Dynamic Incentives in Incompletely Specified Environments

Gabriel Carroll, Stanford University
gdc@stanford.edu
April 15, 2020

Abstract

Consider a repeated interaction where it is unknown which of various possible stage games will be played each period. This framework captures the logic of intertemporal incentives even though numeric payoffs to any strategy profile are indeterminate. A natural solution concept is ex post perfect equilibrium (XPE): strategies must form a subgame-perfect equilibrium for any realization of the stage game process. When (i) there is one long-run player and others are short-run, and (ii) public randomization is available, we can adapt the standard recursive approach to determine the maximum sustainable gap between reward and punishment. This leads to an explicit characterization of what outcomes are supportable in equilibrium, and an optimal penal code that supports any such outcome. Any non-XPE-supportable outcome fails to be an SPE outcome for some specific stage game process. In contrast to standard repeated games, restrictions (i) and (ii) are crucial.

*** PRELIMINARY VERSION. ***
*** PLEASE DO NOT CIRCULATE WITHOUT PERMISSION. ***

Thanks to (in random order) Drew Fudenberg, Andrzej Skrzypacz, Paul Milgrom, and Takuo Sugaya for discussions, as well as seminar audiences at UCSD and Columbia.
The author is supported by an NSF CAREER award.
1 Introduction

This paper studies a model of repeated interactions in which it is unknown what stage game will be played in each period. The stage game may vary from one period to the next, and there is no prior over the process determining its evolution.

Repeated games are the classic theoretical paradigm for studying how, and to what extent, non-myopic behavior can be incentivized by the promise of future rewards. The canonical model is well-understood: given any specification of the stage game and discount factor, one can, at least in principle, fully characterize the outcomes that can arise in equilibrium, as those that are supportable by punishing deviations using the worst possible SPE payoff (Abreu, 1988), which in turn can be identified using the recursive method of Abreu, Pearce and Stacchetti (1990) (henceforth APS). However, this canonical model makes the rigid assumption that players play exactly the same game over and over. In reality, the nature of interactions between players may vary over time, and the players may not be able to precisely describe or agree on its future evolution. We might expect that the basic logic underlying dynamic incentives—that, say, a player can be induced to forego a short-run temptation of 3 payoff units if he can be promised a reward of $3/\delta$ in the future—should remain valid even when the future evolution of the environment cannot be fully specified. It is therefore natural to explore a model of this sort to try to express such intuitions. And once we have formulated such a model, it is natural to ask more generally to what extent central results and tools from the standard theory carry over to this more general setting.

An important feature of our setting is that a strategy profile does not determine a numeric payoff for each player, since the payoffs depend on what stage games arise. Much of the usual analysis of repeated games, such as the recursive characterization of APS, is done in the space of continuation payoffs; the idea that “player 1 has foregone a temptation of 3” is actually expressed as “player 1’s continuation value has changed from 500 to 500 + $3/\delta$.” Hence, another perspective on the analysis performed here is that it sheds some light on the extent to which standard intuitions and results for repeated games depend on being able to talk about explicit continuation values.

Our focus here will be on repeated interactions with a fixed discount factor $\delta < 1$, and the following four features:

(i) One long-run strategic player interacts with a series of short-run players (as in e.g. Fudenberg, Kreps and Maskin (1990)).
(ii) A public randomization device is available.

(iii) Attention is restricted to pure strategies (conditional on the public randomization).

(iv) Actions are perfectly observed.

We will show how, in this class of environments, the recursive method can be adapted and used to characterize what outcomes are supportable in equilibrium.

Feature (i) is of course restrictive, but it still gives plenty of scope for studying the basic question of how to sustain non-myopic behavior through future incentives. As usual in the literature, this formulation allows multiple interpretations: the short-run players may be different individuals in each period, they may be long-lived but completely impatient individuals, or they may be a continuum of players all facing the same action by the long-run player, whose individual deviations are not detectable. This assumption can be varied somewhat; for example, the techniques developed here should also be applicable when there are multiple long-run players, and all but one have commitment power. The crucial feature is that there is only one player who needs to be given dynamic incentives. As will be explained below, the analysis does not extend to multiple such players.

Assumption (ii) is also crucial. This is in contrast to the usual setting of repeated games, where allowing for public randomization is mostly a technical convenience (for example, the APS recursive analysis can be carried out with or without public randomization). Here, this assumption is a substantive necessity, as described below.

Assumptions (iii) and (iv) do not seem to be as essential. They are made mostly for simplicity of exposition. The goal is to describe what outcomes are supportable, and this is easiest to do when an “outcome” consists simply of a sequence of stage game and action profiles. Without (iii) and (iv), the description of an outcome would have to condition on realizations of mixing or of the monitoring signal, leading to much more notational overhead.

The interaction assumed here has the following structure: There is a known set of stage games that may arise. Each period, the players all observe what stage game comes up. They then choose their actions and receive their payoffs. The long-run player maximizes the $\delta$-discounted sum of stage payoffs. Although the process governing the stage games in each period is unspecified, it is common knowledge that future stage games do not depend on players’ past actions (unlike in stochastic games).

The solution concept proposed here is ex-post perfect equilibrium (XPE), which requires that a player should never have any incentive to deviate at any history, no matter
what future stage games may come up. In other words, for any fixed sequence of realized stage games, the strategies should form a subgame-perfect equilibrium. We wish to characterize what outcomes can be supported by XPE strategies.

Note that this solution concept is not rooted in individual maximization, since a player’s payoff from a given strategy profile is not a single number; the payoff depends what stage games get realized. In principle, one could demand that if players have uncertainty about future stage games, they should have beliefs (and higher-order beliefs) about them, or at least have preferences about acts mapping stage game processes to outcomes, and maximize those preferences. One possible interpretation of XPE is that it gives players a simple way to coordinate on self-enforcing strategies, without needing to think about each other’s beliefs. An alternative interpretation is that the players do know the stage game process, and the analyst is uncertain, and she would like to confidently predict that some amount of non-myopic behavior is possible (and specify strategies that will do it) without needing to know more.

The key analytic technique is to adapt the recursive characterization from APS (or rather, a simple “one-dimensional” version of it, given assumption (i)) to our setting. We give a recursive characterization of the set of values of \( w \) such that it is possible to find two XPE profiles, one “reward” and one “punishment,” such that the reward gives the long-run player a payoff at least \( w \) more than the punishment no matter what stage games are realized. Thus, instead of recursing on continuation values, we recurse on the reward-punishment gap.

This leads to our first main result, a characterization of the outcomes that can be supported in XPE. Essentially, these are the outcomes in which, at each period, the “debt” owed to the long-run player for forgoing short-run gains in past periods never accumulates beyond the maximum sustainable reward-punishment gap (and, of course, the short-run players are best-responding). A special case of this result applies when only one stage game is possible, in which case the result characterizes the SPE outcomes of a traditional repeated game with only one long-run player; this description does not seem to exist in the literature and may be independently worthwhile.

Although this result is cast as describing the supportable outcomes for a given specification of uncertainty, we can equivalently view it as characterizing the forms of uncertainty that are consistent with sustaining a particular outcome. For example, in the “product-quality game” traditionally used as an example of an interaction between long-run and short-run players, we can use this to ask: what must the consumers believe about the firm’s future opportunities in order to be persuaded by the firm’s claim that it has enough
incentives to provide high quality in the present period?

A second result, which falls out of the proof of the first, is the existence of an optimal penal code that can be used as punishment to support any XPE outcome, as in Abreu (1988). This optimal penal code gives the long-run player his worst outcome among all XPE, *no matter* what stage games are realized. These two results together illustrate how classic ideas from repeated games successfully carry over to our framework.

Because, as observed above, the XPE solution concept is not based on individual maximization, one might well ask about its positive content. In particular, it is clear that any outcome is supportable in XPE can be achieved no matter what players believe about the stage game process. But is there any reason that the same isn’t true for outcomes that are *not* supportable in XPE? And if it is, then why should we accept XPE as a useful tool for studying the power of dynamic incentives? Our third result shows that, in fact, for any outcome not supportable in XPE, there is a particular stage game process for which it would not even be supportable in SPE. This provides a “positive foundation” for studying the set of XPE-supportable outcomes. The proof relies directly on the characterization in the first result.

As mentioned above, our recursive analysis relies both on having a single long-run player and on the availability of public randomization. Dropping either of these assumptions would lead the analysis of this paper to break down. Although it seems difficult to formalize an assertion about the nonexistence of any kind of recursive characterization, we can show concretely that the theory fails to carry over by demonstrating that optimal penal codes can fail to exist when either assumption is dropped. Since the optimal penal code plays a central role in the proof of the characterization of supportable outcomes, this suggests that such a characterization in general, if it exists, would have to look quite different. Section 5 provides examples documenting that dropping either assumption can result in nonexistence of optimal penal codes, and offers some discussion of why both assumptions are crucial to the recursive technique. This suggests that the usual analysis of repeated games relies on specific assumptions in the canonical model that go beyond the more basic idea of inducing non-myopic actions via future rewards.

The rest of the paper proceeds linearly: model, analysis, results, discussion. Literature will be discussed as it comes up.
2 Model

We begin by developing the model in terms of a standard repeated-game setup. The notation will largely follow Mailath and Samuelson (2006), suitably adapted for our framework of uncertainty. We then introduce some adaptations to notation that will be a bit more convenient for our focus on a single player's long-run incentives.

2.1 Standard formulation

There are \( n \geq 2 \) players. Player 1 is a long-run player, and the others are short-run players. As usual, we can be agnostic as to whether player \( i > 1 \) in period \( t \) is physically the same person (or persons) as player \( i \) in period \( t' \), but it is notationally simpler to use the same label \( i \) for both. There is a nonempty set \( \mathcal{G} \) of possible stage games. In any stage game \( G \in \mathcal{G} \), we denote the set of actions available to player \( i \) as \( A_i(G) \). We assume that actions are labeled so that \( A_i(G) \) and \( A_i(G') \) are disjoint for \( G \neq G' \); this makes the definitions up front slightly cumbersome but will simplify notation later. Write \( A(G) = \times_{i=1}^n A_i(G) \). Also, write \( A_i = \cup_{G \in \mathcal{G}} A_i(G) \) for the set of all actions that \( i \) can ever play, and likewise \( A = \cup_{G} A(G) \). Then, player \( i \)'s stage payoff function is simply written \( u_i : A_i(G) \to \mathbb{R} \). We assume a uniform bound \( M \) on the possible stage payoffs: \( |u_i(a)| \leq M \) for all \( i, a \). All these objects are exogenously given primitives. We assume that each \( A_i(G) \) is a compact metric space, and that \( u_i(a) \) is continuous on \( A(G) \) for each \( G \). (Finite action sets are a special case.) We equip \( A_i \) and \( A \) with their disjoint union topologies.

In the repeated game, in each period \( t = 0, 1, 2, \ldots \), the players observe the realized stage game \( G^t \in \mathcal{G} \), as well as the public randomization signal \( \omega^t \sim U[0, 1] \), and then they simultaneously choose actions. Thus, a history at time \( t \) consists of the stage games, public random signals, and actions at past dates, together with the stage game and random signal at the present date. So the set of time-\( t \) histories is

\[
H^t = (\cup_{G \in \mathcal{G}} (\{G\} \times [0, 1] \times A(G)))^t \times (\mathcal{G} \times [0, 1])
\]

with representative element

\[
h^t = (G^0, \omega^0, a^0; G^1, \omega^1, a^1; \ldots; G^{t-1}, \omega^{t-1}, a^{t-1}; G^t, \omega^t).
\]

We focus on pure strategies; thus, a strategy for player \( i \) is a measurable function \( s_i : \cup_{t=0}^\infty H^t \to A_i \), such that \( s_i(h^t) \in A_i(G^t) \) whenever the history \( h^t \) ends in \( G^t \). A strategy
profile takes the form \( s = (s_1, \ldots, s_n) \), or can be equivalently written \( s : \cup_i H^t \to A \), with the corresponding restriction \( s(h^t) \in A(G^t) \).

