, the whole space is obtained, namely:

$$\boldsymbol{p}\text{-}SRO\,(G) = (\Delta\,(A))^I\,.$$

In particular any pure strategy profile in A is consistent with subjective $\boldsymbol{p}$-rationality, even when $\boldsymbol{p} = \mathbf{1}$.

Remark 3 (Comparison with Further Solution Concepts) The sets of $\boldsymbol{p}$-rational outcomes define sets of probability distributions of play that are broader than the correlated equilibria that follow their own logic. As the examples show, they are distinct from ϵ-neighbourhoods of the correlated equilibria, thus putting further structure on the types of deviations from the set $CE(G)$ that occur as $\boldsymbol{p}$ departs from $\mathbf{1}$. At the same time, they are distinct from the ϵ-correlated equilibria, reflecting the fact that they impose no constraints on the *type* of departure from rationality assumed – unlike with the ϵ-correlated equilibria, which assume the agents are ϵ-optimizers. A similar remark applies to the quantal response equilibria of McKelvey and Palfrey [13] or other models such as the level-k reasoning models (e.g, [8]) that put specific restrictions on how players can deviate from rationality. It remains an empirical question to what extent the $\boldsymbol{p}$-rational outcomes bound observed behavior in a robust and useful manner.

Remark 4 (Learning to Play *p*-Rational Outcomes) Clearly, all learning dynamics that lead to correlated equilibria (see e.g., [11]) will also lead to $\boldsymbol{p}$-rational outcomes. The question arises as to what further dynamics (not necessarily converging to correlated equilibria) may converge to $\boldsymbol{p}$-rational outcomes and whether they include interesting dynamics that for example allow for faster or more robust convergence.

References

[1] Aliprantis, C.D. and K.C. Border (2006) *Infinite Dimensional Analysis: A Hitchhiker's Guide*, (Third Ed.), Berlin, Springer Verlag.

[2] Aumann, R.J. (1974) "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics*, **1**: 67–96.

[3] Aumann, R.J. (1987) "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, **55**: 1–18.

[4] Aumann, R.J. and J.H. Dreze (2008) "Rational Expectations in Games," *American Economic Review*, **98**: 72–86.

[5] Pearce, D. (1984) "Rationalizable Strategic Behavior," *Econometrica*, **52**: 1007–1028.

[6] Börgers, T. (1996) "Weak Dominance and Approximate Common Knowledge," *Journal of Economic Theory*, **64**: 265-276.

[7] Brandenburger, A., and E. Dekel (1987) "Rationalizability and Correlated Equilibria," *Econometrica*, **55**: 1391–1402.

[8] Camerer, C.F. (2003) *Behavioral Game Theory: Experiments in Strategic Interaction*, New York, Princeton University Press.

[9] Conlisk, J. (1996) "Why Bounded Rationality?," *Journal of Economic Literature*, **34**: 669-700.

[10] Harsanyi, J.C. (1967–1968) "Games with Incomplete Information Played by 'Bayesian' Players. I–III," *Management Science*, **14**: 159–182, 320–334, 486–502.

[11] Hart, S. (2005) "Adaptive Heuristics," *Econometrica*, **73**: 1401–1430.

[12] Hu, T.W. (2007) "On p-Rationalizability and Approximate Common Certainty of Rationality," *Journal of Economic Theory*, **136**: 379-391.

[13] McKelvey, R.D., and T.R. Palfrey (1995) "Quantal Response Equilibria in Normal Form Games," *Games and Economic Behavior*, **7**: 6–38.

[14] Monderer, D. and D. Samet (1989) "Approximating Common Knowledge with Common Beliefs," *Games and Economic Behavior*, **1**: 170–190.

[15] Pearce, D. (1994) "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, **52**: 1029–1050.

[16] Rubinstein, A. (1998) *Modelling Bounded Rationality*, Cambridge, MA, MIT Press.

[17] Tan, T.C., and S.R. Werlang (1988) "The Bayesian Foundation of Solution Concepts of Games," *Journal of Economic Theory*, **45**: 370-391.

