Dynamic Incentives in Incompletely Specified Environments

Gabriel Carroll, University of Toronto
gabriel.carroll@utoronto.ca

August 11, 2024

Abstract

Consider a repeated interaction where it is unknown which of various stage games will be played each period. This framework separates the basic logic of intertemporal incentives from the requirement that any given strategy profile yields a well-defined payoff vector. A natural solution concept is ex post perfect equilibrium: strategies must form a subgame-perfect equilibrium for any realization of the sequence of stage games. When there is one long-run player and others are short-run, and public randomization is available, we can adapt the standard recursive approach to determine the maximum sustainable gap between reward and punishment for the long-run player, and use this to fully characterize what outcomes are supportable in equilibrium. With multiple long-run players or no public randomization, the approach fails, because optimal penal codes may not exist.

Thanks to (in random order) Drew Fudenberg, Daniel Garrett, Alex Wolitzky, Andrzej Skrzypacz, Wiola Dziuda, Piotr Dworczak, George Mailath, Paul Milgrom, Kristof Madarasz, Bruno Strulovici, Johannes Hörner, Rohit Lamba, Takuro Yamashita, and Takuo Sugaya for discussions and comments, as well as audiences at many seminars and conferences. Ellen Muir provided valuable research assistance. This work was partially supported by an NSF CAREER award. The author also thanks the economics department at the University of Oxford for their hospitality during a sabbatical.

1 Introduction

Repeated games provide a canonical formal framework for studying the power, and the limits, of self-enforcing dynamic incentives. An analyst interested in some situation where

such incentives matter can model the situation as a repeated game, and then can draw on a well-developed set of mathematical tools to describe what outcomes can arise in equilibria of the game. Moreover, these tools can tell not only *what* but also *why*: that is, for the outcomes that can be supported in equilibrium, how to do so (i.e. what off-path play deters deviations); and for outcomes that cannot be supported, a explanation of why not.

For a familiar classroom example, suppose the analyst expects two players, who share a common discount factor δ, to play a repeated version of the prisoner's dilemma depicted below. Can they sustain the cooperative outcome of (c, c) in every period?

	c	d
c	$2, 2$	$0, 3$
d	$3, 0$	$1, 1$

The answer is that they can if $\delta \geq 1/2$, and not otherwise. A routine explanation is as follows: if $\delta \geq 1/2$, cooperation can be sustained by threatening deviators with (d, d) forever (and one can verify directly that this punishment is sufficient); if $\delta < 1/2$ then no punishment can deter deviation, since the worst punishment we could hope to inflict is for the opponent to play d forever, penalizing the deviator by a loss of 1 unit in each future period, but this does not outweigh the one-period gain of 1 from deviation.

Much more generally, for *any* specification of the stage game and any target outcome the analyst might be interested in, the standard toolkit provides, at least in principle, not only a recipe for checking whether the target outcome is supportable in equilibrium, but also an explanation for why or why not: the recursive analysis of Abreu, Pearce and Stacchetti (1990) (henceforth APS) can be used to find the worst punishment that can be inflicted on any player in equilibrium (the "optimal penal code"), and then we know from Abreu (1988) that the possible equilibrium outcome paths are exactly those for which this optimal penal code is sufficient to deter deviations. Then, for any outcome that can be supported, we know how to support it; and for any that cannot, we can unwind the steps of the APS recursion to describe why no stronger punishment is available.

However, the APS machinery takes as input a complete description of the game matrix. On the other hand, the verbal explanation in the example above does not use all the information in the matrix; nor does it even require that the matrix should be the same in every period. For example, the payoff from (d, d) could be replaced by some other (possibly time-varying) number instead of 1; as long this number is positive but low enough—specifically, always below $3 - 1/\delta$—the threat of reversion to (d, d) still supports

cooperation; and as long as it is above $3 - 1/\delta$, the argument for cooperation being unsupportable goes through. This simple reasoning seems to accord with everyday intuition about dynamic incentives. And yet the APS recursive machinery does not readily express it. Indeed, that machinery operates on vectors of payoffs from the whole infinite-horizon game, and it is not clear how to apply it when the payoffs from a given strategy profile are indeterminate. Given that analysts (and indeed players themselves) are unlikely ever to actually know the exact parameters of the game being played, this suggests that there is something basic about the logic of dynamic incentives that the standard tools are not capturing.

This paper aims to get at this basic logic by explicitly stripping away the assumption of a completely specified environment. We take up a framework where, at each moment in time, the nature of rewards and punishments available in the future is not fully specified, and yet it is specified in enough detail to enable some reasoning about the possibilities and limits of dynamic incentives. We then ask whether a version of the standard tools for repeated games can be carried over to analyze such a setting. Ideally, we would like an analysis that is as comprehensive as in the standard framework. Specifically, we seek an exact answer to the question: which outcomes can be incentivized in a way that doesn't require the analyst's missing information to be filled in?

There may be many ways to formulate such an incompletely specified model. We adopt here a minimal departure from the standard framework—one that is also adopted in a recent paper by Krasikov and Lamba (2023) (discussed further below) and is there termed "uncertain repeated games." Instead of a single stage game repeated each period, there is a *set* of stage games that are considered possible. Each period, a stage game is drawn from the set and observed by the players, who then take actions. But future stage games need not be the same as the present one, and no probabilistic belief is given as to the process governing their evolution. It is, however, assumed that the future stage games do not depend on players' past actions. This modeling framework is thus rich enough to express the concept of dynamic incentives, while the nonprobabilistic aspect makes it impossible to define expected continuation values and thus precludes a totally mechanical application of the usual APS recursion.

What is an appropriate equilibrium concept for this prior-free environment? We would like to express the following basic intuition: a player should be willing to forgo a present gain of some amount, say 3 payoff units, if he can be promised a reward whose net present value is 3 regardless of what stage games will arrive in the future. Thus, we focus on strategy profiles for which this principle is sufficient to ensure compliance. These are the

strategy profiles that constitute a subgame-perfect equilibrium for every sequence of stage games that may be realized. We refer to such a profile as an *ex-post perfect equilibrium* (XPE). The central goal of our formal analysis, then, is to characterize what outcomes can be supported by an XPE.

It should be noted that XPE is *not* interpreted here as a positive prediction—there is no claim that players in an uncertain repeated game will always play an XPE.[1] Instead, it is a particular class of equilibria that is convenient to analyze, as doing so does not require probing into players' beliefs about the future stage games. In this way, it is similar to the study of dominant-strategy mechanisms in mechanism design—or to belief-free equilibria that have been studied elsewhere in the repeated games literature (Ely, Hörner and Olszewski, 2005; Hörner and Lovo, 2009). More specifically relevant here, since our interest is in adapting the toolkit from standard repeated games, XPE form a natural class to study because they have a familiar recursive structure: any XPE continues to be an XPE after any history. For readers wanting a more literal interpretation, we can also think of XPE as recommendations to players for how to play, such that players will be willing to follow the recommendations regardless of their beliefs about future stage games.

The question at hand is whether we can adapt the standard APS machinery to the uncertain repeated games framework to completely characterize all possible XPE outcomes. We will show how this can be done in a setting with 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.

Feature (i) is of course restrictive, but it still encompasses many situations of interest. As usual in the literature, this formulation allows multiple interpretations: the short-run players may be different individuals each period; they may be long-lived but completely

[1] In particular, XPE is not simply subgame-perfect equilibrium with players having some "non-standard" preferences. For example, assuming that players evaluate payoffs by the worst-case over future stage games would not lead to the concept of XPE.

We can actually use our results to give a positive interpretation to the set of XPE outcomes, but some subtlety is involved; see Section 5.4.

impatient individuals; or they may represent a continuum of small players, whose individual deviations are not detectable and so cannot be punished. What matters is that only one player can be given dynamic incentives. This long-run/short-run setting thus still has many applications; for example, it is used in much of the reputation literature following Fudenberg and Levine (1989), as well as the sustainable planning literature in macroeconomics following Chari and Kehoe (1990) (drawing on the many-small-players interpretation, where the one long-run player is the government).

Assumption (ii) is also crucial, as we shall see. In the standard setting, the recursive analysis can be performed either with or without this assumption (of course, the two cases may give different answers), but for us the assumption will be necessary.

Assumptions (iii) and (iv) are made mostly for ease of exposition: we wish to characterize equilibrium outcomes, and even describing outcomes becomes more complex the more sources of randomization there are, so it makes sense to focus on a setting with fewer such sources. However, many of the ideas here could be developed without these restrictions.

When either assumption (i) or (ii) is dropped, the analysis breaks down, and it does so in a specific way: optimal penal codes are no longer guaranteed to exist, as examples in Section 6 will show. Indeed, the appropriate concept of an optimal penal code for our setting is an XPE that is worst (for the player being punished) among all XPE's for every possible sequence of stage games simultaneously; its existence is not a foregone conclusion. To some readers, these negative results on the portability of the APS machinery will constitute the more important finding of the paper. Most of the paper's text will be devoted to proving the positive results under assumptions (i)–(iv). But the intended message lies in the combination of positive and negative results, rather than solely one or the other.

The analysis here is more than a formal exercise in generality; it also suggests revisiting standard intuitions about dynamic incentives. In particular, one of the major qualitative lessons from the theory of repeated games is that cooperation can be supported just as long as the gains from deviation do not exceed the available scope for punishment. Our analysis suggests this conclusion should be qualified, since the nonexistence of optimal penal codes with multiple long-run players (or no public randomization) means it is unclear what "scope for punishment" should mean in general.

To characterize the possible XPE outcomes under assumptions (i)–(iv), the key idea is as follows. We characterize the set of values of w such that it is possible to find two XPE profiles, one "reward" and one "punishment," such that the long-run player's payoff from

reward exceeds the payoff from punishment by at least w no matter what stage games are realized. This can be done by a version of the APS recursion: rather than perform the recursion in payoff space as usual, we do it in the space of "gaps" between two equilibria, and we show that it successfully identifies the largest such feasible gap. Action profiles for which the temptation to deviate exceeds this largest feasible gap can never occur in an XPE, since deviations cannot be punished. The remaining actions can occur, and we can use the worst such actions to construct an optimal penal code.