We refer to a realization of the sequence of stage games as an environment, \( E = (G^0, G^1, \ldots) \). A history \( h^t \) is consistent with the environment \( E \) if the stage games appearing at all periods \( 0, 1, \ldots, t \) in \( h^t \) are the same as those specified in \( E \). Given a strategy profile \( s \), an environment \( E \), and a history \( h^t = (G^0, \omega^0, a^0; \ldots; G^t, \omega^t) \) that is consistent with \( E \), we define subgame payoffs as follows. For any realization path \((\omega^{t+1}, \omega^{t+2}, \ldots)\) for the subsequent random signals, we can recursively define the action profiles \( a^t' = s(G^0, \omega^0, a^0; \ldots; G^t', \omega^t') \) for each \( t' \geq t \). Then, player 1’s subgame payoff at \( h^t \) is the (normalized) discounted sum of stage payoffs

\[
U_1(s|E, h^t) = (1 - \delta) \mathbb{E} \left[ \sum_{t' = t}^{\infty} \delta^{t'-t} u_1(a^t') \right],
\]

where the expectation is due to the public randomization. Player \( i \)'s payoff, for each \( i > 1 \), is simply

\[
U_i(s|E, h^t) = u_i(a^t).
\]

Given environment \( E \), strategy profile \( s \) is a subgame-perfect equilibrium (SPE) for \( E \) if, for each player \( i \), each history \( h^t \) consistent with \( E \), and each alternative strategy \( s_i' \),

\[
U_i(s|E, h^t) \geq U_i(s_i', s_{-i}|E, h^t). \tag{2.1}
\]

The usual arguments for the one-shot deviation principle apply: it suffices to have (2.1) hold for all \( h^t \) consistent with \( E \) and all \( s_i' \) that differ from \( s_i \) only at the history \( h^t \).

We can also define player 1’s continuation payoff in environment \( E \), following a history \( h^t \) consistent with \( E \) and an action profile \( a^t \), as

\[
U_1(s|E, h^t, a^t) = (1 - \delta) \mathbb{E} \left[ \sum_{t' = t+1}^{\infty} \delta^{t'-(t+1)} u_1(a^t') \right],
\]

where the expectation is over all the public random signals \((\omega^{t+1}, \omega^{t+2}, \ldots)\), and the future actions are determined by beginning from \( h^t \) followed by \( a^t \) and then playing according to \( s \). This quantity is not part of the definition of SPE, but it is relevant to player 1’s incentives to deviate: (2.1) is satisfied for player 1 at \( h^t \) if and only if

\[
(1 - \delta) u_1(s(h^t)) + \delta U_1(s|E, h^t, s(h^t)) \geq (1 - \delta) u_1(a'_1, s_{-1}(h^t)) + \delta U_1(s|E, h^t, (a'_1, s_{-1}(h^t)))
\]
for all deviations $a'_1$. Similarly, we can define

$$U_1(s|E) = (1 - \delta) \mathbb{E} \left[ \sum_{t=0}^{\infty} \delta^t u_1(a^t) \right],$$

the expected payoff from the beginning of the game in environment $E$.

Strategy profile $s$ is an ex-post perfect equilibrium (XPE) if it is an SPE for every environment.\(^1\) Later, we will indicate sufficient conditions on primitives to ensure that an XPE exists.

### 2.2 More convenient notation

We can apply a standard simplification for games with short-run players (e.g. Fudenberg, Kreps and Maskin, 1990): For each $G \in \mathcal{G}$, let $A^*(G)$ be the set of action profiles at which no short-run player wishes to deviate,

$$A^*(G) = \{a \in A(G) | u_i(a) \geq u_i(a'_i, a_{-i}) \text{ for all } i > 1, a'_i \in A_i(G)\}.$$  

Evidently, the constraints (2.1) for the short-run players are satisfied iff $s(h^t) \in A^*(G^t)$ for all histories $h^t$ (consistent with the environment $E$).

With this in mind, we can now dispense with explicit consideration of the short-run players’ incentives, focusing only on the long-run player. We accordingly drop the player subscript for payoffs: henceforth, we write $u$ and $U$ rather than $u_1$ and $U_1$ unless there is ambiguity.

We can summarize the above as

**Lemma 2.1.** Strategy profile $s$ is an XPE if and only if both the following conditions hold:

1. for every history $h^t$, $s(h^t) \in A^*(G^t)$;

2. for every environment $E$, every history $h^t$ consistent with $E$, and every possible deviation $s'_1$ by player 1 that differs from $s_1$ only at history $s'_1$, we have $U(s|E, h^t) \geq U(s'_1, s_{-1}|E, h^t)$.

\(^1\)The terminology is inspired by that of Fudenberg and Yamamoto (2010), who study a repeated game in which the stage game is fixed over time but unknown; their equilibrium concept requires subgame-perfection for each such game. Some literature has used the name “belief-free equilibrium” for related concepts, e.g. Ely, Hörner and Olszewski (2005); Hörner and Lovo (2009).
Notice that the set of XPE has a recursive structure: s is an XPE if it meets conditions (1)–(2) at every period-0 history and each continuation strategy profile starting from date 1 is an XPE.

In addition, when \( a \in A(G) \), let us write \( \hat{u}(a) = \max_{a'_1 \in A_1(G)} u(a'_1, a_{-i}) \) for the stage payoff that would result from the myopically optimal deviation from \( a \) (where, here and henceforth, we take “myopically optimal deviation” to mean “conforming” when 1’s action is already a best reply in the stage game). Clearly \( \hat{u}(a) \geq u(a) \), and \( \hat{u} \) is again continuous on \( A(G) \). Although it makes no difference formally, a conceptual reframing may be helpful: rather than think of action profiles as consisting specifically of an action by each player, and contemplating explicit deviations by player 1, we may think of action profiles (that may arise in equilibrium) in a stage game \( G \) simply as abstract objects belonging to a set \( A^*(G) \), and focus on \( \hat{u}(a) \) as the quantity relevant to player 1’s incentive to deviate.

Finally, if \( E = (G^0, G^1, G^2, \ldots) \), it will be useful to write \( E^-t = (G^t, G^{t+1}, \ldots) \), the continuation environment starting in period \( t \), and to further abbreviate \( E^-1 \) as simply \( E^- \).

3 Analysis

3.1 Recursive technique

Player 1 can be dissuaded from a deviation that earns a short-term gain of \( g \) only if doing so reduces the continuation payoff by at least \( \frac{1-\delta}{\delta} g \) in every possible environment. This suggests trying to find the largest “gap” \( w \geq 0 \) such that there exist two XPE’s, say \( \bar{s} \) and \( \tilde{s} \), such that \( U(\bar{s}|E) - U(\tilde{s}|E) \geq w \) for every environment \( E \); doing so then lets us rule out some action profiles because deviation cannot be prevented.

We adapt the recursive machinery from APS to describe the set of such values \( w \). Their \( B \) operator for \( n \)-player games maps subsets of \( \mathbb{R}^n \) to subsets of \( \mathbb{R}^n \). Here, we are concerned only with one long-run player, so the recursion is done on subsets of \( \mathbb{R} \). Moreover, public randomization makes our set convex, hence an interval, and its lower bound is zero. So we only need to keep track of the upper bound, i.e. a single number.

With this in mind, we first define, for any \( w \geq 0 \) and \( G \in \mathcal{G} \),

\[
A^*(G, w) = \left\{ a \in A^*(G) \mid \hat{u}(a) - u(a) \leq \frac{\delta}{1-\delta} w \right\}. 
\]
Now define

\[ B(w; G) = (1 - \delta) \left( \max_{a \in A^*(G,w)} u(a) - \min_{a' \in A^*(G,w)} \tilde{u}(a') \right) + \delta w. \quad (3.1) \]

(If \( A^*(G, w) \) is empty, then take \( B(w; G) = -\infty \). Note that as long as \( A^*(G, w) \) is nonempty, it is closed, and the max and min exist by continuity.)

Intuitively, this \( B(w; G) \) represents the largest possible gap in 1’s payoff between two different strategy profiles, given that \( G \) is played at date 0, the date-0 incentive constraints must be satisfied, and all continuation payoffs starting from date 1 must lie within an interval of width \( w \). Indeed, these last two requirements together imply that both profiles must specify an action in \( A^*(G; w) \) at date 0. Moreover, the payoff from following the “bad” strategy profile cannot be less than the payoff from a date-0 deviation; thus the payoff gap between the good and bad strategy profiles is at most the gap between conforming to the good profile and deviating from the bad profile. Decomposing this gap into its period-0 component and its continuation component produces the two terms on the right side of (3.1).

The above argument sketches why the expression in (3.1) is an upper bound on the payoff gap between two strategy profiles, and suggests how to attain it: Normalizing the interval of allowable continuation payoffs to \([0, w]\), specify that the “good” profile begins with the \( a \) attaining the max in (3.1) and promises a continuation payoff of \( w \) if 1 conforms; the “bad” profile begins with the \( a' \) attaining the min and promises a continuation payoff of 0 if 1 deviates. To ensure the correct gap in the on-path payoffs, the continuation payoff after conforming in the “bad” profile should be set so that 1 is indifferent between initially conforming and deviating. This can indeed be done (the fact that \( a' \in A^*(G, w) \) ensures that this continuation is at most \( w \)). Note that public randomization is essential for this, as it ensures that the set of allowable continuation payoffs is an interval. We will revisit this point in Section 5.

Now define

\[ B(w) = \inf_{G \in \mathcal{G}} B(w; G). \]

This is the maximum payoff gap that can be guaranteed regardless of what stage game arrives in the initial period, given that continuation payoffs lie in an interval of width \( w \).

Notice that \( B(w; G) \) is strictly increasing in \( w \) at a rate of at least \( \delta \) (the first term of (3.1) is weakly increasing because \( A^*(G; w) \) is increasing in \( w \), and the second term is clearly increasing at rate \( \delta \)). Therefore, \( B(w) \) is as well.
We now adopt an assumption that will be maintained for the rest of the paper:

**Assumption 3.1.** There exists \( w \geq 0 \) such that \( B(w) \geq w \).

As we shall see, this assumption will imply that an XPE exists (and in fact, the converse is also true).

As an aside, either of the following sufficient conditions on primitives implies that Assumption 3.1 is satisfied:

1. For every \( G \in \mathcal{G} \), there exists \( a \in A^*(G) \) such that \( \bar{u}(a) = u(a) \) (i.e. a stage Nash equilibrium).
   
   (This ensures the assumption holds with \( w = 0 \).)

2. There exists \( \epsilon > 0 \) such that, for every \( G \in \mathcal{G} \), there exist \( a, a' \in A^*(G) \) with \( u(a) \geq \bar{u}(a') + \epsilon \), and \( \delta \geq \frac{2M}{2M+\epsilon} \).
   
   (In this case, \( A^*(G; \epsilon) = A^*(G) \) for all \( G \), and then \( B(\epsilon; G) \geq \epsilon \) for all \( G \), so we can take \( w = \epsilon \).)

However, rather than adopt either of these, we will just make Assumption 3.1 directly.

Let \( w^* \) be the largest value such that \( B(w) \geq w \). It is straightforward that this maximum indeed exists, and that in fact \( B(w^*) = w^* \).

This \( w^* \) is the limiting value of a recursion. To show this, we need a continuity argument (our analogue to Theorem 5 of APS):

**Lemma 3.2.** The functions \( B(w; G) \) and \( B(w) \) are right-continuous in \( w \).

(The proof of this result, and all others not given in the text, are in Appendix A.)

With this property, one can readily show that starting with a value of \( w \) large enough to be an upper bound for \( w^* \), for example any \( w_0 > 2M \) (note that indeed \( w > 2M \) implies \( B(w; G) < w \) for each \( G \)), and then iterating \( B \) gives a decreasing sequence that converges to \( w^* \). However, for technical reasons, it will be useful to take a slightly different sequence, one in which \( w_{k+1} \) is strictly above \( B(w_k) \). Specifically:

**Lemma 3.3.** Define a sequence as follows: \( w_0 > 2M, w_1 \in (B(w_0), w_0) \), and for \( k = 2, 3, \ldots \), put \( w_k = (B(w_{k-1}) + B(w_{k-2}))/2 \). Then:

1. \( w_0 > w_1 > w_2 > \cdots \);

2. \( w_k > B(w_{k-1}) \) for \( k \geq 1 \);
3. \( w_k \to w^* \).

We can now show that there is no way to guarantee a payoff gap between two different XPE’s of more than \( w^* \). In fact, a stronger statement is true: For any \( \epsilon > 0 \), we can find an “adversarial” environment such that, in this environment, even if any SPE is allowed, the largest and smallest attainable payoffs differ by less than \( w^* + \epsilon \).

**Lemma 3.4.** Given any \( \epsilon > 0 \), there exists a finite \( T \) and a sequence of stage games \( G^0, G^1, \ldots, G^T \in \mathcal{G} \) with the following property: For any environment \( E \) that begins with stage games \( G^0, \ldots, G^T \), and any two SPE’s \( \bar{s} \) and \( \bar{s} \) for this environment,

\[
U(\bar{s}|E) - U(s|E) < w^* + \epsilon.
\]

The proof uses the sequence from Lemma 3.3. We show by induction that there is an adversarial environment that prevents the payoff gap from exceeding \( w_k \). In particular, since \( w_k > B(w_{k-1}) \), we can choose a stage game \( G \) such that \( B(w_{k-1}; G) < w_k \). Then, if \( G \) is played in the initial period, and subsequent periods feature the sequence of stage games that prevents a gap of more than \( w_{k-1} \) (which exists by the induction hypothesis), then the total payoff gap cannot exceed \( w_k \).