APPENDIX

A Proofs of Lemma 1

(a) By definition, $CB^{\boldsymbol{p}}(R) \subseteq \bigcap_{i\in I} B_i^{p_i}(R)$, and therefore, for any $i \in I$, $CB^{\boldsymbol{p}}(R) \subseteq B_i^{p_i}(R_i)$. But now:

$$(t_{-i}; t_i) \in B_i^{p_i}(R_i) \iff f_i(t_i)(R_i) = \frac{f(R_i \cap (T_{-i} \times \{t_i\}))}{f(T_{-i} \times \{t_i\})} \geq p_i.$$

So, as $p_i > 0$, $T_{-i} \times \{t_i\} \subseteq R_i$ and therefore, $B_i^{p_i}(R_i) \subseteq R_i$.

(b) First, if $T = \bigcap_{i\in I} B_i^p(R_{-i})$, then:

$$R = R \cap T = \bigcap_{i\in I} (B_i^p(R_{-i}) \cap R_i) = \bigcap_{i\in I} B_i^{p_i}(R) \subseteq CB^p(R),$$

and therefore, $R \subseteq CB^p(R)$. Now, again, since $T = \bigcap_{i\in I} B_i^p(R_{-i})$, we have both:

$$f(R_{-i}) \geq p, \text{ and } f(R_{-i} \cap R_i) \geq pf(R_i).$$

The fact that, for any $j \neq i$, $f(R) = f(R_{-i}|R_i) f(R_i) \geq pf(R_i) \geq pf(R_{-j}) \geq p^2$ completes the proof.

(c) As mentioned in the paper, it is easy to check that:

$$T = (CB^{\boldsymbol{p}})_{|\boldsymbol{p}=\boldsymbol{1}}(R) \iff T = \bigcap_{i\in I} B_i^{\boldsymbol{1}}(R).$$

Now, take $t = (t_i)_{i\in I}$. Then,

$$t \in \bigcap_{i\in I} B_i^{\boldsymbol{1}}(R) \iff \forall i \in I, f_i(t_i)(R) = 1 \iff$$

$$\iff \forall i \in I, f_i(R \cap (T_{-i} \times \{t_i\})) = f_i(T_{-i} \times \{t_i\}) \overset{*}{\iff}$$

$$\overset{*}{\iff} \forall i \in I, f_i(R_{-i} \cap (T_{-i} \times \{t_i\})) = f_i(T_{-i} \times \{t_i\}) \iff$$

$$\iff t \in \bigcap_{i\in I} B_i^1(R_{-i}).$$

* The left implication is immediate; the right implication is a consequence of the non-emptiness of $R \cap (T_{-i} \times \{t_i\})$.

B Proof of Theorem 1

We first introduce and prove the following lemma:

Lemma 2 *Let G a finite game in strategic form, $\boldsymbol{p} \in [0,1]^I$, $n \in \mathbb{N}$, $k \leq n$, and $\hat{\pi} \in (A \times \{k\}, \boldsymbol{p})\text{-}CE(nG)$. Let, $B = \langle (nA_i)_{i \in I}, (s_i)_{i \in I}, \hat{\pi} \rangle$ where $s_i(a_i, \alpha_i) = a_i$ for any $(a_i, \alpha_i) \in nA_i$, and any $i \in I$. B is a $\boldsymbol{p}$-rational belief system for G such that $\pi_B = \text{Proj}_A(\hat{\pi})$.*

Proof. Since $\hat{\pi}$ is a $(A \times \{k\}, \boldsymbol{p})$-correlated equilibrium of nG:

- From the incentive constraints we know that for any $i \in I$, $A_i \times \{k\} \subseteq R_i$.
- From the $\boldsymbol{p}$-belief constraints we know that for any $i \in I$, and any $(a_i, \alpha_i) \in nA_i$:

$$\hat{\pi}\left((A_{-i} \times \{\boldsymbol{k}_{-i}\}) \times \{(a, \alpha_i)\}\right) \geq p_i \hat{\pi}\left(nA_{-i} \times \{(a_i, \alpha_i)\}\right).$$

  Thus, we obtain that for any $(a_i, \alpha_i) \in nA_i$, $\hat{\pi}(R_{-i} \cap (nA_{-i} \times \{(a_i, \alpha_i)\})) \geq p_i \hat{\pi}(nA_i \times \{(a_i, \alpha_i)\})$, and therefore, that B is $\boldsymbol{p}$-rational.