This leads directly to our main result, an explicit characterization of the outcomes that can be supported in XPE. As in the standard analysis, they are the outcomes for which the potential gain from deviating is always small enough to be deterred by the optimal penal code. In the leading case where on-path behavior does not condition on the public randomization, there is another natural description of such outcomes: they are the ones for which, at each period, the "debt" owed to the long-run player for forgoing short-run gains in past periods never accumulates beyond the amount that can definitely be repaid—that is, the maximum feasible gap. As a special case, when only one stage game is possible, this 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.

This paper constitutes an initial attempt at extending the toolkit for repeated games to an incompletely specified setting. Since the first version of the paper was circulated, a couple of other works have built on the framework developed here. Most directly related is Krasikov and Lamba (2023), mentioned above, which adopts the same framework but considers multiple long-run players. That paper considers the $\delta \to 1$ limit, rather than the fixed-δ setting studied here, and proves two versions of a folk theorem—one version that considers outcomes; and another in which the model is interpreted as uncertainty only on the *players'* part, while the analyst expects the stage games to follow a known Markov process, and the folk theorem therefore concerns long-run average payoffs. Additional work in progress by the same pair of authors considers instead multiple long-run players and fixed δ, and gives methods to construct various kinds of equilibria. (Example 6.1 of the present paper suggests that a complete characterization of the equilibrium outcomes may be unattainable in this setting.) Kostadinov (2023) also studies uncertain repeated games but assumes that players have a minmax-regret objective.

The next section presents an extended example that illustrates the formal framework and the questions under study. The rest of the paper proceeds linearly: the general model, analysis, and main results, all formulated under assumptions (i)–(iv) above; followed by

the counterexamples showing the difficulties with extending to multiple long-run players or dropping public randomization, and some brief discussion. Proofs omitted from the main text are in Appendix A.

2 Illustration

This example demonstrates the central question under study, as well as some key features of the analysis. The example is cast in the long-run/short-run setting that will occupy most of the paper.

Example 2.1. Consider first the game G shown in the left panel of Figure 1.[2] This payoff matrix represents the following story: Player 1 is a government agency (say, a tax collection agency), while player 2 is a citizen. The agency chooses the level of enforcement: high (h), low (l), or no enforcement (n). At the same time, the citizen chooses her level of compliance with the law: high (h), low (l), or no compliance (n). The agency prefers higher compliance by the citizen. It also finds high enforcement costly and prefers low enforcement, but no enforcement is undesirable (we might imagine, say, that with no enforcement, the agency cannot cover its own operating expenses). The citizen, for her part, prefers lower levels of compliance, except that if her compliance is below the enforcement level then she is caught and gets a payoff of 0. So her best reply is to match her compliance to the agency's enforcement level.

G :

	h	l	n
h	$6,1$	$3,0$	$0,0$
l	$8,1$	$5,2$	$2,0$
n	$0,1$	$0,2$	$0,3$

G' :

	h	l	n
h	$8,1$	$6,0$	$2,0$
l	$9,1$	$7,2$	$3,0$
n	$0,1$	$0,2$	$0,3$

Figure 1: Two versions of the enforcement game.

If this is played as a one-shot game, action l is dominant for the agency, so the unique Nash equilibrium is (l, l). Now suppose it is a repeated game, with past actions observed: player 1, the agency, is long-lived and maximizes the δ-discounted sum of stage payoffs; each period, there is a different citizen in the role of player 2 (or, in a perhaps more natural interpretation, there are many long-lived but small citizens playing as player 2,

[2] A game with a similar structure, presented in terms of a monetary policy application, appears in Stokey (1991).

who take 1's future actions as given and so just best-reply to 1's stage action each period). For each value of the parameter δ, the analyst would like to know whether good outcomes for the agency can be sustained in equilibrium. The payoff of 8 from (l, h) is off the table: given that player 2 myopically best-replies each period (and given our focus on pure strategies), only (h, h), (l, l), (n, n) can ever occur. So, as the next best hope, the analyst asks whether there is an equilibrium where (h, h) is played each period.

If $\delta \geq 2/3$, then this can indeed be achieved, by specifying that any deviation by the agency will be punished by reversion to stage-Nash (l, l) in all future periods. On the other hand, if $\delta < 2/3$, this doesn't work, since the short-run gain of 2 from deviating to l outweighs the loss of 1 in each future period from punishment.

However, one of the major lessons from the theory of repeated games is that more complex punishments can often be more effective than Nash reversion. This is true in the long-run/short-run setting, just as it is with multiple long-run players. And, in particular, it applies in game G: As long as $\delta \geq 1/3$, the (h, h) outcome can be supported by the "carrot-and-stick" punishment wherein, if the agency ever deviates, then (n, n) is played for one period, followed by a return to (h, h) in subsequent periods. If, during the punishment period, the agency again deviates, then we again punish with (n, n) for one more period (and then return to (h, h)), and so forth. This works because, both in the cooperation and in the punishment phases, the short-run gain of 2 from deviating is deterred by the loss of 6 in the following period due to punishment. Thus, in particular, when $1/3 \leq \delta < 2/3$, we can support the desirable outcome with carrot-and-stick punishments but not with Nash reversion. Note also that if $\delta < 1/3$ then there is no way to support the (h, h) outcome. Indeed, the agency can always secure itself at least 2 in every period by playing l, whereas its on-path payoff cannot be more than 6 each period; hence, a deviation can never be punished by more than a loss of 4 in all future periods. When $\delta < 1/3$, a short-run gain of 2 by deviating from (h, h) (or, for that matter, from (n, n)) outweighs this punishment, so in fact the only equilibrium outcome is stage-Nash (l, l) forever.

The game G', on the right in Figure 1, has a similar structure, but slightly different payoffs for the agency. Again, the citizen best-replies by matching the agency's action each period, and l is dominant for the agency in the stage game, so the stage Nash is (l, l). In the repeated version of G', an analysis similar to the above finds that (h, h) each period can be supported by carrot-and-stick punishments if $\delta \geq 3/8$—here, the short-run gain from deviation is 1 in the cooperative phase and 3 in the punishment phase, both of which are deterred by a loss of 8 next period. On the other hand, if $\delta < 3/8$, only stage

Nash is possible: First, the same logic as in game G shows that (n, n) can never be played in equilibrium, because there is not enough room to punish a deviation. Consequently, if we hope to support (h, h), the strongest possible punishment for deviation is reversion to Nash (l, l). But, as long as $\delta < 1/2$, this punishment is insufficient. So in fact the only equilibrium outcome is stage-Nash forever.

To summarize so far: if G is played forever, the high-compliance outcome (h, h) can be supported iff $\delta \geq 1/3$; if G' is played forever, it is supportable iff $\delta \geq 3/8$.

Now we turn to the setting of this paper: Suppose that each period, either G or G' will be played, but the analyst is unsure which one, and the stage game may alternate unpredictably across periods.[3] Each period, the players observe the current matrix and then choose their actions. The analyst would like to know, for each δ, whether the outcome of always playing (h, h) can be sustained, deterring deviations by punishments that are not sensitive to the agency's expectations about the future payoff matrices. That is: is the all-(h, h) outcome supportable in XPE?

The carrot-and-stick strategies still do the trick if $\delta \geq 1/2$. Indeed, in either matrix and in either phase (cooperation or punishment), the short-run gain from deviating is at most 3, whereas the loss the following period from being punished with (n, n) is at least 6, which outweighs the gain. Thus, we can see that there is never an incentive to deviate, even though we cannot actually compute the agency's total payoff without knowing the realized sequence of payoff matrices. Conversely, for $\delta < 1/2$, this punishment is not assured to work: when the time for punishment comes, if the current matrix is G' but the agency expects to be in G next period, then the short-run gain from playing l instead of the intended n outweighs the subsequent punishment. Consequently, the agency cannot be trusted to play h on-path, since it may not be possible to inflict the intended punishment following a deviation.

In fact, if $\delta < 3/7$, only the stage-Nash outcome is supportable in XPE. To see this, first suppose G' is to be played today. There is no way to get the players to play (n, n): if the agency anticipates G in all future periods, then the maximum punishment it could face is a loss of 4 in each future period (since the on-path payoff is at most 6, and the agency can secure at least 2 by playing l), which is not enough to deter the short-run gain of 3 from deviating. This means, in turn, that there is also no way to ever get (h, h) to be played when the current payoff matrix is G', nor is there any way to get (h, h) or (n, n)

[3]Given the motivation in the introduction, it might be more natural for the analyst to consider a set of payoff matrices defined by some bounds on the payoffs, but here it is expositionally simpler to consider just two possible matrices.

when the current matrix is G: If the agency were to expect G' in all future periods, given that (n, n) will never be played, the worst possible punishment for current deviations is (l, l); but in all three cases, this punishment is insufficient to deter the deviation. This leaves (l, l) as the only action profile that can ever be played in equilibrium, in either matrix.

(In particular, for $\delta \in [3/8, 3/7)$, the (h, h) outcome cannot be supported under uncertainty, even though it could be supported both in the repeated game with G played every period and with G' played every period. Actually, this phenomenon is not novel to the setting of uncertain repeated games; it can arise even if each period's stage game is drawn independently from a known distribution. A similar observation appears in Rotemberg and Saloner (1986, p. 406).)

Finally, what about $\delta \in [3/7, 1/2)$? It turns out that always playing (h, h) (both in G and in G') can indeed be supported in equilibrium, even though the carrot-and-stick punishments no longer do the trick. As we shall see later, it can be supported with more complex punishments that randomize the timing of the return to the cooperative phase (thus using the public randomization).