**Proof of Lemma 3.4.** For each \( k = 1, 2, \ldots \), let \( \overline{G}_k \in \mathcal{G} \) be such that \( B(w_{k-1}; \overline{G}_k) < w_k \); this exists by Lemma 3.3 part 2. We will show that in any environment that begins with the stage games \( \overline{G}_k, \overline{G}_{k-1}, \ldots, \overline{G}_1 \) (in that order), the payoffs from any two SPE’s differ by less than \( w_k \). Since \( w_k \to w^* \), the lemma then follows, by taking \( k \) large enough relative to \( \epsilon \).

We prove the statement by induction on \( k \). The base case \( k = 0 \) is trivial, since in any environment at all, the payoffs of any two action profiles within a stage differ by at most \( 2M < w_0 \), and therefore the same is true for the payoffs of any two SPE’s. Now suppose the statement holds for \( k - 1 \). Consider an environment \( E \) beginning with \( \overline{G}_k, \overline{G}_{k-1}, \ldots, \overline{G}_1 \).

Let \( s \) be any SPE. Let \( a^0 \) be the action profile played at some date-0 history \( h^0 = (\overline{G}_k, \omega^0) \), and \( a'_1 \) be player 1’s myopically optimal deviation; the incentive constraint reads

\[
(1 - \delta)u(a^0) + \delta U(s|E, h^0, a^0) \geq (1 - \delta)u(a'_1, a^0_{-1}) + \delta U(s|E, h^0, (a'_1, a^0_{-1}))
\]

or, rearranging,

\[
(1 - \delta)(u(a'_1, a^0_{-1}) - u(a^0)) \leq \delta(U(s|E, h^0, a^0) - U(s|E, h^0, (a'_1, a^0_{-1}))).
\]
The left side is \((1 - \delta)(\hat{u}(a^0) - u(a^0))\), while the right side is \(\delta\) times the difference of two SPE payoffs in the continuation environment \(E^-\), and so is less than \(\delta w_{k-1}\) by the induction hypothesis. Hence, \(a^0\) must lie in \(A^*\(\overline{G}_k, w_{k-1}\)\). That is, only action profiles in \(A^*\(\overline{G}_k, w_{k-1}\)\) can be played at date 0 in SPE.

Now let \(\overline{s}, \overline{s}\) be two different SPE’s. The payoff from \(\overline{s}\) is

\[
\mathbb{E}[(1 - \delta)u(a^0) + \delta U(\overline{s}|E, \overline{G}^k, \omega^0, a^0)]
\]

(where the expectation is over the random signal \(\omega^0\) and resulting action profile \(a^0\))

\[
\leq (1 - \delta) \max_{a \in A^*\(\overline{G}_k, w_{k-1}\)} u(a) + \delta \sup_{s' \text{ is SPE for } E^-} U(s'|E^-).
\]

Likewise, the payoff from \(\overline{s}\) is at least the payoff from deviating to the myopically action \(a'_1\) in date 0, which is

\[
\mathbb{E}[(1 - \delta)\hat{u}(a^0) + \delta U(\overline{s}|E, \overline{G}^k, \omega^0, (a'_1, a^0_{-1}))]
\]

(note that \(a^0\) is now determined by \(\overline{s}\) instead of \(\overline{\pi}\))

\[
\geq (1 - \delta) \min_{a \in A^*\(\overline{G}_k, w_{k-1}\)} \hat{u}(a) + \delta \inf_{s' \text{ is SPE for } E^-} U(s'|E^-).
\]

Subtracting, and using the fact that two different SPE payoffs in environment \(E^-\) differ by at most \(w_{k-1}\) by induction, gives us exactly

\[
U(\overline{s}|E) - U(\overline{s}|E) \leq B(w_{k-1}, \overline{G}_k).
\]

Since this is less than \(w_k\), the desired statement follows.

\[
\square
\]

This result partially justifies an understanding of \(w^*\) as the largest reward-punishment gap that can be sustained in XPE. We say “partially” because it shows that a higher gap cannot be sustained, but it does not show that \(w^*\) is attainable; this will follow from Section 3.3.

As a consequence of the preceding analysis, we can return to make good on the promise at the beginning of this section, to rule out some actions where deviation is too tempting:

**Lemma 3.5.** In any XPE, at any history \(h^t\) ending in a current stage game \(G^t\), the action
profile played must be in $A^*(G^i, w^*)$.

Proof. It suffices to prove this for date-0 histories. Consider any initial stage game $G$ and any $\epsilon > 0$. Consider any environment $E$ that begins with $G$ followed by the finite sequence of stage games given by Lemma 3.4. For any SPE $s$ for this environment, any action profile $a^0$ played at date 0 must satisfy

$$(1 - \delta)(\hat{u}(a^0) - u(a^0)) \leq \delta (w^* + \epsilon),$$

by the same logic used in the proof of Lemma 3.4 (and the fact that continuation payoffs of two different SPE’s from period 1 onward differ by at most $w^* + \epsilon$).

Therefore, if $s$ is an XPE, then at any date-0 history with any stage game $G^0$, the action profile to be played must satisfy $\hat{u}(a^0) - u(a^0) \leq \frac{\delta}{1 - \delta} (w^* + \epsilon)$. Since $\epsilon > 0$ is arbitrary, the right side can be replaced by $\frac{\delta}{1 - \delta} w^*$, giving the desired result.

As a side observation, there may be nontrivial interactions between the different stage games in $G$ in determining the value of $w^*$. That is: Suppose that for each $G \in \mathcal{G}$, we define $w^*(G)$ as the highest fixed point of $w \mapsto B(w; G)$ (thus, we repeat the process above, but in a world in which the set of possible stage games consists only of $G$). Then, $w^*$ may be bounded strictly below all of the $w^*(G)$. This also implies that the adversarial environments constructed in Lemma 3.4 may need to have the stage game vary from one period to the next.

Example 3.1. Suppose there are two possible stage games, with $n = 2$, as shown in Figure 1, and suppose $\delta = 3/4$. The two games are called $G$ and $G'$. Part (a) of the figure illustrates them in the standard matrix form, whereas part (b) rearranges them to a form more suitable for us, by showing the action profiles in $A^*(G)$ and $A^*(G')$, and the player-1 payoffs $u$ and $\hat{u}$ for each.

It is straightforward to check that the $B(\cdot; \cdot)$ functions are given by

$$B(w; G) = \begin{cases} 
\frac{2}{4} + \frac{3}{4} w & \text{if } w \geq \frac{1}{3}, \\
-\infty & \text{if } w < \frac{1}{3}
\end{cases}$$

and

$$B(w; G') = \begin{cases} 
\frac{3}{4} + \frac{3}{4} w & \text{if } w \geq \frac{7}{3}, \\
\frac{1}{4} + \frac{3}{4} w & \text{if } w < \frac{7}{3},
\end{cases}$$
In particular, their highest fixed points are $w^*(G) = 2$ and $w^*(G') = 3$ respectively. Part (c) of the figure shows these two functions ($B(w, G)$ as the solid line, $B(w, G')$ as dashed) and their respective intersections with the $45^\circ$ line. However, $w^*$ is the largest value for which the lower of the two functions crosses the $45^\circ$ line, and this happens only at $w^* = 1$. \[ \triangle \]

### 3.2 Quasi-minmax payoffs

Lemma 3.5 leads to bounds on the payoffs that can arise in any XPE. In particular, for each stage game $G$, let us pick “most effective reward” and “most effective punishment” action profiles

$$\pi(G) \in \arg\max_{a \in A^*(G, w^*)} u(a); \quad q(G) \in \arg\min_{a \in A^*(G, w^*)} \hat{u}(a).$$

(As before, these exist, by compactness, and by the fact that $B(w^*) \neq -\infty$ implying $A^*(G, w^*)$ is nonempty.)

The latter give a lower bound on payoffs in XPE. For any environment $E = (G^0, G^1, \ldots)$, define

$$U(E) = (1 - \delta) \sum_{t=0}^{\infty} \delta^t \hat{u}(a(G^t)).$$

**Lemma 3.6.** If $s$ is an XPE, then for any environment $E$,

$$U(s|E) \geq U(E).$$

**Proof.** Fix the environment $E$. Suppose that players 2, \ldots, $n$ follow the strategy $s$, whereas 1 simply plays the myopically optimal deviation at each history. By Lemma 3.5, at each period $t$, regardless of the past history, $s$ specifies playing an action in $A^*(G^t, w^*)$. Therefore, by myopically deviating, player 1 gets a payoff of at least $\hat{u}(a(G^t))$ in this period. Summing across all periods shows that 1’s payoff from the repeated deviation is at least $U(E)$. Hence, the payoff from conforming to $s$ is at least this much. $\square$

We can think of $\hat{u}(a(G))$ as a “quasi-minmax” payoff for player 1 when the stage game is $G$, providing a straightforward lower bound on 1’s equilibrium payoffs. Although it involves a minimum over action profiles of 1’s myopic best-reply payoff, it differs from the usual minmax in two ways. First, the min is taken only over a restricted set of action
Figure 1: An example where the sustainable reward-punishment gap \( w^* \) is lower than the gap sustained by any individual stage game.
profiles, those in $A^*(G, w^*)$. This is natural; because we are not in a folk theorem setting but are considering a fixed $\delta$, some action profiles are ruled out as unsustainable. And second, the action profile that actually produces the stage payoff of $\tilde{u}(a(G))$ typically cannot be played in equilibrium, as it does not satisfy the incentive constraints of the short-term players. This is again familiar from the literature on repeated games with short-term players, such as Fudenberg, Kreps and Maskin (1990) (though they nonetheless use the term “minmax value” for the analogous quantity).

Now that we have a lower bound on XPE payoffs, our next step is to develop an XPE strategy profile whose payoffs exceed this bound by a controlled amount.

### 3.3 Automaton strategies

We present the strategy profile in the form of an automaton, as in (Mailath and Samuelson, 2006, Section 2.3). The automaton enters each period $t$ in some state. After the stage game $G^t$ and public random signal $\omega^t$ are realized, the automaton specifies the action profile $a^t$ to be played, and then the automaton transitions to a new state for period $t+1$ depending on the actions observed. In fact, since dynamic incentives are irrelevant for players 2, ..., $n$, we can focus on state transitions that depend only on 1’s action.

More specifically, we consider an automaton whose state space is the interval $W = [0, w^*]$. The state $w \in W$ is to be interpreted as a promise that the payoff will exceed the lower bound $U(E)$ by exactly $w$. The main elements to be specified are the action (output) function $f : W \times G \times [0, 1] \to A$ (which, of course, must output an action in $A(G)$ when the input involves stage game $G$) and the state transition function $\tau : \cup_{G \in \mathcal{G}} (W \times \{G\} \times [0, 1] \times A(G)) \to W$. These objects, together with a choice of initial state $w \in W$, determine a strategy profile in the natural way.

For any $G \in \mathcal{G}$, define

$$\lambda(G) = \frac{1}{(1 - \delta)(u(\tilde{u}(G)) - \tilde{u}(a(G))) + \delta w^*}.$$ 

The denominator equals $B(w^*; G) \geq B(w^*) = w^*$, so we have $\lambda(G) \in [0, 1/w^*]$.

Now, for any $w, G, \omega, a$:

---

2Section 5.7 of Mailath and Samuelson (2006) develops automata for dynamic games. The formalism used there is in some sense closer to our setting, since actions and transitions depend on the current game state, which is analogous to our stage game $G^t$. However, in their setup, automaton state transitions happen after the game state in period $t$ is realized and before actions at time $t$ are chosen, whereas for our purposes it is more convenient to have state transitions between periods.
• If $\omega \leq \lambda(G)w$: Put $f(w,G,\omega) = \overline{a}(G)$, and
  $$\tau(w,G,\omega,a) = \begin{cases} 
  w^* & \text{if } a_1 = \overline{a}_1(G), \\
  w^* - \frac{1-\delta}{\delta}(\hat{u}(\overline{a}(G)) - u(\overline{a}(G))) & \text{otherwise};
  \end{cases}$$

• If $\omega > \lambda(G)w$: Put $f(w,G,\omega) = \underline{a}(G)$, and
  $$\tau(w,G,\omega,a) = \begin{cases} 
  \frac{1-\delta}{\delta}(\hat{u}(\underline{a}(G)) - u(\underline{a}(G))) & \text{if } a_1 = \underline{a}_1(G), \\
  0 & \text{otherwise}
  \end{cases}.$$

In words, we use public randomization to play $\overline{a}(G)$ with probability $\lambda(G)w \in [0,1]$ and play $\underline{a}(G)$ with complementary probability, and then transition to a new state depending on which of the two action profiles was to be played and on whether player 1 deviated. We need to check that all the possible values specified for $\tau$ are indeed valid states (i.e. they lie in the interval $[0,w^*]$); this follows from the fact that $\overline{a}(G)$ and $\underline{a}(G)$ are in $A^*(G,w^*)$.