Finally, $\pi_B = \text{Proj}_A(\hat{\pi})$ is an obvious identity. ■

Let's go on with the proof of the theorem. For $k \in \{0,1\}$; by $-k$ we denote the element complementary to k in this set.

$\subseteq$

Let $B = \langle (T_i)_{i \in I}, (s_i)_{i \in I}, f \rangle$ a $\boldsymbol{p}$-rational belief system for G. For each player $i \in I$ we can define:

$$\begin{array}{rccc} \beta_{k,i}: & T_i & \longrightarrow & 2A_i \\ & t_i \in R_i & \to & (s_i(i), k) \\ & t_i \notin R_i & \to & (s_i(i), -k) \end{array}$$

Now, for any $(a, \alpha) \in 2A$, let $\pi(a, \alpha) = f\left(\prod_{i \in I} \beta_{k,i}^{-1}(a_i, \alpha_i)\right)$. It is immediate that $\text{Proj}_A(\pi) = \pi_B$. Let's check that π is a $(A \times \{k\}, \boldsymbol{p})$-correlated equilibrium of $2G$. Let $i \in I$:

- Let $(a_i, k) \in A_i \times \{k\}$ and $(\bar{a}_i, \alpha_i) \in 2A_i$. Then:

$$\sum_{(a_{-i}, \alpha_{-i}) \in 2A_{-i}} \pi((a_{-i}, \alpha_{-i}); (a_i, k)) h_{2,i}((a_{-i}, \alpha_{-i}); (\bar{a}, \alpha_i)) =$$

$$= \sum_{t_i \in [\beta_{k,i}]^{-1}(a_i, k)} f\left(T_{-i} \times \{t_i\} \,|\, [\beta_{k,i}]^{-1}(a_i, k)\right) \sum_{a_{-i} \in A_{-i}} \mathbb{E}_B\left[h_i(a_{-i}; \bar{a}_i) \,|\, f_i(t_i)\right].$$

  Since by construction, $T_{-i} \times [\beta_{k,i}]^{-1}(a_i, k) \subseteq R_i$, (a_i, k) is maximizer of the above.

- Let $(a_i, \alpha_i) \in 2A_i$. Then:

$$\pi\left(\left(A_{-i} \times \{\boldsymbol{k}_{-i}\}\right) \times \{a_i, \alpha_i\}\right) =$$

$$= f\left(R_{-i} \cap \left(T_{-i} \times [\beta_{k,i}]^{-1}(a_i, \alpha_i)\right)\right) \geq p_i f\left(T_{-i} \times [\beta_{k,i}]^{-1}(a_i, \alpha_i)\right) = p_i \pi\left(2A_{-i} \times \{(a_i, \alpha_i)\}\right).$$

$\boxed{\supseteq}$

Just apply Lemma 2 to $n = 2$.

C Proof of Theorem 2

Nonemptiness follows from the fact that correlated equilibria always exist for any finite game G and constitute $\boldsymbol{p}$-rational outcomes for any $\boldsymbol{p} \in [0,1]^I$. Given that the set of $\boldsymbol{p}$-rational outcomes is a projection (under Proj_A) of the $(X, \boldsymbol{p})$-correlated equilibria of $2G$, with $X = A \times \{k\}$ a copy of the action space of the original game G, the remaining properties follow once they have been shown for the latter. This is what we do next. For the given game G, define the $(X, \boldsymbol{p})$-correlated equilibrium correspondence, where $X = A \times \{k\}$, $k \in \{0, 1\}$, is fixed:

$$\begin{array}{rccc} \rho: & [0,1]^I & \longrightarrow & \Delta(2A) \\ & \boldsymbol{p} & \to & (X, \boldsymbol{p})\text{-}CE(2G). \end{array}$$

Clearly ρ is convex- and compact-valued, it remains to show that it is also continuous. We do this by showing that it is upper- and lower-hemicontinuous (respectively, *uhc* and *lhc*) as a correspondence of $\boldsymbol{p}$.

$\boxed{\boldsymbol{uhc}}$

Since $2A$ is finite, $\Delta(2A)$ is compact, and hence upper-hemicontinuity is equivalent to showing that ρ has a closed graph. But this is immediate from inspection of the inequalities defining the sets $(X, \boldsymbol{p})$-$CE(2G)$. In particular, the inequalities are all weak inequalities, linear in $\boldsymbol{p}$. Moreover, the domain $[0,1]^I$ is compact.