△

The following sections study the general version of the question explored here. The primitives are the set of possible stage game matrices and the discount factor. The analyst has in mind a particular outcome (in the example, this outcome involved always playing the same actions, but this need not be the case in general). Our goal is to give a test to discern whether or not this outcome is supportable in equilibrium.

3 Model

We first develop the model, using notation that will largely follow Mailath and Samuelson (2006), suitably adapted for our framework of uncertainty. We present the formalism under the assumptions of one long-run player and public randomization available, but the adaptations to drop these assumptions (for Section 6) are straightforward. We will then introduce some convenient notational simplifications.

3.1 Basic formulation

There are $n \geq 2$ players. Player 1 is a long-run player, with a discount factor $\delta \in (0, 1)$, 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 anyhow we 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 $u_i : A \to \mathbb{R}$. 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 full past history and the current 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 $s : \cup_t H^t \to A$, with the corresponding restriction $s(h^t) \in A(G^t)$. It will sometimes be useful to abbreviate a finite history of random signals $(\omega^0, \ldots, \omega^t)$ by $\omega^{0,\ldots,t}$, and to write $\mathbb{E}^t[\cdots]$ for the time-t expectation operator (i.e. the expectation conditional on signals $\omega^{0,\ldots,t}$).

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

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

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',

$$U_i(s|E, h^t) \geq U_i(s_i', s_{-i}|E, h^t). \tag{3.1}$$

The usual one-shot deviation principle applies: it suffices to have (3.1) 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

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

where, again, the expectation is over 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: (3.1) is satisfied for one-shot deviations by 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 $a_1' \in A_1(G^t)$ (where G^t is the period-t stage game). 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.[4] Later, we will indicate sufficient conditions on primitives to ensure that an XPE exists.

3.2 More convenient notation

We can apply a common 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,

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

Evidently, the constraints (3.1) 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 and focus 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 3.1. *Strategy profile s is an XPE if and only if both the following conditions hold:*

1. *for every history h^t, with stage game G^t arising at time t, we have $s(h^t) \in A^*(G^t)$;*
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 h^t, we have $U(s|E, h^t) \geq U(s'_1, s_{-1}|E, h^t)$.*

Notice that the set of XPE has a recursive structure: s is an XPE if it meets conditions 1 and 2 of the lemma 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 $\widehat{u}(a) = \max_{a'_1 \in A_1(G)} u(a'_1, a_{-1})$ for the stage payoff that would result from the myopically optimal deviation from a. (Here and henceforth, we take "myopically optimal deviation" to mean "conforming" when 1's action is

[4]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.

already myopically optimal.) Clearly $\widehat{u}(a) \geq u(a)$, and $\widehat{u}$ 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 $\widehat{u}(a)$ 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^-.

4 Analysis

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 $\overline{s}$ and $\underline{s}$, such that $U(\overline{s}|E) - U(\underline{s}|E) \geq w$ for every environment E; doing so then lets us rule out some action profiles because deviation cannot be prevented. We can find this largest w by applying the recursive machinery of APS in the space of gaps. This may appear to be simply a renormalization of the usual recursion done in payoff space, but the failure of the technique in the more general settings in Section 6 shows that this is not the case.

We can then use the surviving action profiles to construct our optimal penal code: Each period, we hit player 1 with the most punitive among these surviving action profiles; to ensure player 1's compliance with this punishment, we also have to offer him future rewards that are just barely adequate to deter deviation. Although this construction may seem straightforward, it is notably different than optimal penal codes in the standard setting, as will be discussed further in Section 6.

As a side note, the gap approach here has antecedent in Cronshaw and Luenberger (1994). They study symmetric games (without uncertainty) and characterize SPE payoffs of strongly symmetric equilibria. They also observe that incentives depend on the difference between the best and worst continuation payoffs, rather than the payoff levels; their "maximal deterrence" corresponds to the "maximal feasible gap" here. Although they perform the recursion in payoff space, their equation (7) characterizing the maximal deterrence corresponds to our fixed-point condition $w^* = B(w^*)$ below.

4.1 Recursive technique

In APS, the recursive 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 (the smallest gap between two XPE's) is zero. So the set is described by a single number—its upper bound—and thus our operator will just map nonnegative numbers to nonnegative numbers.

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

$$A^*(G, w) = \left\{ a \in A^*(G) \;\middle|\; \widehat{u}(a) - u(a) \leq \frac{\delta}{1-\delta} w \right\}.$$

Now define

$$B(w; G) = (1-\delta) \left(\max_{a \in A^*(G,w)} u(a) - \min_{a' \in A^*(G,w)} \widehat{u}(a') \right) + \delta w. \tag{4.1}$$

(If $A^*(G, w)$ is empty, then take $B(w; G) = -\infty$. As long as $A^*(G, w)$ is nonempty, it is closed, so the max and min exist by continuity; note also $B(w; G) \geq 0$ in this case.)

Intuitively, $A^*(G, w)$ is the set of action profiles that are enforceable when G is played in the current period, given that the continuation gap available starting next period between "good" and "bad" payoffs is at most w. In turn, $B(w; G)$ represents an upper bound on the current gap, given that G is played initially and the continuation gap is at most w. To see the latter: 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 (4.1).

The above argument sketches why the expression in (4.1) is an upper bound on the payoff gap between two strategy profiles. To show that this upper bound is attainable, we will need to be able to set the *on-path* continuation payoff from the "bad" strategy profile appropriately, which can be done using the public randomization; we will make this idea precise when it is needed in Section 4.3.

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 the available continuation gap is at most w.

Notice that $B(w;G)$ is strictly increasing in w at a rate of at least δ. Therefore, $B(w)$ is as well.

We now adopt an assumption that will be maintained throughout the rest of this section and the next:

Assumption 4.1. *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 4.1 is satisfied:

1. For every $G \in \mathcal{G}$, there exists $a \in A^*(G)$ such that $\widehat{u}(a) = u(a)$ (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 $u(a) \geq \widehat{u}(a') + \epsilon$, 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 4.1 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:

Lemma 4.2. *The functions $B(w;G)$ and $B(w)$ are right-continuous in w.*

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^*.

For technical reasons, it will actually be useful to take a slightly different sequence, one in which w_{k+1} is strictly above $B(w_k)$. Specifically:

Lemma 4.3. *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 $w^* + \epsilon$. (This stronger statement is useful for results discussed later, in Section 5.4.)

Lemma 4.4. *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 $\overline{s}$ and $\underline{s}$ for this environment,*

$$U(\overline{s}|E) - U(\underline{s}|E) < w^* + \epsilon.$$

The proof uses the sequence from Lemma 4.3. 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} (given by the induction hypothesis), then the total payoff gap cannot exceed w_k.

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 4.3.

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

Proof. It suffices to prove this for date-0 histories. Consider any initial stage game G^0 and any $\epsilon > 0$. Consider any environment E that begins with G^0 followed by the finite sequence of stage games given by Lemma 4.4. Suppose s is an SPE for this environment,

and let a^0 be the action profile played at some time-0 history $h^0 = (G^0, \omega^0)$. Player 1's payoff is then

$$U(s|E, h^0) = (1-\delta)u(a^0) + \delta U(s|E, h^0, a^0),$$

whereas if player 1 deviates to the myopically action a_1' at time 0 and then follows s, the payoff is

$$(1-\delta)\widehat{u}(a^0) + \delta U(s|E, h^0, (a_1', a_{-1}^0)).$$

Deviating cannot be beneficial; this gives (after rearranging)

$$(1-\delta)(\widehat{u}(a^0) - u(a^0)) \leq \delta \left(U(s|E, h^0, (a_1', a_{-1}^0)) - U(s|E, h^0, a^0) \right).$$

But both $U(s|\cdots)$ terms on the right-hand side are SPE payoffs in the continuation environment, so by Lemma 4.4, they 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 $\widehat{u}(a^0) - u(a^0) \leq \frac{\delta}{1-\delta}(w^* + \epsilon)$. Since $\epsilon > 0$ is arbitrary, the right side can be replaced by $\frac{\delta}{1-\delta}w^*$, implying the result.

□

As a side observation, there may be nontrivial interactions between the different stage games in $\mathcal{G}$ 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 $w \mapsto B(w; G)$. Then, w^* may be bounded strictly below all of the $w^*(G)$. This also implies that the adversarial environments constructed in Lemma 4.4 may need to have the stage game vary from one period to the next.

In fact, we effectively saw this in the opening Example 2.1, with the two stage games G and G'. Suppose $\delta = 2/5$. Then, the high-compliance outcome (h, h) is supportable in either stage game individually, but not in the uncertain repeated game. This is reflected in the w^* values: one can check that $w^*(G) = 4$ and $w^*(G') = 5$, but in the uncertain repeated game, $w^* = 0$ (and thus any action profile other than stage-Nash is ruled out according to Lemma 4.5).

4.2 Quasi-minmax payoffs

Lemma 4.5 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

$$\overline{a}(G) \in \operatorname*{argmax}_{a \in A^*(G,w^*)} u(a); \qquad \underline{a}(G) \in \operatorname*{argmin}_{a \in A^*(G,w^*)} \widehat{u}(a).$$

(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

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

Lemma 4.6. *If s is an XPE, then for any environment E,*

$$U(s|E) \geq \underline{U}(E).$$

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 4.5, 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 $\widehat{u}(\underline{a}(G^t))$ in this period. Summing across all periods shows that 1's payoff from the repeated deviation is at least $\underline{U}(E)$. Hence, the payoff from conforming to s is at least this much. □

We can think of $\widehat{u}(\underline{a}(G))$ as a "quasi-minmax" payoff for player 1 when the stage game is G, providing a straightforward lower bound on 1's equilibrium payoffs. It differs from the usual minmax in two ways. First, the min is taken only over a restricted set of action profiles, those in $A^*(G, w^*)$. This is natural; because we are not in a folk theorem setting but are considering a fixed δ, some action profiles are ruled out as unenforceable. And second, the action profile that actually produces the stage payoff of $\widehat{u}(\underline{a}(G))$ typically cannot be played in equilibrium, as it does not satisfy the incentive constraints of the short-run players. This is again familiar from the literature on repeated games with short-run players, such as Fudenberg, Kreps and Maskin (1990) (though they nonetheless use the term "minmax value" for the analogous quantity).