Starting in any state $w \in [0,w^*]$ and proceeding according to the automaton defines a strategy profile. Denote this strategy profile by $s[w]$. The following is a key step in our analysis.

**Proposition 3.7.** Pick any $w \in [0,w^*]$, and let $E$ be any environment. Then:

1. For each $w$, $U(s[w]|E) = \overline{U}(E) + w$.

2. If the short-run players are following $s[w]$, then at any history $h^t$, player 1 is indifferent between following $s[w]$ and playing the myopically optimal (one-shot) deviation.

3. Strategy profile $s[w]$ is an XPE.

**Proof.** 1: Suppose $G = G^0$ is the first stage game encountered in $E$. By directly considering the possible cases depending on public randomization, and splitting each case into the initial stage and the continuation payoff, we have

\begin{equation}
U(s[w]|E) = \lambda(G)w \times ((1 - \delta)u(\overline{a}(G)) + \delta U(s[w^*]|E^-)) + (1 - \lambda(G)w) \times \left((1 - \delta)u(\underline{a}(G)) + \delta U\left(s\left[\frac{1-\delta}{\delta}(\hat{u}(\underline{a}(G)) - u(\underline{a}(G)))\right]|E^-\right)\right).
\end{equation}

In contrast, write $\tilde{U}(w|E) = \overline{U}(E) + w$. We will show that $\tilde{U}$ satisfies the same
recurrence:

\[ \tilde{U}(w|E) = \lambda(G)w \times \left( (1 - \delta)u(\pi(G)) + \delta\tilde{U}(w^*|E^-) \right) + \\
(1 - \lambda(G)w) \times \left( (1 - \delta)u(a(G)) + \delta\tilde{U} \left( \frac{1 - \delta}{\delta} (\hat{u}(a(G)) - u(a(G))) \right| E^- \right) \]

To see this, expand both the \(\tilde{U}\) terms on the right-hand side of (3.3) and obtain (after slightly simplifying the second line)

\[ \lambda(G)w \times \left( (1 - \delta)u(\pi(G)) + \delta\bar{U}(E^-) + \delta w^* \right) + \\
(1 - \lambda(G)w) \times \left( (1 - \delta)\hat{u}(a(G)) + \delta\bar{U}(E^-) \right). \]

Now by combining the terms with the \(\lambda(G)w\) coefficient, this rearranges to

\[ ((1 - \delta)\hat{u}(a(G)) + \delta\bar{U}(E^-)) + \lambda(G)w \times ((1 - \delta)(u(\pi(G)) - \hat{u}(a(G))) + \delta w^*). \]

But the first parenthesized term is simply \(\bar{U}(E)\) from the definition, and the second term is \(\lambda(G)w/\lambda(G) = w\), so the whole expression reduces to \(U(E) + w = \bar{U}(w|E)\) as claimed.

Now a standard contraction argument shows that the solution to the recurrence is unique: Write \(\Delta(w|E) = U(s[w]|E) - \tilde{U}(w|E)\). Subtracting (3.3) from (3.2) gives

\[ \Delta(w|E) = \lambda(G)w \times \delta \Delta(w^*|E^-) + (1 - \lambda(G)w) \times \delta \Delta \left( \frac{1 - \delta}{\delta} (\hat{u}(a(G)) - u(a(G))) \right| E^- \right). \]

Put \(C = \sup_{w,E} |\Delta(w|E)|\), and note the supremum is finite since both \(U\) and \(\tilde{U}\) are bounded. Using \(C\) to bound each of the \(\Delta(\cdot \cdot \cdot)\) terms in the previous equation gives

\[ |\Delta(w|E)| \leq \lambda(G)w \times \delta C + (1 - \lambda(G)w) \times \delta C = \delta C. \]

Thus, for all \(w\) and \(E\), we have \(|\Delta(w|E)| \leq \delta C\). In other words, \(C \leq \delta C\), which forces \(C = 0\). Therefore, \(U(s[w]|E) = \bar{U}(w|E)\) for all \(w\) and \(E\), which completes the proof of part 1.

2: It suffices to prove the statement at period-0 histories. So suppose the date-0 history is \(h^0 = (G^0, \omega^0)\). Assume that the automaton specifies an action profile \(a^0\) for which 1’s action is not already a myopic best reply (otherwise there is nothing to prove). There are two cases:

- If \(\omega^0 \leq \lambda(G^0)w\), then the action profile to be played is \(\pi(G^0)\). If player 1 conforms,
the state next period is \( w^* \), so the continuation payoff will be \( U(E^-) + w^* \) by part 1, and therefore the total payoff is

\[
(1 - \delta)u(\pi(G^0)) + \delta (U(E^-) + w^*).
\]

If player 1 deviates (optimally) then the stage payoff is \( \hat{u}(\pi(G^0)) \) and the state next period is \( w^* - \frac{1 - \delta}{\delta} (\hat{u}(\pi(G^0)) - u(\pi(G^0))) \), so by a similar calculation, the total payoff is

\[
(1 - \delta)\hat{u}(\pi(G^0)) + \delta \left( \frac{U(E^-) + w^* - \frac{1 - \delta}{\delta} (\hat{u}(\pi(G^0)) - u(\pi(G^0)))}{\delta} \right) = (1 - \delta)u(\pi(G^0)) + \delta (U(E^-) + w^*).
\]

- If \( \omega^0 > \lambda(G^0)w \), then the action profile to be played is \( a(G^0) \). Similar calculations show that the total payoff if 1 conforms is

\[
(1 - \delta)u(a(G^0)) + \delta \left( \frac{U(E^-) + 1 - \delta}{\delta} (\hat{u}(\pi(G^0)) - u(\pi(G^0))) \right) = (1 - \delta)\hat{u}(a(G^0)) + \delta U(E^-)
\]

and if player 1 deviates is

\[
(1 - \delta)\hat{u}(a(G^0)) + \delta U(E^-).
\]

So in each case, the payoffs from conforming and deviating are equal.

3: We have just shown that at every history, player 1 is indifferent to the myopically optimal one-shot deviation. Playing a non-optimal deviation cannot do better, since it leads to the same next-period state (and so the same continuation payoff) as the optimal deviation while giving a lower stage payoff. (Note that if the action profile \( a \) specified is such that 1’s action is already a best reply, then \( \hat{u}(a) = u(a) \), so by inspection of the formulas, the next-period state after a deviation is the same as after conforming, and the same argument applies.) So, player 1 cannot benefit from a one-shot deviation of any sort, and 1’s incentive constraint is satisfied.

The other players’ incentives are also satisfied, since whenever a stage game \( G \) is to be played, the automaton specifies an action profile in \( A^*(G, w^*) \subseteq A^*(G) \). So we have an XPE. □
With the result, we are now justified in thinking of \( w^* \) as the largest sustainable reward-punishment gap (as mentioned in Section 3.1), since we do indeed have two XPE’s—namely, \( s[w^*] \) and \( s[0] \)—whose payoffs differ by \( w^* \) in every environment.

## 4 Main results

With this machinery in hand, we are ready to take up what is arguably the main question of interest in repeated games: what outcomes might arise in equilibrium?

### 4.1 Defining outcomes

A first question is how outcomes should be defined, in this setting without a prior over environments. One option is to take the perspective of an observer watching the game unfold over time. From this perspective, there is a true (but initially unknown) environment, and the strategy profile determines what action profile is played in each period. Thus, we define an observable outcome to be a sequence of the form \( z = (G^0, a^0, G^1, a^1, \ldots) \in (G \times A)^\infty \), specifying a stage game and action profile at each date, such that \( a^t \in A(G^t) \) for each \( t \).

An alternative perspective is to view an outcome as a full description of the actions that may be played “on-path,” or, to take a more decision-theoretic viewpoint, an act defined on possible realizations of the environment. Accordingly, define a full outcome to be a function \( z : \cup_{t=0}^\infty G^{t+1} \rightarrow A \), specifying an action profile \( z(G^0, \ldots, G^t) \in A(G^t) \) for each possible initial sequence of stage games. An observable outcome \( z' = (G^0, a^0, G^1, a^1, \ldots) \) belongs to the full outcome \( z \) if \( z(G^0, \ldots, G^t) = a^t \) for each \( t \).

We will mostly focus on observable outcomes for the sake of expository simplicity. Section 4.5 will state the corresponding results for full outcomes.

Let us say that a strategy profile \( s \) supports the observable outcome given by \( z = (G^0, a^0, G^1, a^1, \ldots) \) if, for all realizations of \( \omega^0, \omega^1, \ldots \), we have

\[
s(G^0, \omega^0, a^0; G^1, \omega^1, a^1; \ldots, G^t, \omega^t) = a^t \quad \text{for all } t.
\]

Thus, in the environment specified by \( z \), the actions specified by \( z \) are played. Note that this definition requires that players do not condition on the public randomization along the specified path, but they may do so elsewhere—either after a deviation, or on-path in other environments.

This notion is admittedly not fully satisfactory: why should we assume that the true
environment happens to be one in which randomization is not used? A better perspective
is that randomization should be allowed, i.e. a more complete definition of an observ-
able outcome would specify an environment and a stochastic process over action profiles,
but that when we study below the question of which such outcomes can arise in equilib-
rium, we confine ourselves for simplicity to studying the special case where the process is
deterministic. Indeed, asking the same question for outcomes that do condition on pub-
lic randomization does not seem conceptually more complicated but would require more
notation.

Say that \( s \) supports a full outcome \( z \) if it supports every observable outcome that
belongs to \( z \). (In this case, strategies condition on public randomization only after a
deviation has occurred.)

### 4.2 Supportable outcomes

It is not hard to see that the following are necessary conditions for an observable outcome
\( z = (G^0, a^0, \ldots) \) to be supported by an XPE \( s \):

\[
a^t \in A^*(G^t, w^*) \quad \text{for all } t; \tag{4.1}
\]

\[
(\hat{u}(a^t) - u(a^t)) + \sum_{t=\bar{t}+1}^{\bar{t}} \delta^{\bar{t}-t} (\hat{u}(a(G^t)) - u(a^t)) \leq \frac{\delta^{\bar{t}+1-\bar{t}}}{1-\delta} w^* \quad \text{for all } \bar{t} < \bar{t}. \tag{4.2}
\]

(Note that an equivalent formulation is simply that \( a^t \in A^*(G^t) \) for all \( t \) and (4.2)
holds for all \( \bar{t} \leq \bar{t} \), where the sum is empty if \( \bar{t} = 0 \).)

Indeed, we have already seen that (4.1) is necessary. For (4.2), consider any \( \epsilon > 0 \),
and consider the environment \( E' \) that consists of \( (G^0, \ldots, G^{\bar{t}}) \), followed by the sequence
of stage games identified in Lemma 3.4 for this \( \epsilon \) (and any other stage games thereafter).
Consider player 1’s decision at time \( \underline{t} \), with history \( h^{\underline{t}} \). Conforming to \( s \) gives a payoff

\[
U(s|E', h^{\underline{t}}) = (1 - \delta) \left( \sum_{t=\underline{t}}^{\bar{t}} \delta^{\bar{t}-t} u(a^t) \right) + \delta^{\bar{t}+1-\bar{t}} \mathbb{E}[U(s|E', (h^{\bar{t}}, a^{\bar{t}}))]. \tag{4.3}
\]

(Here, \( h^{\bar{t}} \) represents the history arising at period \( \bar{t} \), which is random due to the uncertainty
about the intervening public random signals. In fact, the \( \mathbb{E} \) operator is unnecessary
because we know that future play does not condition on the random signals, but it is
useful to be explicit.)
An alternative strategy $s'_1$ would play a myopic best reply to $s_{-1}(h^t)$ at each period $t = \bar{t}, \ldots, T$, and then follow $s_1$ from date $\bar{t} + 1$ onward. This would give a stage payoff of $\hat{u}(a^t)$ in period $t$, and would guarantee at least $\hat{u}(a(G^t))$ in each period $t = \bar{t} + 1, \ldots, T$. So player 1’s deviation payoff satisfies

$$U(s'_1, s_{-1}|E', h^\bar{t}) \geq (1 - \delta) \left( \hat{u}(a^\bar{t}) + \sum_{t=\bar{t}+1}^T \delta^{t-\bar{t}} \hat{u}(a(G^t)) \right) + \delta^{T+1-\bar{t}} \mathbb{E}[U(s|E', (h^\bar{t}, \tilde{a}^\bar{t}))] \quad (4.4)$$

(where $\tilde{h}^\bar{t}$ and $\tilde{a}^\bar{t}$ denote the history and period-$\bar{t}$ actions produced by 1’s deviations; note that the randomness now may affect future play since we are now off-path). Since the deviation should not be profitable, subtracting (4.3) from (4.4) and dividing by $1 - \delta$ gives

$$\left( \hat{u}(a^\bar{t}) - u(a^\bar{t}) \right) + \sum_{t=\bar{t}+1}^T \delta^{t-\bar{t}} \left( \hat{u}(a(G^t)) - u(a^t) \right) + \frac{\delta^{T+1-\bar{t}}}{1 - \delta} \left( \mathbb{E}[U(s|E', (h^\bar{t}, \tilde{a}^\bar{t}))] - \mathbb{E}[U(s|E', (h^\bar{t}, \tilde{a}^\bar{t}))] \right) \leq 0.$$

However, the two $U(s|\cdots)$ terms both represent SPE payoffs in the environment starting at date $\bar{t} + 1$, and so by Lemma 3.4, they differ by less than $w^* + \epsilon$. Applying this bound, rearranging, and taking $\epsilon \to 0$ gives (4.2).