$\boxed{\boldsymbol{lhc}}$

Denote by $\Gamma_\rho \subset [0,1]^I \times \Delta(2A)$ the graph of the correspondence ρ. Fix $(\boldsymbol{p}, \hat{\pi}) \in \Gamma_\rho$ and let $(\boldsymbol{p}^n)_n \subset [0,1]^I$ be a sequence converging to $\boldsymbol{p}$. We need to show that there exists a sequence $((\hat{\pi})^n)_n$ converging to $\hat{\pi}$ such that $(\hat{\pi})^n \in \rho(\boldsymbol{p}^n)$ for sufficiently large n. Take the point $(\boldsymbol{p}, \hat{\pi})$. Clearly this satisfies all inequalities defining $\rho(\boldsymbol{p})$, in particular also the $\boldsymbol{p}$-rationality constraints. Consider the following sequence $(\boldsymbol{p}^n, \hat{\pi})_n \subset [0,1]^I \times \Delta(2A)$. If for sufficiently large n the elements are contained in Γ_ρ we are done. So consider the case where they are not. Consider the family of projections $\Pi_\rho : [0,1]^I \times \Delta(2A) \longrightarrow [0,1]^I \times \Delta(2A)$ that map, for fixed $\bar{\boldsymbol{p}} \in [0,1]^I$, any element $(\bar{\boldsymbol{p}}, \bar{\pi}) \in [0,1]^I \times \Delta(2A)$ to the point in the set $\{\bar{\boldsymbol{p}}\} \times \rho(\bar{\boldsymbol{p}})$ that is closest to $(\bar{\boldsymbol{p}}, \bar{\pi})$. Since the sets $\rho(\cdot)$ are always nonempty, convex, compact polyhedra, we have that $\Pi_\rho(\boldsymbol{p}^n, \hat{\pi})$ is uniquely defined and moreover, $\Pi_\rho(\boldsymbol{p}^n, \hat{\pi}) \in \Gamma_\rho$ for all points in the sequence $(\boldsymbol{p}^n, \hat{\pi})_n$. It remains to show that the sequence $(\Pi_\rho(\boldsymbol{p}^n, \hat{\pi}))_n$ converges to the point $(\boldsymbol{p}, \hat{\pi})$.

Apart from the $\boldsymbol{p}$-belief constraints all other constraints defining $\rho(\boldsymbol{p})$ are independent of $\boldsymbol{p}$. Hence, if $(\boldsymbol{p}, \hat{\pi})$ satisfies those constraints, then so must any other point in the sequence $(\boldsymbol{p}^n, \hat{\pi})_n$. Therefore the only constraints that can be violated by elements of the sequence $(\boldsymbol{p}^n, \hat{\pi})_n$ are the $\boldsymbol{p}$-belief constraints. Consequently, any point in the sequence $(\Pi_\rho(\boldsymbol{p}^n, \hat{\pi}))_n$ lies on the boundary of the polyhedra defined by the $\boldsymbol{p}$-belief constraints. As mentioned, these constraints are linear in $\boldsymbol{p}$, and since they also define nonempty, convex, compact polyhedra, the sequence $(\Pi_\rho(\boldsymbol{p}^n, \hat{\pi}))_n$ indeed converges to $(\boldsymbol{p}, \hat{\pi})$. This shows the continuity of ρ and hence also of $\boldsymbol{p}$-$RO(G)$ in $\boldsymbol{p}$.

Finally, the claims that, for $\boldsymbol{p} = 0$, we have 0-$RO(G) = \Delta(A)$, and for $\boldsymbol{p} = 1$, we have 1-$RO(G) = CE(G)$, are immediate. To see that for any $\boldsymbol{p} \in [0, 1)$, we have $\dim[\boldsymbol{p}\text{-}RO(G)] = \dim[\Delta(A)]$, notice that the $(X, \boldsymbol{p})$-correlated equilibria with $X = A^1$ and $\boldsymbol{p} < 1$ entail distributions that put strictly positive weight on all strategies in A^2 as well as all convex combinations of such distributions. Projecting onto the original space $\Delta(A)$ implies distributions with strictly positive weights on all strategies in A as well as all possible convex combinations. This concludes the proof.