4.3 Automaton strategies

We will next show that the lower bound in Lemma 4.6 is tight. In fact, we will construct a *single* XPE that attains this lower bound simultaneously for all environments E. This

will imply, in particular, that it is an optimal penal code—that is, it is the harshest punishment available to deter deviations by player 1, regardless of the environment.

Given the definition of $\underline{U}(E)$, we might wish to construct this XPE by simply giving player 1 a payoff of $\widehat{u}(\underline{a}(G^t))$ in each period t. But, again, this payoff comes from a deviation from $\underline{a}(G^t)$, and the deviated action profile generally cannot be played on the equilibrium path since it is not incentive-compatible for the short-run players. Instead, we specify that $\underline{a}(G^t)$ is played in each period t, but then keep track of the fact that player 1 needs to be "compensated" in the future for the difference $\widehat{u}(\underline{a}(G^t)) - u(\underline{a}(G^t))$. We provide this compensation by sometimes playing the reward $\overline{a}(G^t)$ instead of $\underline{a}(G^t)$, using the public randomization to ensure that just the right amount of reward is given.

This idea can naturally be formalized as an automaton, as in Mailath and Samuelson (2006, Section 2.3),[5] where the automaton's state variable represents the amount of compensation that is currently owed to player 1. Thus, the automaton enters each period t in some state. After the stage game G^t and public random signal ω^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 $2, \ldots, n$, we can focus on state transitions that depend only on 1's action.

More specifically, the state space for our automaton is the interval $W = [0, w^*]$. The main elements to be specified are the action (output) function $f : W \times \mathcal{G} \times [0,1] \to A$ (which, of course, must output an action profile 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

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

The denominator equals $B(w^*; G) \geq B(w^*) = w^*$, so for any $w \in W$, we have $\lambda(G)w \in [0,1]$. (The denominator of $\lambda(G)$ may be zero, but only if $w^* = 0$, in which case $w = 0$. In this case, interpret $\lambda(G)w$ as 0 throughout the following.)

Now, for any w, G, ω, a:

[5]Section 5.7 of Mailath and Samuelson (2006) may seem to be the more relevant section to cite: that section develops automata for dynamic games. However, it is assumed there that 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

$$\tau(w, G, \omega, a) = \begin{cases} w^* & \text{if } a_1 = \overline{a}_1(G), \\ w^* - \frac{1-\delta}{\delta}(\widehat{u}(\overline{a}(G)) - u(\overline{a}(G))) & \text{otherwise;} \end{cases}$$

- If $\omega > \lambda(G)w$: Put $f(w, G, \omega) = \underline{a}(G)$, and

$$\tau(w, G, \omega, a) = \begin{cases} \frac{1-\delta}{\delta}(\widehat{u}(\underline{a}(G)) - u(\underline{a}(G))) & \text{if } a_1 = \underline{a}_1(G), \\ 0 & \text{otherwise.} \end{cases}$$

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 τ 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 $s[w]$. We then have the following result (whose proof is a bit lengthy but routine):

Proposition 4.7. *Pick any $w \in [0, w^*]$, and let E be any environment. Then:*

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

As promised, this result gives us an XPE that attains the lower bound payoff $\underline{U}(E)$ in every environment E—namely, $s[0]$. Moreover, with this result we are now fully justified in thinking of w^* as the largest sustainable reward-punishment gap, since we do indeed have two XPE's whose payoffs differ by w^* in every environment—namely, $s[w^*]$ and $s[0]$.

5 Main results

With this machinery in hand, we are ready to return to our main question of interest: what outcomes can arise in equilibrium?

5.1 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 that there exists a "true" (but initially unknown) environment; the outcome should then consist of this realized environment, together with the action profiles played in each period. Of course, the latter may be random, since they can condition on the public signals. Accordingly, we define a *realizable outcome* as a pair (E, z), where $E = (G^0, G^1, \ldots)$ is an environment, and $z : \cup_{t=0}^{\infty}[0,1]^{t+1} \to A$, specifying an action profile $z(\omega^{0,\ldots,t}) \in A(G^t)$ for each date and history of public signals.[6] In the special case where the time-t action is independent of the random signals for each t, we call the outcome *deterministic*; such an outcome can simply be described by a single stage game G^t and action profile $a^t \in A(G^t)$ in each period. The reader may find it useful to focus on deterministic outcomes for concreteness, but we will state results for the general case since they are not much more involved.

An alternative perspective is to view an outcome as a full description of the actions that may be played "on-path," for whatever environment may be realized. Accordingly, define a *full outcome* to be a function $z : \cup_{t=0}^{\infty}(\mathcal{G} \times [0,1])^{t+1} \to A$, specifying an action profile $z(G^0, \omega^0, \ldots, G^t, \omega^t) \in A(G^t)$ for each possible initial sequence of stage games and public signals. We again say that such an outcome is *deterministic* if actions are always independent of the signals. A realizable outcome (E, z'), where $E = (G^0, G^1, \ldots)$, *belongs to* the full outcome z if $z(G^0, \omega^0, \ldots, G^t, \omega^t) = z'(\omega^{0,\ldots,t})$ for all t and $\omega^{0,\ldots,t}$.

The outcome examined in Example 2.1 was an example of a full outcome. However, both definitions are arguably reasonable. It will be simpler in terms of exposition to begin by focusing on realizable outcomes. Section 5.3 will state the corresponding results for full outcomes.

Let us say that a strategy profile s *supports* the realizable outcome given by (E, z) if, for all t and $\omega^{0,\ldots,t}$, if we define $a^{t'} = z(\omega^{0,\ldots,t'})$ for each $t' \leq t$ then s satisfies

$$s(G^0, \omega^0, a^0; G^1, \omega^1, a^1; \ldots, G^t, \omega^t) = a^t.$$

Thus, in the environment E, actions on-path are chosen as specified by z. We similarly say that s *supports* a full outcome z if it supports every realizable outcome that belongs to z.

[6] One might alternatively just define a realizable outcome as consisting of an environment and a joint distribution over $(a^0, a^1, \ldots)$. This description would be incomplete, since it does not specify, for example, whether the uncertainty about a^5 is resolved based on the realization of ω^5, or is resolved already by observing ω^0, or anywhere in between.

Note that even if an outcome is deterministic, randomization off-path may be needed to support it in equilibrium.

5.2 Supportable outcomes

The following are readily shown to be necessary conditions for a realizable outcome $(G^0, G^1, \ldots; z)$ to be supported by an XPE s:

$$z(\omega^{0,\ldots,t}) \in A^*(G^t, w^*) \qquad \text{for all } t \text{ and } \omega^{0,\ldots,t}; \tag{5.1}$$

$$\left(\widehat{u}(a^{\underline{t}}) - u(a^{\underline{t}})\right) + \sum_{t=\underline{t}+1}^{\bar{t}} \delta^{t-\underline{t}} \left(\widehat{u}(\underline{a}(G^t)) - \mathbb{E}^{\underline{t}}[u(a^t)]\right) \;\leq\; \frac{\delta^{\bar{t}+1-\underline{t}}}{1-\delta} w^* \tag{5.2}$$
$$\text{for all } \underline{t} < \bar{t} \text{ and } \omega^{0,\ldots,\underline{t}},$$

where $a^{\underline{t}} = z(\omega^{0,\ldots,\underline{t}})$, and likewise a^t for $t > \underline{t}$.

(An equivalent formulation is simply that $a^t \in A^*(G^t)$ for all t and (5.2) holds for all $\underline{t} \leq \bar{t}$, where the sum is empty if $\underline{t} = \bar{t}$.)

Indeed, we have already seen that (5.1) is necessary. To understand (5.2), take any interval of dates from $\underline{t}$ from $\bar{t}$, and imagine that at date $\underline{t}$, player 1 deviates by myopically best-replying to the short-run players' actions; then suppose that, at each period $t = \underline{t}+1, \ldots \bar{t}$, we try to punish player 1 by playing $\underline{a}(G^t)$, and player 1 continues myopically best-replying. The left-hand side of (5.2) represents the total discounted net gain from the repeated deviation over this interval (from the perspective of time $\underline{t}$). The right-hand side represents the maximum feasible gap, suitably discounted. Thus, condition (5.2) says that the gain from the repeated deviation should be small enough to be deterred even if the environment after time $\bar{t}$ turns out to be one where punishment is difficult, instead of the target environment. (We formalize this argument below.) Notice also that the terms $\widehat{u}(\underline{a}(G^t)) - \mathbb{E}^{\underline{t}}[u(a^t)]$ may be positive or negative, so it is unknown a priori for which pairs $(\underline{t}, \bar{t})$ the constraint will be tightest.

Our first main result is that conditions (5.1)–(5.2) actually give a complete characterization of the realizable outcomes that can be supported in XPE.

Theorem 5.1. *Let $E = (G^0, G^1, \ldots)$ be an environment. A realizable outcome (E, z) is supported by some XPE s if and only if it satisfies the necessary and sufficient conditions (5.1)–(5.2).*

To see sufficiency, we can construct a strategy profile that starts out following z and punishes deviations using the optimal penal code $s[0]$. However, we also need to specify what happens if "Nature deviates," that is, if the environment ends up differing from the target E. In this case, we swap to $s[w^*]$: the automaton strategy in the high-reward state. This ensures that, in *every* environment and starting at any date, player 1's future payoff from conforming is high enough so that the threat of the $s[0]$ punishment deters deviations.