Condition (4.2) essentially says that the payoff gains from repeated myopic deviation across any interval of periods must be bounded by $w^*$ (suitably discounted). Notice that the terms $\hat{u}(a(G^t)) - u(a^t)$ may be positive or negative, so it is unknown a priori for which pairs $(\bar{t}, T)$ the constraint will be tightest.

Our first main result is that conditions (4.1)–(4.2) actually give a complete characterization of the observable outcomes that can be supported in XPE. At first, this may be surprising because (4.2) seems to pertain to incentives for repeated deviation over an interval of periods. An intuition comes from the indifference result of Proposition 3.7: If deviations are optimally punished using the automaton strategies $s[0]$, then the payoff from deviating repeatedly is the same as from a single deviation, so the condition suffices to rule out even one-shot deviations.

**Theorem 4.1.** An observable outcome $z = (G^0, a^0, \ldots)$ is supported by some XPE $s$ if and only if it satisfies the necessary and sufficient conditions (4.1)–(4.2).

**Proof.** Necessity was just argued, so we prove sufficiency. Construct a strategy profile $s$ as follows:

23
• At any history \( h^t \) such that all stage games and action profiles so far have been consistent with \( z \) (i.e. \((G^0, a^0, \ldots, G^t)\) have been observed), play \( a^t \).

• For any history \( h^t \) where the previous stage games and action profiles \((G^0, a^0, \ldots, G^{t-1}, a^{t-1})\) were all as specified by \( z \), but the period-\( t \) stage game is different, play according to \( s[w^*] \) from \( h^t \) onward.

• For any history \( h^t \) where past stage games and all action profiles prior to the previous period, i.e. \((G^0, a^0, \ldots, G^{t-1})\), were as specified by \( z \), but an action profile different from \( a^{t-1} \) was observed, play according to \( s[0] \) from period \( t \) onward.

Notice that at every history, either all stage games and action profiles so far agreed with \( z \), or there was a unique first stage game or action profile that did not agree with \( z \), so this description does specify a well-defined strategy profile.

At any history where any stage game or past action has differed from \( z \), there is no incentive to deviate; this follows because we already know that \( s[w^*] \) and \( s[0] \) are XPE's. Moreover, the incentives of the short-run players are automatically satisfied since an action profile in \( A^*(G) \) is indicated at every history. So we only need to check the incentives of player 1 to deviate at histories \( h^t \) that have so far agreed with \( z \).

Consider such a history \( h^t \), and any environment \( E \) consistent with it. Let \( \bar{t} + 1 \) be the earliest period in which the stage game specified by \( E \) differs from that appearing in \( z \). So \( E \) begins \((G^0, G^1, \ldots, G^\bar{t}, \tilde{G}^{\bar{t}+1}, \ldots)\). Evidently \( \bar{t} \geq t \).

If \( \bar{t} = t \), then by conforming, player 1 achieves a payoff (from the period-\( t \) vantage point) of \((1 - \delta)u(a^t) + \delta(U(E^{-(\bar{t}+1)}) + w^*)\), since play transitions to \( s[w^*] \) next period. By deviating, player 1’s payoff is \((1 - \delta)(\hat{u}(a^t) - u(a^t)) - \delta w^*\), which is \( \leq 0 \) by condition (4.1).

If \( \bar{t} > t \), then by conforming, player 1 achieves a payoff of

\[
(1 - \delta) \left( \sum_{t' = t}^{\bar{t}} \delta^{t'-t}u(a^{t'}) \right) + \delta^{\bar{t}+1-t} \left( U(E^{-(\bar{t}+1)}) + w^* \right).
\]

By deviating, player 1’s payoff is

\[
(1 - \delta)\hat{u}(a^t) + \delta U(E^{-(t+1)}).
\]

By expanding using the definition of \( U \), we get \( U(E^{-(t+1)}) = (1 - \delta) \left( \sum_{t' = t+1}^{\bar{t}} \delta^{t'-(t+1)}u(a^{t'}(G^{t'})) \right) + \)
$$\delta^{\bar{t}-t}U(E^{-\bar{t}+1})$$, and so the deviation payoff is

$$(1 - \delta) \left( \tilde{u}(a_t) + \sum_{t'=t+1}^{\bar{t}} \delta^{t'-t} \tilde{u}(a(G'_{t'})) \right) + \delta^{\bar{t}+1-t}U(E^{-\bar{t}+1}).$$

Now condition (4.2) (with $t$ in place of $\bar{t}$) implies that the deviation is unprofitable.

One loose end remains: what if the environment $E$ never disagrees with $z$? In this case, for each $t' > t$, let $E_{t'}$ be an alternative environment that agrees with $E$ until period $t'$ and disagrees with it starting at $t' + 1$. History $h^t$ is then consistent with $E_{t'}$. Taking limits as $t' \to \infty$, we have $U(s|E_{t'}, h^t) \to U(s|E, h^t)$ and, for any proposed deviating strategy $s_1'$, $U(s_1', s_{-1}|E_{t'}, h^t) \to U(s_1', s_{-1}|E, h^t)$. So the fact that the deviation is not profitable in any $E_{t'}$ (which we have already shown) implies, by taking limits, that it is not profitable in $E$ either.

A few remarks are in order.

First, we can compare the conditions for an XPE outcome against those for an SPE outcome in standard repeated games. By taking the limit as $\bar{t} \to \infty$ in (4.2), we get

$$(\hat{u}(a^\ell) - u(a^\ell)) + \sum_{t=\bar{t}+1}^{\infty} \delta^{t-\bar{t}} (\hat{u}(a(G^t)) - u(a^t)) \leq 0. \tag{4.5}$$

This condition says that the payoff from following the proposed outcome, beginning in period $\bar{t}$, is at least as high as that from a one-period deviation followed by the ensuing punishment. In repeated games, the corresponding condition is sufficient for supportability in SPE (Abreu, 1988, Proposition 4). Here, we need a condition indexed both by $\bar{t}$ and $\bar{t}$ because of the possibility of different stage games arising in future periods. That is, the proposed observable outcome may satisfy (4.5) if there is a large temptation to deviate at period $\bar{t}$ but large rewards promised at some future period. Such an outcome may not be supportable because, when the future comes along, the stage games may be one in which large rewards are impossible, and thus the deviation at $\bar{t}$ cannot be discouraged.

Second, a slight rewriting of the conditions leads to an alternative interpretation. Given the observable outcome $z$, recursively define $r_{-1}(z) = 0$ and

$$r_t(z) = \max \left\{ \frac{1}{\delta} r_{t-1}(z) + (\hat{u}(a(G^t))) - u(a^t) \right\}.$$
for \( t = 0, 1, \ldots \). Then we have (proof in Appendix A):

**Proposition 4.2.** An observable outcome \( z = (G^0, a^0, \ldots) \) is supported by some XPE if and only if it satisfies \( a^t \in A^*(G^t) \) for all \( t \), and \( r_t(z) \leq \frac{\delta}{1 - \delta} w^* \) for all \( t \).

We can think of \( r_t \) as the “reward” owed to player 1 after period \( t \) for refraining from deviation in the past. The proposition then says that an observable outcome can be supported in XPE just so long as the reward owed never exceeds the amount that can be promised. In each period \( t \), the reward promised in the future needs to be large enough to cover the previously promised reward, with “interest,” adjusted by whatever portion is being delivered in the present period (this is the first term of the max); it also needs to be large enough to outweigh the gains from a one-time deviation at \( t \).

Third, we have so far been viewing the set of possible stage games as fixed and asking what outcomes are supportable. But we could equally well flip things around and ask: given a proposed outcome, what sets of stage games allow it to be supported? This question might be of interest, for example, to a long-run player who is confident about the environment and has a desired outcome in mind, but who worries that the short-run players are more uncertain about the environment, and who wants to know what the short-run players need to know in order for them to be assured that the long-run player is willing to follow the plan.

More formally, let \( \mathcal{G} \) be some “universe” of potential stage games, and \( u : \mathcal{A} \to \mathbb{R} \) the corresponding payoff function, satisfying the assumptions of Section 2 (where \( \mathcal{A} \) is the disjoint union of the sets \( A(G) \) for \( G \in \mathcal{G} \)). Let \( z = (G^0, a^0, \ldots) \) be an observable outcome and \( \mathcal{G} \subseteq \mathcal{G} \) with each \( G^t \in \mathcal{G} \). We consider possible sets of stage games \( \mathcal{G} \), with \( \mathcal{G} \subseteq \mathcal{G} \subseteq \mathcal{G} \). Under what conditions on \( \mathcal{G} \) will it be the case that \( z \) is supportable in XPE over \( \mathcal{G} \)? Proposition 4.2 gives an answer: this happens if and only if the value of \( w^* \) for \( \mathcal{G} \) is greater than or equal to \( \frac{1 - \delta}{\delta} \sup_t r_t(z) \); equivalently, if and only if there exists some \( w \geq \frac{1 - \delta}{\delta} \sup_t r_t(z) \) such that \( B(w; G) \geq w \) for each \( G \in \mathcal{G} \). As a side note, observe that this condition is not closed under taking unions: that is, it may be that a set of stage games \( \mathcal{G} \) supports \( z \) as an XPE outcome, and another set \( \mathcal{G}' \) does also, but their union \( \mathcal{G} \cup \mathcal{G}' \) does not, because the value of \( w \) that works for \( \mathcal{G} \) is different than the one that works for \( \mathcal{G}' \). (One can generate examples along the lines of Example 3.1.)

### 4.3 Universal penal codes

Another central result from standard repeated games that does carry over to our setting is the existence of “worst punishments” that can be used to support any equilibrium
outcome path. Explicitly, let us say that a strategy profile $s$ is a *universal penal code* if it has the following property: For every observable outcome $z = (G^0, a^0, \ldots)$ that is supportable in XPE, there is in particular an XPE supporting $z$ where, following any initial deviation by player 1 (i.e., a history of the form $(G^0, \omega^0, a^0; \ldots; G^t, \omega^t)$ and then some action profile $(a'_1, a^t_{-1})$ with $a'_1 \neq a^t_1$), continuation play is given by $s$. We then have the following result:

**Theorem 4.3.** There exists a universal penal code.

*Proof.* It follows from the proof of Theorem 4.1 that $s[0]$ is a universal penal code, since, for any observable outcome meeting conditions (4.1)–(4.2), that proof constructs an XPE supporting it with player 1’s deviations punished by $s[0]$.

(However, unlike the repeated-game setting, here the statement that $z$ should be played with deviations punished by $s$ does not give a full description of the strategy profile, since it does not specify what happens if the history departs from $z$ due to a different stage game being realized.)

As a brief note on literature, Abreu (1988) is usually credited for the notion of penal codes. The relevant definition there is that of an *optimal penal code*, which is a specification of an SPE for each player that delivers to that player the lowest payoff among all SPE’s. Although it is also true here that our $s[0]$ is an optimal penal code, in the strong sense of delivering the lowest XPE payoff in *any* environment, this definition does not explicitly relate to its use as a punishment, which is why we have instead emphasized the definition of universal penal codes here. In more general settings (such as those examined in Section 5), the notions of an optimal penal code and a universal penal code need not coincide.