D Proof of Proposition 1

Fix G and let $A^n = \Pi_{i \in I} A_i^n$ denote the space of all pure strategy profiles that survive n rounds of iterated elimination of strictly dominated strategies in G, and similarly for the individual sets A_i^n. Let G^n denote the subgame of G with strategies restricted to A^n. Because G is finite, the limit sets A_i^∞, A^∞, and G^∞ are well defined (and are obtained after finitely many iterations). Also, for any subset $Y \subset A$, let $Y^c = A \setminus Y$ denote the complement of Y in A.

For any given $p \in [0, 1]$, let $\boldsymbol{p} \in [0, 1]^I$ be such that $\min \boldsymbol{p} \geq p$. We show that for p sufficiently close to 1, behavior is supported with high probability (p) in A^∞. Specifically, we construct a $\bar{p} < 1$ such that for any $p \in [\bar{p}, 1]$, if $\pi \in \boldsymbol{p}$-$RO(G)$, then $\pi((A^\infty)^c) \leq 1 - p$.

Consider the game $G^0 = G$ and pick some $p^1 < 1$. It immediately follows from p-rationality that for $p \in [p^1, 1]$, if $\pi \in \boldsymbol{p}$-$RO(G)$, we have $\pi\left(\left(A^1\right)^c\right) \leq 1 - p$.

Suppose now that the above statement is true for $n - 1$, namely there exists $p^{n-1} < 1$ such that for $p \in [p^{n-1}, 1]$, if $\pi \in \boldsymbol{p}$-$RO(G)$, then we have $\pi\left(\left(A^{n-1}\right)^c\right) \leq 1 - p$. We show that the statement also holds for n.

Fix the game G^{n-1}. It follows from finiteness of G and continuity of the payoffs that there exists $p^n \in [p^{n-1}, 1)$ such a strategy in $A^{n-1} \setminus A^n$ that is strictly dominated in G^{n-1} (by some strategy in G^{n-1} and hence in G) is also strictly dominated in G (by the same strategy) given a $\boldsymbol{p}$-rational belief system with $\min \boldsymbol{p} \geq p$ and $p \geq p^n$; (this follows from $p^n \geq p^{n-1}$, and because $\pi \in \boldsymbol{p}$-$RO(G)$ with $p \geq p^{n-1}$ implies $\pi\left(\left(A^{n-1}\right)^c\right) \leq 1 - p$). This implies that for any $p \in [p^n, 1]$ and any $\pi \in \boldsymbol{p}$-$RO(G)$, we also have $\pi((A^n)^c) \leq 1 - p$.

Finiteness of the game implies that the process ends after finitely many steps implying that indeed there

exists $p^\infty < 1$ such that for $p \in [p^\infty, 1]$ and any $\pi \in \boldsymbol{p}\text{-}RO(G)$, we have $\pi\left((A^\infty)^c\right) \leq 1 - p$. Taking $\bar{p} = p^\infty$ shows the claim.

E Proof of Theorem 3

We prove the first statement. The one concerning the $\boldsymbol{p}$-rational expectation of rational types becomes obvious after that proof. The statement is given for $k = 0$, and we suppose we are taking some player i_0's expectation.

$\subseteq$

Let $B = \left\langle (T_i)_{i \in I}, (s_i)_{i \in I}, f \right\rangle$ a $\boldsymbol{p}$-rational belief system for G, $i_0 \in I$, and $t_{i_0} \in T_{i_0}$. For any $i \in I \setminus \{i_0\}$, we define:

$$\begin{array}{rccc} \beta_{k,i}: & T_i & \longrightarrow & 3A_i \\ & t_i \in R_i & \to & \left(s_i\left(t_i\right), k\right) \\ & t_i \notin R_i & \to & \left(s_i\left(t_i\right), k+1 \bmod 3\right) \end{array}$$

and:

$$\begin{array}{rccc} \beta_{k,i_0}: & T_i & \longrightarrow & 3A_{i_0} \\ & t'_{i_0} \in R_{i_0} \setminus \{t_{i_0}\} & \to & \left(s_{i_0}\left(t_{i_0}\right), k\right) \\ & t'_{i_0} \notin R_{i_0} \cup \{t_{i_0}\} & \to & \left(s_{i_0}\left(t_{i_0}\right), k+1 \bmod 3\right) \\ & t'_{i_0} = t_{i_0} & \to & \left(s_{i_0}\left(t_{i_0}\right), k+2 \bmod 3\right) \end{array}$$