Proof. For necessity of the conditions: (5.1) is given by Lemma 4.5. For (5.2), take any $\epsilon > 0$, and consider the environment E' that consists of $(G^0, \dots, G^{\bar{t}})$, followed by the sequence of stage games identified in Lemma 4.4 for this ϵ (and arbitrary stage games thereafter). Consider player 1's decision at time $\underline{t}$, with history $h^{\underline{t}}$. Conforming to s gives a payoff

$$U(s|E', h^{\underline{t}}) = (1-\delta)\left(\sum_{t=\underline{t}}^{\bar{t}} \delta^{t-\underline{t}} \mathbb{E}^{\underline{t}}[u(a^t)]\right) + \delta^{\bar{t}+1-\underline{t}} \mathbb{E}^{\underline{t}}[U(s|E', h^{\bar{t}}, a^{\bar{t}})]. \tag{5.3}$$

(Here, $h^{\bar{t}}$ represents the history arising at period $\bar{t}$.)

An alternative strategy s_1' would play a myopic best reply to the short-run players' anticipated actions at each period $t = \underline{t}, \dots, \bar{t}$, and then follow s_1 from date $\bar{t}+1$ onward. This would give a stage payoff of $\widehat{u}(a^{\underline{t}})$ in period $\underline{t}$, and would guarantee at least $\widehat{u}(\underline{a}(G^t))$ in each period $t = \underline{t}+1, \dots, \bar{t}$. So player 1's deviation payoff satisfies

$$U(s_1', s_{-1}|E', h^{\underline{t}}) \geq (1-\delta)\left(\widehat{u}(a^{\underline{t}}) + \sum_{t=\underline{t}+1}^{\bar{t}} \delta^{t-\underline{t}} \widehat{u}(\underline{a}(G^t))\right) + \delta^{\bar{t}+1-\underline{t}} \mathbb{E}^{\underline{t}}[U(s|E', \widetilde{h}^{\bar{t}}, \widetilde{a}^{\bar{t}})] \tag{5.4}$$

(where $\widetilde{h}^{\bar{t}}$ and $\widetilde{a}^{\bar{t}}$ denote the history and period-$\bar{t}$ actions produced by 1's deviations). Since the deviation should not be profitable, (5.3) should be greater than or equal to (5.4). Subtracting, dividing by $1-\delta$, and rearranging terms, we find that the left-hand side of (5.2) is bounded above by

$$\frac{\delta^{\bar{t}+1-\underline{t}}}{1-\delta}\left(\mathbb{E}^{\underline{t}}[U(s|E', h^{\bar{t}}, a^{\bar{t}})] - \mathbb{E}^{\underline{t}}[U(s|E', \widetilde{h}^{\bar{t}}, \widetilde{a}^{\bar{t}})]\right).$$

However, the two $U(s|\cdots)$ terms both represent SPE payoffs in the environment starting at date $\bar{t}+1$, so by Lemma 4.4, they differ by less than $w^* + \epsilon$. Applying this bound and taking $\epsilon \to 0$ gives (5.2).

For sufficiency, suppose realizable outcome (E, z) satisfies (5.1) and (5.2). Construct a strategy profile s as follows:

(i) At any history h^t such that all stage games $(G^0, \ldots, G^t)$ so far have been consistent with E and all actions so far $(a^0, \ldots, a^{t-1})$ have been as prescribed by z, play as specified by z.

(ii) For any history h^t where the stage games and action profiles through time $t-1$ were all as specified by (E, z), but the period-t stage game is different, play according to $s[w^*]$ from h^t onward.

(iii) For any history h^t where all past stage games through time $t-1$ and all action profiles through time $t-2$ were as specified by (E, z), but the action profile observed at $t-1$ was different from that prescribed by z, play according to $s[0]$ from period t onward.

Notice that every history either falls into case (i) or has a unique initial subhistory covered by case (ii) or (iii), so this description does specify a well-defined strategy profile. By construction it supports (E, z); we need to check that it is an XPE.

At any history where any stage game or past action has differed from (E, z), there is no incentive to deviate; this follows because we already know that $s[w^*]$ and $s[0]$ are XPE's. So we only need to check incentives to deviate at histories in case (i). Moreover, the short-run players' incentives are satisfied by (5.1), so we need only check player 1's incentives.

Consider such a history h^t, and any environment $\widetilde{E}$ consistent with it. Suppose $\widetilde{E} \neq E$. Let $\bar{t}+1$ be the earliest period in which $\widetilde{E}$ and E differ. So, writing $E = (G^0, G^1, \ldots)$ as usual, then $\widetilde{E}$ begins $(G^0, G^1, \ldots, G^{\bar{t}}, \widetilde{G}^{\bar{t}+1}, \ldots)$. Evidently $\bar{t} \geq t$.

If $\bar{t} = t$, then by conforming when asked to play a^t, player 1 achieves a payoff (from the period-t vantage point) of $(1-\delta)u(a^t) + \delta(\underline{U}(\widetilde{E}^{-(\bar{t}+1)}) + w^*)$, since play transitions to $s[w^*]$ next period. By deviating, player 1's payoff is $(1-\delta)\widehat{u}(a^t) + \delta\underline{U}(\widetilde{E}^{-(\bar{t}+1)})$ (or less, if a non-optimal deviation is chosen). So the overall gain from deviating is $(1-\delta)(\widehat{u}(a^t) - u(a^t)) - \delta w^*$, which is ≤ 0 by condition (5.1).

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

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

By deviating, player 1's payoff is $(1-\delta)\widehat{u}(a^t)+\delta\underline{U}(\widetilde{E}^{-(t+1)})$. Expanding using the definition of $\underline{U}$, we get $\underline{U}(\widetilde{E}^{-(t+1)}) = (1-\delta)\left(\sum_{t'=t+1}^{\bar{t}}\delta^{t'-(t+1)}\widehat{u}(\underline{a}(G^{t'}))\right) + \delta^{\bar{t}-t}\underline{U}(\widetilde{E}^{-(\bar{t}+1)})$, and so the deviation payoff equals

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

Now condition (5.2) (with t in place of $\underline{t}$) implies that the deviation is unprofitable in environment $\widetilde{E}$.

Finally, what about if $\widetilde{E} = E$? In this case, for each $t' > t$, let $\widetilde{E}^{t'}$ 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 $\widetilde{E}^{t'}$. So, we have already shown that, for any proposed deviating strategy s_1', $U(s|\widetilde{E}^{t'}, h^t) \geq U(s_1', s_{-1}|\widetilde{E}^{t'}, h^t)$. Taking limits as $t' \to \infty$ gives $U(s|E, h^t) \geq U(s_1', s_{-1}|E, h^t)$: thus, the deviation is not profitable in environment E.

□

A few remarks are in order.

First, as a special case when $\mathcal{G} = \{G\}$ is a singleton, we can cover the case of a standard repeated game with a single long-run player and public randomization; our analysis so far identifies the long-run player's worst SPE payoff (which reduces simply to $\widehat{u}(\underline{a}(G))$) and characterizes the supportable outcomes. This does not seem to be noted in existing literature.

Second, 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 (5.2), we get

$$\left(\widehat{u}(a^{\underline{t}}) - u(a^{\underline{t}})\right) + \sum_{t=\underline{t}+1}^{\infty}\delta^{t-\underline{t}}\left(\widehat{u}(\underline{a}(G^t)) - \mathbb{E}^{\underline{t}}[u(a^t)]\right) \leq 0 \qquad \text{for all } \underline{t}, \omega^{0,\ldots,\underline{t}}. \tag{5.5}$$

This condition says that the payoff from following the target outcome, beginning in period $\underline{t}$, is at least as high as that from a one-period deviation followed by the ensuing punishment. In repeated games, the corresponding condition is in fact sufficient for supportability in SPE (see Abreu, 1988, Proposition 4). Here, we need a condition indexed both by $\underline{t}$ and $\bar{t}$ because of the possibility that the realized environment eventually differs from E. That is, if the target realizable outcome involves a large temptation to deviate at period $\underline{t}$ but also large rewards promised at some future period t', then it may satisfy (5.5) but nonetheless not be supportable because, at time $\underline{t}$, we cannot yet be sure that

large rewards at time t' will actually be available, and thus the deviation at time $\underline{t}$ cannot be deterred.

Third, at least in the *deterministic* case, we can rewrite the conditions in a way that offers an alternative interpretation. Given a deterministic realizable outcome (E, z), recursively define $d^{-1}(z) = 0$ and

$$d^t(z) = \max\left\{\widehat{u}(a^t) - u(a^t), \frac{1}{\delta}d^{t-1}(z) + \left(\widehat{u}(\underline{a}(G^t)) - u(a^t)\right)\right\}$$

for $t = 0, 1, \ldots$. Then we have:

Proposition 5.2. *A deterministic realizable outcome (E, z), with $E = (G^0, G^1, \ldots)$ and $z = (a^0, a^1, \ldots)$, is supported by some XPE if and only if it satisfies $a^t \in A^*(G^t)$ for all t, and $d^t(z) \leq \frac{\delta}{1-\delta}w^*$ for all t.*

We can think of d^t as measuring the "debt" owed to player 1 for refraining from deviation in the past. In each period t, the debt repayment promised in the future needs to be large enough to deter a one-shot deviation at t; it also be large enough to cover any previously existing debt, with "interest," adjusted by whatever portion is being delivered in the present period. These two considerations correspond to the two terms of the max. The proposition then says that an outcome can be supported in XPE just so long as the debt owed never exceeds the amount that can be promised.