### 4.4 Foundation for XPE outcomes

As mentioned in the introduction, a difficulty with giving a positive interpretation to XPE is that it is not rooted in individual maximization. One might instead argue that agents should play an SPE of the dynamic game induced by whatever process (perhaps random) they believe governs the stage games. (And even if the agents are unsure about this process, and there is asymmetric information about it, they should presumably play an equilibrium of the the resulting incomplete information game, in which they condition their strategies on their private information.) Of course, a strength of XPE is that any
XPE is automatically an equilibrium of any such fully-specified game as well, and so an XPE-supported outcome is one that the analyst can confidently describe as being attainable in whatever world the players actually live in. But perhaps there are other such outcomes as well: that is, some outcome that is supportable in equilibrium for whatever process might govern the stage games, but requires different punishments for different processes, and so is not supportable in XPE.\(^3\)

Fortunately, this turns out not to be the case: for any outcome \(z = (G^0, a^0, \ldots)\) that is not supportable in XPE, there is some stochastic process over stage games such that, if the players believe this to be the true process and behave as expected-utility maximizers, then the outcome \(z\) is not supportable even in SPE. Thus, the set of XPE-supportable outcomes has a distinguished role to play, even for an analyst who is interested in positive predictions and believes that SPE is the relevant solution concept.

To state the result, we should be a bit more careful: If the players believe that some initial sequence of stage games \((G^0, \ldots, G^t)\) has probability zero (for example, if they assign probability 1 to some environment that does not start with this sequence), then this history does not correspond to any node of the game tree, so it is not clear what it would mean for \(z\) to be supportable in SPE. To steer clear of such interpretive difficulties, we will assume for the rest of this subsection that the set of stage games \(G\) is finite, and we will focus on “full-support” processes over stage games.

Here are formal definitions. A stage game process is a stochastic process \(G\) (defined on some probability space), parameterized by \(t = 0, 1, \ldots\), and taking values in \(G\). Evidently, we can equivalently view a stage game process as an \(E\)-valued random variable, where \(E\) is the set of all environments. We denote a realization of such a process by \(E = (G^0, G^1, \ldots)\) as usual. The process has full support if every joint realization \((G^0, \ldots, G^t)\), for each \(t\), has positive probability. Given any full-support stage game process, a dynamic game can be defined in the natural way (with the understanding that the public random signals \((\omega^0, \omega^1, \ldots)\) are independent of the stage game process). Histories and strategies are defined exactly as in the main model. We can now define expected payoffs at each history.

\(^3\)A parallel is the question of foundations for dominant-strategy implementation in mechanism design Bergemann and Morris (2005); Chung and Ely (2007). Dominant-strategy mechanisms, when they exist, allow for a desired outcome to be achieved regardless of agents’ beliefs or higher-order beliefs about each other. This robustness has led to a large literature focusing on such mechanisms. But even when they do not exist, it may be still be possible to implement the desired outcome with a mechanism where agents’ strategies depend on their beliefs.
\( h^t = (G^0, \omega^0, a^0; \ldots; G^t, \omega^t) \): we have

\[
U_i(s|G, h^t) = \mathbb{E} \left[ U_i(s|E, h^t) \middle| G^0, \ldots, G^t \right],
\]

where the expectation is over environments \( E \), conditional on the observed \( G^0, \ldots, G^t \).

Strategy profile \( s \) is an SPE for \( G \) if, for each player \( i \), each history \( h^t \), and each deviation \( s_i' \), we have

\[
U_i(s|G, h^t) \geq U_i(s_i', s_{-i}|G, h^t).
\]

The definitions of observable outcomes, and strategies supporting such outcomes, are unchanged.

Evidently, any XPE is indeed an SPE for each stage game process: since deviating can never increase the payoff in any environment, it cannot increase the payoff in expectation over environments.

Here, then, is the formal result on “positive foundations” for XPE outcomes:

**Theorem 4.4.** If an observable outcome \( z = (G^0, a^0, \ldots) \) is not supported by any XPE, then there exists a full-support stage game process \( G \) such that \( z \) is not supported by any SPE for \( G \).

**Proof.** By Theorem 4.1, \( z \) must violate either (4.1) or (4.2). If \( a^t \notin A^*(G^t, w^*) \) for some \( t \), then the proof of Lemma 3.5 shows that there exists some finite sequence of stage games such that, in any environment beginning with \( G^t \) followed by this finite sequence, \( a^t \) cannot be played in the initial period of any SPE. Otherwise, if (4.2) is violated for some \( t \), the argument given in the text shows that there is a finite sequence of stage games such that, in any environment consisting of \( (G^0, \ldots, G^T) \) followed by this sequence, there is no SPE in which \( a^t \) is played at each date \( t = t, \ldots, T \). Hence, in either case, we have some finite \( \tilde{T} < T \) and initial sequence of stage games \( (G^0, \ldots, G^\tilde{T}, \tilde{G}^{T+1}, \ldots, \tilde{G}^T) \), with the following property: for any environment beginning with this sequence, there is no SPE that begins by playing action profiles \( (a^0, \ldots, a^\tilde{T}) \). Write \( \tilde{G}^t = G^t \) for \( t \leq \tilde{T} \), and extend this sequence to a full environment \( \tilde{E} \) by choosing \( \tilde{G}^{T+1}, \tilde{G}^{T+2}, \ldots \) arbitrarily.

It remains to perturb \( \tilde{E} \) to a full-support stage game process. For small \( \epsilon > 0 \), define the stage game process \( G_\epsilon \) as follows: for each period \( t \) independently, at date \( t \), the stage game \( \tilde{G}^t \) is picked with probability \( 1 - \epsilon \), and otherwise a stage game from \( G \) is chosen uniformly at random. We argue that for small enough \( \epsilon \), \( G_\epsilon \) does not have any SPE that supports the outcome \( z \).

Notice that for each period \( t \), any two subgames that begin at the start of period \( t \) (before \( G^t \) is realized) are isomorphic, since stage games are chosen independently across periods. Consequently, it makes sense to talk about SPE of “the” subgame beginning in
period $t$. We claim that when $\epsilon$ is small enough, for any two SPE’s of the subgame beginning in period $t+1$, their payoffs differ by less than $w_{T-t}$, where $w_{T-t}$ is the appropriate term of the sequence in Lemma 3.3. Indeed, we just need $\epsilon$ small enough so that the inequality $w_k > B(w_{k-1})$ given by Lemma 3.3 is strengthened to $w_k > (1-\epsilon)B(w_{k-1})+\epsilon \cdot 2M$.

A straightforward adaptation of the inductive argument for Lemma 3.4 then shows that, for each $t = T, T-1, \ldots, t+1$ in succession, any two SPE’s of the subgame starting at $t$ have payoffs differing by less than $w_{T-t}$.

Then, assuming that $\epsilon$ is also chosen small enough such that (if $z$ violates (4.1)) we have
\[ \tilde{u}(a^T) - u(a^t) > \frac{\delta}{1-\delta} w_{T-t}, \]
or (if $z$ violates (4.2)) we have
\[ (1-\tilde{t} \epsilon) \left( \left( \tilde{u}(a^{\bar{t}}) - u(a^t) \right) + \sum_{t'=t+1}^{\bar{t}} \delta^{t'-t} \left( \tilde{u}(G_t'(t')) - u(a'_t) \right) \right) - \tilde{t} \epsilon \cdot 2M > \frac{\delta^{\bar{t}+1-\bar{t}}}{1-\delta} w_{T-t}, \]
then the same deviation that makes $z$ not supportable in SPE under $\tilde{E}$ also makes it not supportable under $G_\epsilon$. (More explicitly: if the former condition holds, then player 1 benefits from deviating in stage $\bar{t}$ if $G_{\bar{t}}^T$ is realized; if the latter, then player 1 benefits from myopically deviating for all of periods $t, \ldots, \bar{t}$ if $G_{\bar{t}}^T$ is realized, regardless of what stage games arise in the later periods. Of course, if (4.1) is violated due to one of the short-run players’ incentives instead, then it is immediate that $z$ cannot be supported.)

\[ \square \]

### 4.5 Full outcomes

All of the preceding results of this section have analogues for full outcomes rather than observable outcomes. We briefly develop the statements, leaving proofs to Appendix A.

Evidently, a full outcome $z$ can only be supported in XPE if each of the observable outcomes belonging to it can, or equivalently, if each such observable outcome satisfies (4.1)–(4.2) (or, again equivalently, the sufficient conditions in Proposition 4.2). The converse is also true, and thus:

**Theorem 4.5.** A full outcome $z$ is supported by some XPE if and only if each of the observable outcomes belonging to it can be supported by some XPE.

Actually, more can be said. Recall that (4.2) implied (4.5), by taking the limit as
It turns out that for full outcomes, we can replace (4.2) with this weaker condition:

**Theorem 4.6.** A full outcome $z$ is supported by some XPE if and only if each of the observable outcomes belonging to it satisfies (4.1) and (4.5).

We no longer need to worry about what happens when the realized environment departs from the specified outcome, because a full outcome by definition considers all possible environments.

Say that a strategy profile $\underline{s}$ is a *universal penal code for full outcomes* if it has the following property: For every full outcome $z$ that is supportable in XPE, there is in particular an XPE supporting $z$ where, following any initial deviation by player 1 (i.e. a period-$t$ history with $a^t = z(G^0, \ldots, G^{t'})$ for each $t' = 0, \ldots, t - 1$, followed by an action profile $(a'_1, a'_{t-1})$ with $a'_1 \neq z_1(G^0, \ldots, G^t)$), continuation play is given by $\underline{s}$. Then:

**Theorem 4.7.** There exists a universal penal code for full outcomes.

(In contrast to observable outcomes, here, specifying a full outcome, together with the punishments after 1’s deviations, fully describes the strategy profile—aside from the technicality that a strategy profile should say what happens after deviations by short-run players, but these can simply be ignored.)

Finally:

**Theorem 4.8.** Assume that $\mathcal{G}$ is finite. If a full outcome $z$ is not supported by any XPE, then there exists a full-support stage game process $G$ such that $z$ is not supported by any SPE for $G$.

## 5 Discussion

In some sense, the analysis so far seems quite straightforward: there is some maximum reward gap $w^*$ that can be credibly promised to the long-run player; a contemplated outcome is supportable just so long as the future reward that is owed to compensate for past foregone temptations never grows larger than $w^*$. However, as discussed in the introduction, our framework involves particular features, in particular the restriction that there is only a single long-run player for whom dynamic incentives need to be provided, and also the availability of public randomization. We will show here that the results change if either of these conditions is removed. In particular, a universal penal code may
no longer exist. (Relatedly, an optimal penal code, i.e. an XPE giving the lowest payoff in any environment, can also fail to exist.) Given the central role of the universal penal code in the analysis, this suggests that the theory of XPE’s in these broader settings would have to look very different.

For standard repeated games, the usual recursive analysis by way of continuation values, and its corresponding role for optimal penal codes, applies equally well with or without public randomization, and with any combination of long-run and short-run players. This contrast suggests that the mapping from the logic of debts and rewards emphasized in our analysis to the usual language of continuation values is not immediate in general.

For ease of exposition, the examples in this section are presented in a slightly different framework than the main model: we assume a nonstationary framework (i.e. for each period $t$, there is a different set of possible stage games $G^t$); we also assume a finite horizon, and no discounting. None of these changes matters conceptually. Indeed, we could have redone all of the earlier analysis in a nonstationary framework, at the expense of more notation: there would be a separate $B^t$ operator for each period $t$; then there would be different values $w^*_{t_1}$ for the maximum reward-punishment gap starting in period $t$; and penal codes would be indexed by time (so that Theorem 4.3 would say that for each $t$, there is a punishment $s^t$ such that any supportable outcome can be supported using this punishment for deviations at $t - 1$). Conversely, Appendix B shows how to build on the examples below to express the same ideas while retaining stationarity.

For brevity, we skip over some of the formalities for these examples.

5.1 No public randomization

We first consider a framework without public randomization available. Our example features three periods, $t = 0, 1, 2$. The best way to give the long-run player a low payoff starting at date 1 depends which stage game will be realized at date 2, because the latter determines how expensive it is to deter stage-1 deviations. This means that there is no XPE that gives the lowest payoff starting from date 1 in every environment. This in turn leads to the lack of a universal penal code.

Example 5.1. Consider the sets of stage games shown in Figure 2. There is one possible stage game in each period $t = 0, 1$, and two possible stage games in period 2. Each $A^*(G)$ consists of the diagonal. As with Figure 1, part (a) presents the stage games in standard matrix form, while part (b) shows the sets $A^*(G)$ in each stage game and the values of $u, \hat{u}$.
Let $s_c$ denote the XPE profile for the subgame beginning in period 1 that plays actions $cc, hh, jj$ along the path of play and, if player 1 deviates in period 1, plays $ff$ or $ii$ in period 2 accordingly. (As usual, deviations by 2 can be ignored.) Player 1’s total payoff across the two stages is 10 or 5, depending whether $G^2$ or $G^{2'}$ is realized.