By an identical argument to the one in the first part of the proof of Theorem 1, we can conclude that $\hat{\pi} = f \circ \beta^{-1}$ is a $(A \times \{k\}, \boldsymbol{p})$-correlated equilibrium of G, and it is immediate that $\mathbb{E}_B\left[h_{i_0}\left(a_{-i_0}; s_{i_0}\left(t_{i_0}\right)\right) | f_{i_0}\left(t_{i_0}\right)\right]$ is exactly player i_0's expectation conditional on playing $\left(s_{i_0}\left(t_{i_0}\right), k+2 \bmod 3\right)$ induced by $\hat{\pi}$.

$\supseteq$

It is again a reduction of Lemma 2, this time to $n = 3$.

F Proof of Result in Remark 2

Following Aumann, [2, 3], for any $X = \prod_{i \in I} X_i \subseteq A$, we say that the family $(\pi_i)_{i \in I} \subseteq \left(\Delta\left(A\right)\right)^I$ is a $(X, \boldsymbol{p})$-*subjective correlated equilibrium* of G, if for any $i \in I$:

- For any $a'_i \in X_i$, the following incentive constraints are satisfied:

$$\sum_{a_{-i} \in A_{-i}} \pi_i\left(a_{-i}; a'_i\right)\left[h_i\left(a_{-i}; a'_i\right) - h_i\left(a_{-i}; a_i\right)\right] \geq 0, \text{ for any } a_i \in A_i,$$

- For any $a_i \in A_i$ the following p_i-belief constraint is satisfied:

$$\sum_{a'_{-i} \in X_{-i}} \pi_i \left(a'_{-i}; a_i\right) \geq p_i \sum_{a_{-i} \in A_{-i}} \pi_i \left(a_{-i}; a_i\right).$$

We denote the set of $(X, \boldsymbol{p})$-subjective correlated equilibria of game G by $(X, \boldsymbol{p})\text{-}SCE\left(G\right)$. Given $n \in \mathbb{N}$ and a n-game nG, we have the projection $\text{Proj}_{A^I} : \left(\Delta\left(nA\right)\right)^I \to \left(\Delta\left(A\right)\right)^I$, where for any $\left(\hat{\pi}_i\right)_{i \in I} \in \left(\Delta\left(nA\right)\right)^I$, $\text{Proj}_{A^I}\left(\left(\hat{\pi}_i\right)_{i \in I}\right) = \left(\text{Proj}_A\left(\hat{\pi}_i\right)\right)_{i \in I}$. Then, the proof of the identity:

$$\boldsymbol{p}\text{-}SRO\left(G\right) = \text{Proj}_{A^I}\left[\left(A \times \{0\}\right)\text{-}SCE\left(2G\right)\right]$$

is the same as the one for Theorem 1 after slight modifications (just add sub-indices where needed). To see that the above projections constitute the whole space, let $\left(a^i\right)_{i \in I} \subseteq A$, and for any $i \in I$, $\pi_i = 1_{\{a^i\}}$. Fix $k \in \{0, 1\}$, and define, for any $i \in I$, $\hat{\pi}_i = 1_{\left\{\left(\left(a^i_{-i}, k_{-i}\right); \left(a^i_i, -k\right)\right)\right\}}$. It is immediate that $\text{Proj}_A\left(\left(\hat{\pi}_i\right)_{i \in I}\right) = \left(\pi_i\right)_{i \in I}$. Now, let $i \in I$, then the incentive constraints are trivially satisfied, since $\hat{\pi}_i\left(2A_{-i} \times \left(A_i \times \{k\}\right)\right) = 0$. Moreover, the p_i-belief constraint is also satisfied, because regardless of i's action, the sums are on both sides 1 or 0. We conclude that $\left(1_{\{a^i\}}\right)_{i \in I} \in \text{Proj}_A\left(\left(A \times \{k\}, \boldsymbol{p}\right)\text{-}SCE\left(2G\right)\right)$ for any $\left(a^i\right)_{i \in I} \subseteq A$, so by convexity, $\text{Proj}_A\left(\left(A \times \{k\}, \boldsymbol{p}\right)\text{-}SCE\left(2G\right)\right) = \left(\Delta\left(A\right)\right)^I$.