Fourth, 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 $\overline{\mathcal{G}}$ be some "universe" of potential stage games, and $u_i : \overline{A} \to \mathbb{R}$ the corresponding payoff functions, satisfying the assumptions of Section 3 (where $\overline{A}$ is the disjoint union of the sets $A(G)$ for $G \in \overline{\mathcal{G}}$). Let (E, z) be a realizable outcome and $\underline{\mathcal{G}} \subseteq \overline{\mathcal{G}}$ such that each stage game G^t of E lies in $\underline{\mathcal{G}}$. We consider various sets $\mathcal{G}$ with $\underline{\mathcal{G}} \subseteq \mathcal{G} \subseteq \overline{\mathcal{G}}$; any choice of such a $\mathcal{G}$ specifies the possible stage games in such a way that the environment E can occur. Under what conditions on $\mathcal{G}$ will it be the case that (E, z) is supportable in XPE over $\mathcal{G}$? For the deterministic case, Proposition 5.2 gives

an answer: this happens if and only if the value of w^* for $\mathcal{G}$ is greater than or equal to $\frac{1-\delta}{\delta} \sup_t d^t(z)$; equivalently, if and only if there exists some $w \geq \frac{1-\delta}{\delta} \sup_t d^t(z)$ such that $B(w; G) \geq w$ for each $G \in \mathcal{G}$. (And likewise, for the more general case, this condition must hold for some w large enough to satisfy (5.1)–(5.2) along all signal histories.)

5.3 Full outcomes

Our main characterization, Theorem 5.1, has an analogue for full rather than realizable outcomes.

Evidently, a full outcome z can only be supported in XPE if each of the realizable outcomes belonging to it can, or equivalently, if each such realizable outcome satisfies (5.1)–(5.2). The converse is also true, and thus:

Theorem 5.3. *A full outcome z is supported by some XPE if and only if each of the realizable outcomes belonging to it can be supported by some XPE.*

Actually, more can be said. Recall that (5.2) implied (5.5), by taking the limit as $\bar{t} \to \infty$. For full outcomes, we can replace (5.2) with this weaker condition:

Theorem 5.4. *A full outcome z is supported by some XPE if and only if each of the realizable outcomes belonging to it satisfies (5.1) and (5.5).*

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

The proof that these conditions are sufficient is quite simple: play proceeds according to z on-path, and any deviation by player 1 is punished by using the optimal penal code $s[0]$. If this strategy profile fails to constitute an XPE, then it fails to be an SPE in some particular environment. This implies that no other punishment can support the corresponding realizable outcome of z in that environment either—just as in standard repeated games—and thus z cannot be an XPE outcome.

5.4 Relating XPE to SPE outcomes

Although this paper has introduced the concept of XPE as methodologically motivated and not presented it as a positive prediction of play, we can actually use the characterization results to draw a tight connection between XPE outcomes and positive predictions.

This discussion is peripheral to our main goal, so we give just an overview here; the details are in Section S-1 of the Supplemental Material.

Suppose that an analyst is interested in predicting the possible outcomes of the repeated game. From this analyst's point of view, there is no a priori reason to focus on XPE; instead, the long-run player presumably has some attitude toward the uncertainty about the future stage games, and optimizes with respect to it. A natural model would be that the long-run player has some belief about the stochastic process governing the stage games, and he plays so as to maximize expected discounted utility with respect to this belief. (Short-run players' beliefs are irrelevant since they act myopically.) Any such belief then defines a dynamic game, and we can consider subgame-perfect equilibria of this game. The question is then: for the analyst, who believes SPE is the correct description of behavior but doesn't know what the long-run player's belief is, does the set of XPE outcomes have a meaningful interpretation?

Every XPE strategy profile is automatically an SPE regardless of the long-run player's belief, and therefore, every *outcome* supportable in XPE is also supportable in SPE regardless of the belief. If the converse were also true, we would have a positive interpretation for the set of XPE outcomes: they are precisely the outcomes that the analyst is sure can be supported, despite her ignorance of the long-run player's belief.

Unfortunately, this converse statement is not true: there can be outcomes that can be supported in SPE for every possible belief but not in XPE. (To put it differently, they can always be supported by some punishments for devations, but the punishments have to depend on what the belief is.) Example S-1.1 in the Supplemental Material illustrates this. A rough intuition is as follows. The argument for necessity of (5.2) for XPE outcomes relied on two things: first, that at time $\underline{t}$, the subsequent stage games $G^{\underline{t}+1}, \ldots, G^{\bar{t}}$ are likely to arise over the ensuing periods (so that the payoffs from the target outcome are indeed relevant); and second, that at each time t during this interval, the adversarial environment from Lemma 4.4 has a chance of arriving starting in the next period (otherwise, time-t actions are not confined to $A^*(G^t, w^*)$, so punishments worse than $\underline{a}(G^t)$ might be available). But a probabilistic belief cannot simultaneously place high probability on both of these continuation environments. Thus, for SPE outcomes, we no longer get (5.2) as a necessary condition.

However, there is nothing about our setup that obliges us to assume a probabilistic belief about the stage games. It is natural to instead allow that the long-run player treats the future stage games as ambiguous. In the supplement, we formulate this more general model. There are many possible ways one might specify the nature of the ambiguity and

the long-run player's attitude to it. We therefore consider the broad class of *dynamic variational preferences* (Maccheroni, Marinacci and Rustichini, 2006): this is a large class of ambiguity-averse preferences that is naturally compatible with the discounting structure of repeated games. It includes many common specifications of ambiguity aversion, such as maxmin expected utility and multiplier preferences. For any such preferences, it remains the case that every XPE is automatically an SPE,[7] and so every XPE outcome is an SPE outcome. And here we do have the converse: Theorem S-1 shows that every outcome that is supportable in SPE for all dynamic variational preferences is in fact supportable in XPE. (Our focus is mainly on realizable outcomes, but a corresponding result, Theorem S-2, holds for full outcomes.) Thus, the set of XPE outcomes does indeed have an interpretation, even for an analyst who views SPE as the appropriate prediction.

The result uses our characterization in Theorem 5.1: if an outcome violates one of the conditions (5.1)–(5.2), we construct a specific dynamic variational preference for which the outcome is not supportable in SPE. Moreover, this preference has a simple interpretation. If (5.1) is violated for some date t, then it suffices to have expected utility where the long-run player expects that, from time $t+1$ onward, the stage games follow the adversarial environment from Lemma 4.4. The more interesting case is when (5.2) is violated for some pair of dates $\underline{t}, \bar{t}$. We then construct the preference as follows: first, additively renormalize the payoffs in every stage game G so that $\widehat{u}(\underline{a}(G)) = 0$; then, specify that long-run player entertains ambiguity such that any sequence of stage games is possible up until time $\bar{t}$, the stage games after $\bar{t}$ definitely follow the environment from Lemma 4.4, and payoff streams are evaluated using the worst case over the uncertain early stages.

6 Discussion

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. As we will show here, the results change if either of these conditions is removed; in particular, an optimal penal code may no longer exist.

The construction of an optimal penal code in Section 4.3 involved a simple idea: to punish player 1, play the maximally punitive actions among those that are enforceable, and give just enough future rewards to player 1 to deter deviation. If either of our restrictions

[7] This observation also appears in Krasikov and Lamba (2023).

are dropped, this construction fails because it may be necessary to "overreward" player 1: with multiple long-run players, it may be necessary to reward other players in future periods for participating in 1's punishment, and it may be impossible to do so without also being generous to player 1; even with just one long-run player but no public randomization, ensuring 1's own compliance may require overcompensating due to the discreteness of available rewards. Indeed, in the standard repeated-games framework, these issues can arise and, in general, the optimal punishment for player 1 does not necessarily begin with the action profile that minimizes 1's myopic-deviation payoff among enforceable action profiles.

In uncertain repeated games, if we wish to punish player 1 starting at time t for a deviation at time $t-1$, whether or not the overreward issue arises depends on the future stage games. Consequently, to deliver the harshest punishment from time t forward, which action should be played at t can depend on what stage game will arise at $t+1$. This is why optimal penal codes need not exist in general.

When an optimal penal code does exist, it also serves as what we might call a "universal" penal code: any outcome that can be supported by some punishment can, in particular, be supported by this one. The notions of an optimal penal code and a universal penal code are closely related but not identical; one can give examples of repeated games where there are universal penal codes that are not optimal. And the key role played by $s[0]$ in characterizing equilibrium outcomes in Theorems 5.1 and 5.4 really relies on its universality, not its optimality per se. However, we will show in the examples below that universal penal codes also fail to exist. That is, there is one outcome that can be supported by one off-path punishment, and a different outcome that can be supported by a different punishment, but no single punishment that supports both. Given the lack of a universal penal code, this suggests that any characterization of XPE outcomes in these settings, if one can be given at all, would have to look quite different.

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 $\mathcal{G}^t$); we also assume a finite horizon, and no discounting. None of these changes matters conceptually. For completeness, Section S-2 of the Supplemental Material gives more elaborate examples that express the same ideas while retaining stationarity and discounting.

For brevity, we skip giving a full formal presentation of the model. Also, action profiles will be notated without punctuation, e.g. aq rather than (a, q).

6.1 Multiple long-run players

We first consider an example with two long-run players, who each act to maximize the sum of payoffs across periods. Our example features three periods, $t = 0, 1, 2$. If we wish to punish player 1 starting at date 1 for a date-0 deviation, we may or may not face the overreward problem, depending which stage game will be realized in period 2. Consequently, there is no XPE that gives the lowest payoff starting from date 1 in every environment, and in turn, no universal penal code.

Example 6.1. 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.

G^0:

	q
a	$0, 0$
b	$1, 0$

$\rightarrow$ G^1:

	r	s
c	$1, 1$	$0, 0$

$\nearrow$ $\searrow$

G^2:

	t	u	v
d	$0, 0$	$0, 0$	$0, 0$
e	$0, 0$	$1, 0$	$0, 0$
f	$0, 0$	$0, 0$	$3, 1$

$G^{2'}$:

	w	x
g	$0, 0$	$1, 0$
h	$0, 1$	$1, 1$

Figure 2: Example with two long-run players. No universal penal code exists.