Let $s_d$ denote the XPE profile beginning in period 1 that plays actions $dd, gg, jj$ on-path, with $ff$ or $ii$ in period 2 if player 1 deviates in period 1. Player 1’s total payoff for the two periods is 7 or 11, respectively.

In the three-period game, the observable outcome with actions $aa, ee, ff$ can be supported in XPE. Namely, specify $aa, ee, ff, jj$ as on-path actions. If player 1 deviates in period 0, then switch to $s_d$ as punishment. (Deviation in period 1 only can be ignored, since it brings no within-period gain.) For both realizations of the period-2 game, the punishment is sufficient to deter the period-0 deviation.

The observable outcome with actions $aa, ee, ii$ can also be supported in XPE: specify $aa, ee, hh, ii$, and deter a period-0 deviation by using $s_c$ as punishment.

However, there is no punishment that can support both of these outcomes at once.
Indeed, to support $aa, ee, ff$, the punishment after a period-0 deviation to $b$ has to deliver total payoff $\leq 7$ across the two remaining periods in environment $(G^0, G^1, G^2)$, which requires beginning with $dd$ (since $ee$ is clearly too generous, and $cc$ would have to be followed by $hh$, otherwise 1 would deviate in period 1, but $hh$ is also too generous). To support $aa, ee, ii$, the punishment after $b$ has to deliver total payoff $\leq 7$ across $G^1$ and $G^{2'}$, but this cannot be done using $dd$ (because $dd$ must be followed by $jj$ to deter a period-1 deviation, but this is again too generous). So, no one punishment can support both outcomes.

This shows that Theorem 4.3 fails without public randomization (and the same is true for Theorem 4.7). Note that it also shows that Theorem 4.5 fails, since the full outcome $aa, ee, ff, ii$ cannot be supported in XPE even though its constituent observable outcomes can. Finally, a minor variant of this example suffices to give a case with a single full outcome that can be supported, but only by using different punishments for different date-1 histories. Namely, create a second stage-0 game $G^{0'}$ that is a copy of $G^0$, and now consider the full outcome that specifies $aa, ee, ff, jj$ if $G^0$ is realized, and $aa, ee, hh, ii$ if $G^{0'}$ is realized. This contrasts with stochastic games, where universal penal codes for full outcomes do exist (Kitti, 2016).

\[\triangle\]

5.2 Multiple long-run players

Let us now restore public randomization, but suppose that there are two long-run players, who both act to maximize the sum of payoffs across periods. We again give a three-period example where there is no universal penal code. Similar to the previous example, the most effective way to give the long-run player a low payoff starting at date 1 depends what stage game will be realized at date 2. Here, the reason is that there is an action profile at date 1 that gives a low payoff to player 1 but also a high temptation to deviate for player 2. It may or may not be possible to reward player 2 in the next period for foregoing this temptation without also giving a high payoff to player 1, depending which period-2 stage game is realized.

Example 5.2. Again, one possible stage game in each period $t = 0, 1$, and two in period 2. The relevant stage games are illustrated in Figure 3.

Let $s_r$ denote the XPE for the subgame starting in period 1 that consists of playing $cr$ followed by $dt$ or $gw$. (Deviation by player 2 in period 1 can be ignored, since it brings no gain.) This delivers to the two players total payoffs of $(1, 1)$ across the two periods,
Figure 3: Example with two long-run players. No universal penal code exists.

Let $s_s$ (with apologies for the notation) be the XPE starting in period 1 that plays $cs$ followed by $fv$ or $hw$ on-path, and that punishes a deviation by player 2 at period 1 by following up with $dt$ or $gw$. This delivers total payoffs across the two periods of $(3, 1)$ if $G^2$ is realized and $(0, 1)$ if $G^{2'}$ is realized.

In the overall game, the observable outcome with actions $aq, cr, eu$ can be supported, with $hx$ being chosen in period 2 if $G^{2'}$ is realized. To see this, we just need to be able to deter deviation to $b$ in period 0 (since the specified play constitutes a stage Nash in each subsequent period), and this can be done using $s_s$ as a punishment. The observable outcome with actions $aq, cr, gw$ can also be supported, with $fv$ being chosen on-path in period 2 if $G^2$ is realized. To do this, again we only need to deter the deviation to $b$, and this can be done by punishing with $s_s$.

However, no punishment can support both of these outcomes in XPE. Such a punishment would need to deliver an (expected) payoff to player 1 of $\leq 1$ across periods 1–2 if $G^2$ is realized and a payoff $\leq 0$ if $G^{2'}$ is realized. The latter, in particular, means that $cs$ must be played with probability 1 in period 1. But then the only way to deter 2 from deviating is to reward her with $fv$ (again with probability 1) if $G^2$ is realized. This means that player 1’s total payoff across the two periods is 3 in this environment, contrary to the requirement. Note also that this argument has accounted for the availability of public randomization.

This shows that the analogue of Theorem 4.3 with two long-run players does not hold. The example also can be modified to show that various other results from the main analysis do not extend, just as was done in Example 5.1.

$\triangle$
5.3 Why the difference?

Why does the recursive approach fail to extend to these settings? It may be helpful to try to imagine a common generalization of the approach taken here and APS. Both seek to characterize the set of appropriately-normalized reward vectors that can be achieved in equilibrium, so as to determine which deviations can be deterred. In APS the normalized reward is just the total payoff, whereas here it is the reward-punishment gap. $B(W)$ is the set of normalized rewards that can be attained, given that the continuation rewards are taken from the set $W$. But what does “attained” mean?

The incentive constraints require inequalities: as long as we can ensure a normalized reward of at least $v$ for each stage game $G$ that might materialize, we can deter a deviation gain of corresponding size. Doing so may involve using different rewards for different realizations of $G$. The promise-keeping constraints, however, require equalities: a lower bound on the continuation reward is not enough, because the total reward must be kept low in order to be usable as a punishment at the next stage of the recursion. Thus, for the recursive technique to apply, the same $B(\cdot)$ operator needs to be able to calculate both

$$\{v \mid \text{can ensure normalized reward } \geq v \text{ for all } G\}$$

and

$$\{v \mid \text{can ensure normalized reward } = v \text{ for all } G\}.$$

In the “one-dimensional” case studied here with public randomization, the two sets are just intervals, and they coincide. In general, this will not be true. (In APS, where there is only one game $G$, the equality set uniquely determines the inequality set, so it suffices to recurse on the equality set only.)

6 Summary

Repeated games are a standard modeling tool for studying dynamic incentives in long-run interactions. The canonical question in such models is: what outcomes can be supported in equilibrium? Typically, some amount of non-myopic behavior can be supported by the threat that deviations will be punished with a reversion to stage Nash behavior in the future. But this does not capture the full range of outcome that can be supported, as usually more can be done with the threat of stronger punishments. However, the punishments need to be credible, and so the question of identifying the exact limits of
what can be supported becomes delicate. The recursive analysis of APS gives an answer to
this question. But the basic version of this analysis is carried out in a stylized framework
in which the same game is played in every period, and the recursion is done in the space of
continuation values. This paper has explored a generalization to a setting where multiple
stage games may arise in each period, and the stage game may vary unpredictably from
one period to the next, to try to understand how much of the standard theory—and
in particular the underlying intuition of incentivizing non-myopic behavior via future
rewards—carries through. In particular, this generalization offers some perspective on
how much of the standard theory relies on being able to refer to numeric continuation
values.

We adopted ex-post perfect equilibrium as a solution concept. Under two significant
restrictions—a single long-run player interacting with a series of short-run players, and
availability of public randomization—the recursive technique adapts to identify the max-
imum gap between future reward and punishment that can credibly be promised to the
long-run player. This leads to a characterization of the outcomes attainable in equilib-
rium, as ones for which the reward owed to compensate past obedience never exceeds
this maximum gap. Any such outcome is supportable by using a single worst punishment
following any deviation. And, using the characterization of supportable outcomes, we are
able to connect the concept of ex-post perfect equilibrium to the more standard one of
subgame-perfect equilibrium, by showing that the ex-post perfect equilibrium outcomes
are exactly the ones that can be supported no matter what Bayesian belief players hold
about the environment.

With multiple long-run players, or without public randomization, the analysis does
not replicate (and, in particular, universal penal codes no longer exist). This raises ques-
tions about what are the right set of concepts for studying future promises and dynamic
incentives in these more general settings, without the rigidity of the standard model of
repeated games. Perhaps some other modeling framework, or some other equilibrium
concept, will lead to new insights.

A Omitted proofs

Proof of Lemma 3.2. Consider any decreasing sequence $w_k \rightarrow w$. Because $B(w; G)$ is in-
creasing, the limit $\lim_{k} B(w_k; G)$ is well-defined (with value $-\infty$ if eventually $B(w_k; G) =
\infty$), and right-continuity will follow if we can show that $B(w; G) \geq \lim B(w_k; G)$.

We can assume that $B(w_k; G) \neq -\infty$ for all $k$, since otherwise monotonicity implies
$B(w; G) = -\infty = \lim B(w_k; G)$. Granted this, take $a_k, a'_k \in A^*(G, w_k)$ attaining the max and min in the definition of $B(w_k; G)$. By compactness, we can pass to a subsequence for which $a_k$ and $a'_k$ converge to limits $a_\infty, a'_\infty$. Continuity of $u$ and $\hat{u}$ then imply that $a_\infty, a'_\infty \in A^*(G; w)$, and

$$B(w; G) \geq (1 - \delta)(u(a_\infty) - u(a'_\infty)) + \delta w$$

$$= \lim_{k} (1 - \delta)(u(a_k) - u(a'_k)) + \delta w_k$$

$$= \lim_{k} B(w_k; G).$$

Thus, $B(w; G)$ is right-continuous. For $B(w)$, again consider a decreasing sequence $w_k \to w$. If $B(w) = -\infty$, then $B(w; G) = -\infty$ for some $G$, hence the previous argument implies $B(w_k) = B(w_k; G) = -\infty$ for nearby $w_k$. Otherwise, if the desired right-continuity fails then there exists $\epsilon > 0$ such that $B(w_k) > B(w) + \epsilon$ for all $k$. Take $G$ such that $B(w; G) < B(w) + \epsilon/2$; then right-continuity of $B(w; G)$ for this specific $G$ implies $B(w_k) \leq B(w_k; G) < B(w) + \epsilon$ for large enough $k$, a contradiction.

Proof of Lemma 3.3. First, note that $w_k > w^*$ by induction: This is clearly true for $w_0$; then $w_1 > B(w_0) > B(w^*) = w^*$ by strict monotonicity of $B$, and for $k \geq 2$ we then have $w_k = (B(w_{k-1}) + B(w_{k-2}))/2 > (B(w^*) + B(w^*)/2 = w^*$ by strict monotonicity and induction hypothesis.

In particular, the terms $w_k$ never fall to $-\infty$. Now we prove the ensuing statements:

1: We have $w_1 < w_0$ from the construction, and then $B(w_1) < B(w_0) < w_1$ by strict monotonicity, from which $w_2 = (B(w_0) + B(w_1))/2 < w_1$. Now proceed by induction: if $k > 2$ and $w_{k-1} < w_{k-2} < w_{k-3}$, then $w_k = (B(w_{k-1}) + B(w_{k-2}))/2 < (B(w_{k-2}) + B(w_{k-3}))/2 = w_{k-1}$ by strict monotonicity.

2: For $k = 1$ this is given; for $k \geq 2$ we have $w_k = (B(w_{k-1}) + B(w_{k-2}))/2 > B(w_{k-1})$ using strict monotonicity of $B$ and the fact that $w_{k-2} > w_{k-1}$.

3: Since the sequence is decreasing and bounded below by $w^*$, it has a limit $w_\infty$. Right-continuity of $B$ implies $w_\infty = \lim_k w_k = \lim_k (B(w_{k-1}) + B(w_{k-2}))/2 = (B(w_\infty) + B(w_\infty))/2 = B(w_\infty)$. But since $w_\infty \geq w^*$, and no value greater than $w^*$ is a fixed point of $B$, we have equality.