Focus first on the subgame starting in period 1. Let s_r denote the XPE of this subgame 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, both when G^2 is realized and when $G^{2'}$ is realized.

Let s_s 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.

Thus, within this two-period subgame, we can deliver to player 1 a payoff of at most 1 if G^2 ends up realized (namely, by playing s_r); we can also deliver to player 1 a payoff at most 0 if $G^{2'}$ is realized (namely, by s_s). However, no single XPE meets both these bounds at once: it would be necessary to play cs in period 1 and therefore to reward player 2 in period 2, but in G^2 this can be done only playing fv (with probability 1). Note also that this argument takes into account the availability of public randomization.

Thus, there is no optimal penal code for this subgame. To see the lack of a *universal* penal code, we further reason as follows. In the overall game, the (deterministic) realizable outcome with actions aq, cr, eu can be supported, with hx being chosen in period 2 if $G^{2\prime}$ 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_r as a punishment. The realizable 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 one punishment can support both of these outcomes in XPE. Such a punishment would need to deliver an (expected) payoff to player 1 of ≤ 1 across periods 1–2 if G^2 is realized and a payoff ≤ 0 if $G^{2\prime}$ is realized, and we observed already that this is impossible.

This shows the lack of a universal penal code with two long-run players. Note that it also shows that Theorem 5.3 fails, since the full outcome aq, cr, eu, gw cannot be supported even though its constituent realizable outcomes can.

Also, a minor variant of this example gives a game 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 $G^{0\prime}$ that is a copy of G^0, and now consider the full outcome that specifies aq, cr, eu, hx if G^0 is realized, and aq, cr, fv, gw if $G^{0\prime}$ is realized. This contrasts with stochastic games, where universal penal codes for full outcomes do exist (Kitti, 2016).

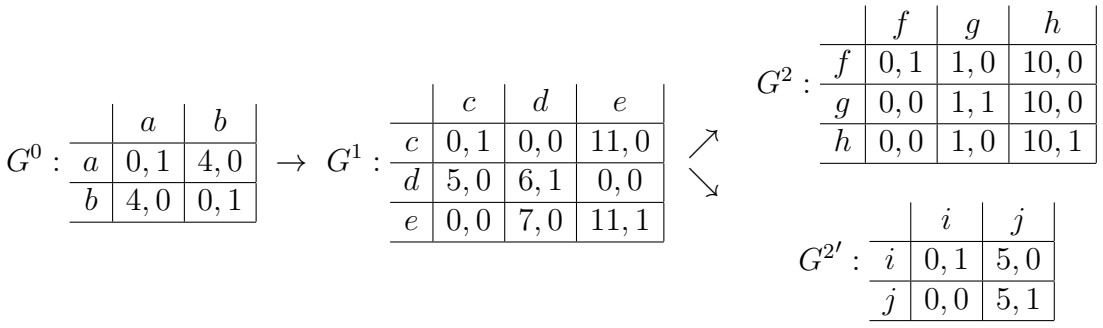
△

6.2 No public randomization

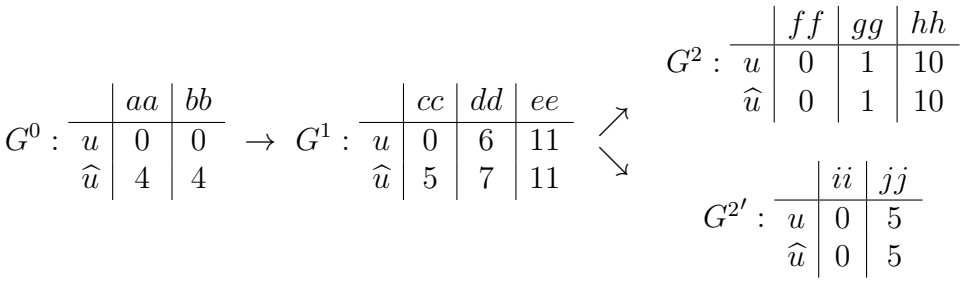
Let us now return to assuming that only player 1 is a long-run player, but remove public randomization. We again give a three-period example.

Example 6.2. Again, one possible stage game in each period $t = 0, 1$, and two in period 2. These games are presented in Figure 3. Part (a) presents the stage games in the usual matrix form, while part (b) extracts the informatino that is relevant for us; it shows the action profiles in $A^*(G)$ for each stage game and the values of $u, \widehat{u}$ for each.

For the subgame beginning in period 1, let s_c denote the XPE profile 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 $G^{2\prime}$ is realized.



(a)



(b)

Figure 3: Example without public randomization. No universal penal code exists.

Let s_d denote the subgame 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.

Thus, in this two-period subgame, one XPE delivers the low payoff of 7 if G^2 arrives in the last period, and another delivers 5 if $G^{2'}$ arrives. But no one XPE gives a payoff ≤ 7 in both environments: playing cc in G^1 would require following up with hh if G^2 arrives to deter deviation (thus giving total payoff 10 across the two periods), while dd in G^1 requires following up with jj in $G^{2'}$ (for total payoff 11).

As in the previous example, we can proceed to infer the nonexistence of a universal penal code as well. In the three-period game, the realizable outcome (deterministic, of course) with actions aa, ee, ff can be supported in XPE. Namely, specify aa, ee, ff, jj as on-path actions, and if player 1 deviates in period 0, switch to s_d as punishment. (Deviation in period 1 only can be ignored, since it brings no within-period gain.) The realizable 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, from the previous

paragraph, no one punishment is able to support both of these outcomes at once.

As in Example 6.1, this further implies that Theorem 5.3 fails in this environment, and it can also be adapted to give a game with a single outcome that can be supported but only using different off-path punishments at different histories.

△

7 Summary

Repeated games provide a widely used modeling framework for studying self-enforcing dynamic incentives. But the standard model assumes a complete specification of the payoff environment. This paper has explored the framework of uncertain repeated games, where, at each moment when players are called on to act, there is only partial information about the nature of rewards and punishments that may be available in future periods. In particular, this framework is able to express the idea that a gain from deviation in one period is outweighed by losses in future periods, without being able to map a given strategy profile to a precise numeric total payoff—something that plays a central role in the standard toolkit. We have adopted the solution concept of ex-post perfect equilibrium, which captures the extent to which the partial specification of the environment is sufficient to design incentives.

Under two significant assumptions—a single long-run player interacting with a series of short-run players, and availability of public randomization—we can carry out a comprehensive analysis of any given uncertain repeated game, in the sense that we can fully describe which outcomes can and cannot be supported in such an equilibrium. Our analysis adapts the recursive technique from APS, applied to the space of reward-punishment gaps rather than the space of total payoffs. Moreover, the analysis provides an explanation of why each outcome is or isn't supportable: for those outcomes that can be supported, we show constructively how to do it (by finding an optimal penal code); and for those that cannot, we show why not (by iteratively ruling out certain actions due to inadequate scope for punishment, and showing that the remaining actions are insufficient to deter deviation in some environment). In particular, one of the major qualitative lessons from the standard theory of repeated games—that the outcomes that can be supported are precisely those for which the deviation gain never exceeds the scope for punishment—holds up in this setting. We also saw, however, that when we allow multiple long-run players, or when we drop public randomization, neither this lesson nor the analytical toolkit readily extends—in particular, optimal penal codes may no longer exist.

This paper has adopted the uncertain repeated games setting, as a parsimonious way of departing from the standard framework. However, it should be clear that both the solution concept and the key techniques generalize in various ways: for example, as in stochastic games, we could allow for a set of possible transition probabilities via which the period-$(t+1)$ stage game may be drawn as a function of the period-t stage game (and even of the actions played as well). In addition, the recursion based on the reward-punishment gap can apply beyond the long-run/short-run setting: what is needed is that the object of the recursion is a one-dimensional interval. Indeed, as Krasikov and Lamba (2023) show, the approach here can be merged with Cronshaw and Luenberger (1994) to study symmetric uncertain repeated games (with multiple long-run players) and characterize the outcomes of strongly symmetric XPE's.

The standard model of repeated games has proven extremely valuable for understanding the possibilities and limitations of providing incentives through repeated interactions. Central elements in the analysis of this model include optimal penal codes and their use to characterize outcomes in payoff space. It seems fitting to ask to what extent these elements are intrinsically connected to the economic concept of dynamic incentives, as opposed to being convenient features of a particular mathematical model. For this distinction to be meaningful, it is necessary to argue that other models are possible. This paper has undertaken a step in this direction. Perhaps some still more general modeling framework in the future will provide further understanding.

A Omitted proofs

Proof of Lemma 4.2. Because $B(w;G)$ and $B(w)$ are increasing, right-continuity is equivalent to upper semi-continuity for both of them.

Because $A^*(G,w)$ is upper hemi-continuous in w, and u and $\widehat{u}$ are continuous functions, the first term in the definition (4.1) of $B(w;G)$ is upper semi-continuous in w. And the second term is clearly continuous. Upper semi-continuity of $B(w;G)$ follows. Then, $B(w)$, being a pointwise infimum of upper semi-continuous functions, is also upper semi-continuous.

□

Proof of Lemma 4.3. First, note that $w_k > w^*$ by an easy induction. 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_∞. 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 Lemma 4.4. 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 4.3 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.

We prove the statement by induction on k. The base case $k = 0$ is immediate, given $2M < w_0$. 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

$$(1-\delta)u(a^0) + \delta U(s|E, h^0, a^0) \geq (1-\delta)\widehat{u}(a^0) + \delta U(s|E, h^0, (a_1', a_{-1}^0))$$

or, rearranging,

$$(1-\delta)(\widehat{u}(a^0) - u(a^0)) \leq \delta(U(s|E, h^0, a^0) - U(s|E, h^0, (a_1', a_{-1}^0))).$$

The right side is δ times the difference of two SPE payoffs in the continuation environment E^-, and so is less than δw_{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 $\overline{s}, \underline{s}$ 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 ω^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 $\underline{s}$ is at least the payoff from deviating to the myopically action a'_1 in date 0, which is

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

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

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

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

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

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

□

Proof of Proposition 4.7. We prove the three assertions in succession.

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

$$\begin{aligned} U(s[w]|E) \;=\; & \lambda(G)w \times \left((1-\delta)u(\overline{a}(G)) + \delta U(s[w^*]|E^-)\right) + \\ & (1-\lambda(G)w) \times \left((1-\delta)u(\underline{a}(G)) + \delta U\left(s\left[\frac{1-\delta}{\delta}(\widehat{u}(\underline{a}(G)) - u(\underline{a}(G)))\right]\middle| E^-\right)\right). \end{aligned} \tag{A.1}$$

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

$$\begin{aligned} \widetilde{U}(w|E) \;=\; & \lambda(G)w \times \left((1-\delta)u(\overline{a}(G)) + \delta \widetilde{U}(w^*|E^-)\right) + \\ & (1-\lambda(G)w) \times \left((1-\delta)u(\underline{a}(G)) + \delta \widetilde{U}\left(\frac{1-\delta}{\delta}(\widehat{u}(\underline{a}(G)) - u(\underline{a}(G)))\middle| E^-\right)\right). \end{aligned} \tag{A.2}$$

To see this, expand both the $\widetilde{U}$ terms on the right-hand side of (A.2) and obtain (after slightly simplifying the second line)

$$\begin{aligned}\lambda(G)w \times \left((1-\delta)u(\overline{a}(G)) + \delta\underline{U}(E^-) + \delta w^*\right) + \\ (1-\lambda(G)w) \times \left((1-\delta)\widehat{u}(\underline{a}(G)) + \delta\underline{U}(E^-)\right).\end{aligned}$$

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

$$((1-\delta)\widehat{u}(\underline{a}(G)) + \delta\underline{U}(E^-)) + \lambda(G)w \times ((1-\delta)(u(\overline{a}(G)) - \widehat{u}(\underline{a}(G))) + \delta w^*).$$

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

Now write $\Delta(w|E) = U(s[w]|E) - \widetilde{U}(w|E)$. Subtracting (A.2) from (A.1) gives

$$\Delta(w|E) = \lambda(G)w \times \delta\Delta(w^*|E^-) + (1-\lambda(G)w) \times \delta\Delta\left(\frac{1-\delta}{\delta}(\widehat{u}(\underline{a}(G)) - u(\underline{a}(G)))\middle| E^-\right).$$

Put $C = \sup_{w,E} |\Delta(w|E)|$; this is finite since both U and $\widetilde{U}$ are bounded. Using C to bound each of the $\Delta(\cdots)$ 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, $U(s[w]|E) = \widetilde{U}(w|E)$ 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 $\overline{a}(G^0)$. If player 1 conforms, the state next period is w^*, so the continuation payoff will be $\underline{U}(E^-) + w^*$ by part 1, and therefore the total payoff is

  $$(1-\delta)u(\overline{a}(G^0)) + \delta\left(\underline{U}(E^-) + w^*\right).$$

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

is

$$(1-\delta)\widehat{u}(\overline{a}(G^0)) + \delta\left(\underline{U}(E^-) + w^* - \frac{1-\delta}{\delta}\left(\widehat{u}(\overline{a}(G^0)) - u(\overline{a}(G^0))\right)\right)$$
$$= (1-\delta)u(\overline{a}(G^0)) + \delta\left(\underline{U}(E^-) + w^*\right).$$

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

$$(1-\delta)u(\underline{a}(G^0)) + \delta\left(\underline{U}(E^-) + \frac{1-\delta}{\delta}\left(\widehat{u}(\underline{a}(G^0)) - u(\underline{a}(G^0))\right)\right)$$
$$= (1-\delta)\widehat{u}(\underline{a}(G^0)) + \delta\underline{U}(E^-)$$

and if player 1 deviates is

$$(1-\delta)\widehat{u}(\underline{a}(G^0)) + \delta\underline{U}(E^-).$$

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

3: We have just shown that in every environment and 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 $\widehat{u}(a) = u(a)$, 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 short-run 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.

□

Proof of Proposition 5.2. As noted in the text, the conditions (5.1)–(5.2) as stated are equivalent to requiring $a^t \in A^*(G^t)$ for all t and (5.2) for all $\underline{t} \leq \overline{t}$. So it suffices to check that, in the deterministic case, the latter is equivalent to $d^t(z) \leq \frac{\delta}{1-\delta}w^*$ for all t. Rewrite (5.2) as

$$\frac{1}{\delta^{\overline{t}-\underline{t}}}\left(\widehat{u}(a^{\underline{t}}) - u(a^{\underline{t}})\right) + \sum_{t=\underline{t}+1}^{\overline{t}} \delta^{t-\overline{t}}\left(\widehat{u}(\underline{a}(G^t)) - u(a^t)\right) \leq \frac{\delta}{1-\delta}w^*. \tag{A.3}$$

(We have removed the expectation operator since a^t is no longer random.) Denoting the left-hand side of (A.3) by $d^{\underline{t},\bar{t}}(z)$, requiring (5.2) for all pairs of dates $\underline{t} \leq \bar{t}$ is then equivalent to $\max_{\underline{t}\in\{0,\ldots,\bar{t}\}} d^{\underline{t},\bar{t}}(z) \leq \frac{\delta}{1-\delta} w^*$ for all $\bar{t}$. But it is easy to see by induction that $d^{\bar{t}}(z) = \max_{\underline{t}\in\{0,\ldots,\bar{t}\}} d^{\underline{t},\bar{t}}(z)$.

□

Proof of Theorem 5.3. Necessity is immediate. For sufficiency, we need to argue that it suffices for each realizable outcome to satisfy (5.1)–(5.2). This follows from sufficiency of the weaker conditions in Theorem 5.4, so we defer to that proof.

□

Proof of Theorem 5.4. Again, we already have necessity, so we focus on sufficiency. Adapting the proof of Theorem 5.1, we construct a strategy profile s as follows: at any history where actions have not yet deviated from z, play as specified by z; 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 (and random signals), 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 $a^{t'}$, for each $t' \geq t$, be the actions specified by z in this environment (which may depend on the already-realized signals $\omega^{0,\ldots,t}$, as well as the random future signals). If player 1 conforms to s, then the payoff starting at h^t from conforming is

$$(1-\delta)\left(\sum_{t'=t}^{\infty} \delta^{t'-t}\mathbb{E}^t[u(a^{t'})]\right). \tag{A.4}$$

If player 1 deviates from s, then subsequent play transitions to $s[0]$ and so the payoff from $t+1$ onward is given by $\underline{U}(E^{-(t+1)})$. Therefore, the payoff from deviating optimally, as measured from h^t, is

$$(1-\delta)\widehat{u}(a^t) + \delta\underline{U}(E^{-(t+1)}) = (1-\delta)\left(\widehat{u}(a^t) + \sum_{t'=t+1}^{\infty} \delta^{t'-t}\widehat{u}(\underline{a}(G^{t'}))\right). \tag{A.5}$$

Rearranging (5.5) tells us exactly that (A.4) is greater than or equal to (A.5). Hence, deviating is never profitable, in any environment.

□

References

Abreu, Dilip (1988) "On the theory of infinitely repeated games with discounting," *Econometrica*, 383–396.

Abreu, Dilip, David Pearce, and Ennio Stacchetti (1990) "Toward a theory of discounted repeated games with imperfect monitoring," *Econometrica*, 1041–1063.

Chari, V V and Patrick J Kehoe (1990) "Sustainable plans," *Journal of Political Economy*, **98** (4), 783–802.

Cronshaw, Mark B and David G Luenberger (1994) "Strongly symmetric subgame perfect equilibria in infinitely repeated games with perfect monitoring and discounting," *Games and Economic Behavior*, **6** (2), 220–237.

Ely, Jeffrey C, Johannes Hörner, and Wojciech Olszewski (2005) "Belief-free equilibria in repeated games," *Econometrica*, **73** (2), 377–415.

Fudenberg, Drew, David M Kreps, and Eric S Maskin (1990) "Repeated games with long-run and short-run players," *Review of Economic Studies*, **57** (4), 555–573.

Fudenberg, Drew and David K Levine (1989) "Reputation and equilibrium selection in games with a patient player," *Econometrica*, **57** (4), 759–778.

Fudenberg, Drew and Yuichi Yamamoto (2010) "Repeated games where the payoffs and monitoring structure are unknown," *Econometrica*, **78** (5), 1673–1710.

Hörner, Johannes and Stefano Lovo (2009) "Belief-free equilibria in games with incomplete information," *Econometrica*, **77** (2), 453–487.

Kitti, Mitri (2016) "Subgame perfect equilibria in discounted stochastic games," *Journal of Mathematical Analysis and Applications*, **435** (1), 253–266.

Kostadinov, Rumen (2023) "Worst-case Regret in Ambiguous Dynamic Games," Unpublished manuscript, McMaster University.

Krasikov, Ilia and Rohit Lamba (2023) "Uncertain Repeated Games," Unpublished manuscript, Pennsylvania State University.

Maccheroni, Fabio, Massimo Marinacci, and Aldo Rustichini (2006) "Dynamic variational preferences," *Journal of Economic Theory*, **128** (1), 4–44.

Mailath, George J and Larry Samuelson (2006) *Repeated games and reputations: long-run relationships*: Oxford University Press.

Rotemberg, Julio J and Garth Saloner (1986) "A supergame-theoretic model of price wars during booms," *American Economic Review*, **76** (3), 390–407.

Stokey, Nancy L (1991) "Credible public policy," *Journal of Economic Dynamics and Control*, **15** (4), 627–656.