Proof of Proposition 4.2. As noted in the text, the conditions (4.1)–(4.2) as stated are equivalent to requiring $a^t \in A^*(G^t)$ for all $t$ and (4.2) for all $t \leq T$. So it suffices to check
that the latter is equivalent to \( r_t \leq \frac{\delta}{1-\delta} w^* \) for all \( t \). Rewrite (4.2) as

\[
\frac{1}{\delta^{t-1}} (\hat{u}(a^*) - u(a^*)) + \sum_{t'=t+1}^{\bar{t}} \delta^{t'-t} (\hat{u}(a(G')) - u(a')) \leq \frac{\delta}{1-\delta} w^*. \tag{A.1}
\]

Writing the left-hand side of (A.1) as \( r_{t,t}(z) \), requiring (4.2) for all \( t, \bar{t} \) is equivalent to \( \max_{0 \leq t \leq \bar{t}} r_{t,t}(z) \leq \frac{\delta}{1-\delta} w^* \) for all \( \bar{t} \). But it is easy to see by induction that \( r_{\bar{t}}(z) = \max_{0 \leq t \leq \bar{t}} r_{t,t}(z) \).

\[ \square \]

**Proof of Theorem 4.5.** Necessity is immediate. For sufficiency, we need to argue that it suffices for each observable outcome to satisfy (4.1)–(4.2). This follows from sufficiency of the weaker conditions in Theorem 4.6, so we defer to that proof.

\[ \square \]

**Proof of Theorem 4.6.** Again, we already have necessity, so we focus on sufficiency. Adapting the proof of Theorem 4.1, we construct a strategy profile \( s \) as follows: at any history where there has not yet been a deviation from \( z \), play as specified by \( z \) (regardless of the public random signals); when a deviation first occurs at some period \( t - 1 \), play according to \( s[0] \) from period \( t \) onward. Since \( z \) specifies an intended action profile for every possible initial sequence of stage games, this description fully specifies a strategy profile.

As in the earlier proof, we just need to check the incentives of player 1 at any history \( h^t \) where a deviation has not yet occurred. Fix any environment \( E = (G^0, G^1, \ldots) \) such that \( h^t \) is consistent with \( E \). Let \( (G^0, a^0, G^1, a^1, \ldots) \) be the observable outcome specified by \( z \) for this environment. If player 1 conforms to \( s \), then the payoff starting at \( h^t \) from conforming is

\[
(1 - \delta) \left( \sum_{t'=t}^{\infty} \delta^{t'-t} u(a') \right). \tag{A.2}
\]

If player 1 deviates from \( s \), then subsequent play transitions to \( s[0] \) and so the payoff from \( t + 1 \) onward is given by \( U(E^{-(t+1)}) \) (with the same notation for shifted environments as in the earlier proof). Therefore, the payoff from deviating optimally, as measured from \( h^t \), is

\[
(1 - \delta) \hat{u}(a^t) + \delta U(E^{-(t+1)}) = (1 - \delta) \left( \hat{u}(a^t) + \sum_{t'=t+1}^{\infty} \delta^{t'-t} \hat{u}(a(G')) \right). \tag{A.3}
\]
Rearranging (4.5) tells us exactly that (A.2) is greater than or equal to (A.3). Hence, deviating is never profitable, in any environment.

Proof of Theorem 4.7. It is immediate from the proof of Theorem 4.6 that $s[0]$ is a universal penal code for full outcomes.

Proof of Theorem 4.8. If $z$ is not supportable in XPE, then by Theorem 4.5, one of its constituent observable outcomes is not either. By Theorem 4.4, there is some full-support stage game process for which this observable outcome is not supportable in SPE, and a fortiori the full outcome $z$ is not either.

B Stationary versions of counterexamples

We sketch here constructions analogous to Examples 5.1 and 5.2, but retaining the stationary structure of the original model (including infinite horizon and discounting).

Example B.1. For this example, we assume one long-run player and no public randomization, as in Example 5.1. We assume $G$ consists of five stage games as shown in Figure 4. The discount factor is $\delta = 1/10$. (This makes the numbers simple, but similar examples can be constructed for $\delta$ arbitrarily close to 1.) For brevity, we avoid writing out the games in traditional matrix form, and instead just directly name the action profiles assumed to comprise $A^*(G)$ and list the values of $u$ and $\hat{u}$, as in Figure 2(b).

$$
\begin{array}{ccc}
G_1 : & a & b & v \\
  & u & 0 & 0 & 10000 \\
 \hat{u} & 0 & 4 & 10000 \\
 & w & 0 & 0 & 10000 \\
\end{array}
\begin{array}{ccc}
G_2 : & c & d & e \\
  & u & 0 & 60 & 110 & 10000 \\
 \hat{u} & 50 & 70 & 110 & 10000 \\
 & w & 0 & 0 & 10000 \\
\end{array}
\begin{array}{ccc}
G_3 : & f & g & x \\
  & u & 0 & 100 & 10000 \\
 \hat{u} & 0 & 100 & 10000 \\
 & w & 0 & 0 & 10000 \\
\end{array}
\begin{array}{ccc}
G_4 : & h & i & y \\
  & u & 0 & 500 & 10000 \\
 \hat{u} & 0 & 500 & 10000 \\
 & w & 0 & 0 & 100000 \\
\end{array}
\begin{array}{ccc}
G_5 : & j & z \\
  & u & 0 & 1000000 \\
 \hat{u} & 0 & 1000000 \\
 & w & 0 & 0 & 1000000 \\
\end{array}
$$

Figure 4: Stationary example of no universal penal code without public randomization.

There exists an XPE that supports the observable outcome $(c, i, j, j, j, \ldots)$ (here we suppress the list of stage games involved, for brevity). In particular, specify that if “Nature
deviates” by ever choosing a stage game different from that specified by the outcome, then reward actions \((v, w, x, y, z)\) are played from then onward; if player 1 deviates from \(c\) in the first period, then the worst stage-Nash actions \((a, e, f, h, j)\) are played subsequently. All other deviations can be ignored since there are no short-run gains.

There exists an XPE that supports the observable outcome \((d, g, j, j, j, \ldots)\). If Nature ever deviates, use reward actions as above; if player 1 deviates from \(d\) in the first period, then use the worst stage-Nash actions in all subsequent periods.

These, in turn, can be used to support two different observable outcomes that start with \(b\) being played in \(G_1\) in period 0. First, we can support \((b, e, f, j, j, j, \ldots)\) by specifying that reward actions are to be played if Nature deviates, and a deviation from \(b\) by player 1 is punished as follows: in period 1, if the stage game drawn is \(G_2\), we commence the \((d, g, j, j, j, \ldots)\) equilibrium, and otherwise we simply play worst stage-Nash in every period. It is straightforward to check that this deters the deviation to \(b\) in every possible environment (note that there are several cases to check depending when the environment first differs from the proposed outcome).

Second, we can support \((b, e, h, j, j, j, \ldots)\) by specifying that reward actions are to be played if Nature deviates, and a deviation from \(b\) by player 1 is punished as follows: in period 1, if \(G_2\) is drawn, then we commence the \((c, i, j, j, j, \ldots)\) equilibrium, and otherwise we play worst stage-Nash in every period.

Finally, we claim there is no XPE punishment \(\xi\) that can support both the \((b, e, f, j, j, j, \ldots)\) and \((b, e, h, j, j, j, \ldots)\) outcomes, thus showing nonexistence of an optimal penal code in this environment. Indeed, to be an effective deterrent, \(\xi\) would have to give a total payoff of at most 59 in both environments \((G_2, G_3, G_5, G_5, G_5, \ldots)\) and \((G_2, G_4, G_5, G_5, G_5, \ldots)\). We show that no XPE \(\xi\) can have this property.

Evidently, if \(G_2\) is drawn initially then either \(c\) or \(d\) must be played. Suppose that \(c\) is played. In the continuation environment \((G_3, G_5, G_5, \ldots)\), the total payoff needs to be at least the next three periods. But this in turn means that if the continuation environment turns out to be instead \((G_3, G_5, G_5, G_5, G_3, G_3, \ldots)\), then the total payoff is at most \((1 - \delta)(100 + (\delta^4 + \delta^5 + \cdots) \cdot 10000) = 91\), which is not enough to deter the deviation from \(c\) in the preceding period. Correspondingly, if \(\xi\) begins by playing \(d\) in \(G_2\), then the continuation in environment \((G_4, G_5, G_5, G_5, \ldots)\) needs to have payoff at most 50. It therefore needs to begin with \(h\) followed by at least three copies of \(j\). This means that if the continuation environment is instead \((G_4, G_5, G_5, G_5, G_3, G_3, \ldots)\) then this continuation has payoff no more than 1, which means it cannot deter the deviation from \(d\) in the initial
Example B.2. We now restore public randomization but consider two long-run players. We build on Example 5.2, using the ideas of Example B.1 to extend to a stationary environment. The possible stage games are shown in Figure 5. We again assume the discount factor is $\delta = 1/10$ (for both players).

<table>
<thead>
<tr>
<th>$G_1$</th>
<th>$q$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>0,0</td>
<td>0,0</td>
</tr>
<tr>
<td>$b$</td>
<td>1,0</td>
<td>0,0</td>
</tr>
<tr>
<td>$c$</td>
<td>0,0</td>
<td>10000,10000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$G_2$</th>
<th>$s$</th>
<th>$t$</th>
<th>$u$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d$</td>
<td>10,10</td>
<td>0,0</td>
<td>0,0</td>
</tr>
<tr>
<td>$e$</td>
<td>0,0</td>
<td>0,0</td>
<td>10000,10000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$G_3$</th>
<th>$v$</th>
<th>$w$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>0,0</td>
<td>0,0</td>
<td>0,0</td>
</tr>
<tr>
<td>$g$</td>
<td>0,0</td>
<td>100,0</td>
<td>0,0</td>
</tr>
<tr>
<td>$h$</td>
<td>0,0</td>
<td>0,0</td>
<td>30000,10000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$G_4$</th>
<th>$y$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i$</td>
<td>0,0</td>
<td>10000,0</td>
</tr>
<tr>
<td>$j$</td>
<td>0,10000</td>
<td>10000,10000</td>
</tr>
</tbody>
</table>

Figure 5: Stationary example of no universal penal code with two long-run players.

We first identify as “reward” the high-payoff profile in each stage game, which is always stage Nash $(cr, eu, hx, jz)$, and identify as “punishment” the stage Nash with the lowest payoff for player 1 $(bq, ds, fv, iy)$.

Let $s_s$ be the XPE that always plays the punishment action profile. Deviations are simply ignored. This is an XPE since it plays a stage Nash in every period and deviations do not affect future play.

Let $s_t$ be the XPE that does the following: If $G_2$ is drawn in the initial period, then $dt$ is to be played. If player 2 does not deviate from $t$, then in the next period, $cr, eu, hx$, or $jy$ is to be played depending on the stage game (i.e. the reward profile, except that we play $jy$ instead of $jz$ in $G_4$); and after that, the punishment profile is played in all subsequent periods. If 2 does deviate in the initial period, then the punishment profile is played in all subsequent periods. If the initial stage game is not $G_2$, then we simply play the punishment profile in every period. All deviations are ignored except deviation by 2 in the initial period as described above. Note that this is an XPE: it specifies stage Nash in every period, except in the initial period if $G_2$ is drawn, but the punishment the next period is sufficient to deter 2 from deviating to $s$.

Now, we can use these to support two different observable outcomes that begin with $aq$.
being played in $G_1$ in period 0. First, the observable outcome $(aq, ds, gw, fv, fv, fv, \ldots)$ can be supported as follows. If Nature deviates, play reward profiles forever. Deviations by players are ignored unless they bring a short-run gain, as usual. So we need only worry about deviation by player 1 to $b$ in period 0, and we specify that this deviation is punished by switching to $s_s$. We can check that this punishment deters the deviation in every environment (as in Example B.1, there are cases to check depending when the stage games first diverge from those in the target outcome).

Second, the observable outcome $(aq, ds, iy, iy, iy, \ldots)$ can be supported by specifying that a deviation by Nature is followed with reward profiles, while a deviation by played 1 in period 0 is punished by following with $s_t$. Again, this punishment deters the deviation in all environments (with several cases to check).

Finally, we cannot support both $(aq, ds, gw, fv, fv, fv, \ldots)$ and $(aq, ds, iy, iy, iy, iy, \ldots)$ using the same XPE $s$ to punish player 1 for a period-0 deviation in both cases. Such an XPE would have to give a payoff to player 1 of at most 10 in the environment $(G_2, G_3, G_3, G_3, \ldots)$ and at most 0 in the environment $(G_2, G_4, G_4, G_4, \ldots)$. The latter implies that in the initial period, in $G_2$, only action profiles with payoffs $(0, 0)$ can be played with positive probability (accounting for the ability to use public randomization). However, player 2 needs to be guaranteed a total payoff at least 9 in environment $(G_2, G_3, G_3, G_3, \ldots)$, since she can get this much by myopically deviating in the initial period. This means that in this environment, $s$ has to give player 1 an expected payoff of at least 27, because 1’s payoff is always at least three times 2’s payoff (in the initial period this holds because both are getting payoff 0, as argued above, and in subsequent periods it holds because every action profile in $G_3$ satisfies this relation). This contradicts the requirement that 1’s payoff from $s$ in this environment should be at most 10.

\[\triangle\]

**References**


