Discounted Stochastic Games with Voluntary Transfers

Susanne Goldlücke* and Sebastian Kranz†

This Version, February 2015 — work in progress —

Abstract

This paper studies discounted stochastic games with perfect or imperfect public monitoring and the opportunity to conduct voluntary monetary transfers. We show that for all discount factors every public perfect equilibrium payoff can be implemented with a simple class of equilibria that have a stationary structure on the equilibrium path and optimal penal codes with a stick and carrot structure. We develop algorithms that exactly compute or approximate the set of equilibrium payoffs and find simple equilibria that implement these payoffs.

JEL-Codes: C73, C61, C63

Keywords: stochastic games, monetary transfers, computation, imperfect public monitoring, public perfect equilibria

1 Introduction

Discounted stochastic games are a natural generalization of infinitely repeated games that provide a very flexible framework to study relationships in a wide variety of applications. Players interact in infinitely many periods and discount

*Department of Economics, University of Konstanz, susanne.goldluecke@uni-konstanz.de.

†Department of Mathematics and Economics, Ulm University, sebastian.kranz@uni-ulm.de.

I would like to thank the German Research Foundation (DFG) for financial support through SFB-TR 15 at the University of Bonn and an individual research grant. Part of the work has been conducted while I was visiting the Cowles Foundation in Yale which provided a very stimulating research environment.

We would like to thank Dirk Bergemann, An Chen, Mehmet Ekmekci, Paul Heidhues, Johannes Hörner, Jon Levin, David Miller, Larry Samuelson, Philipp Strack, Juuso Välimäki, Joel Watson and seminar participants at Arizona State University, UC San Diego and Yale for very helpful discussions.

future payoffs with a common discount factor. Payoffs and available actions in a period depend on a state that can change between periods in a deterministic or stochastic manner. The probability distribution of the next period's state only depends on the state and chosen actions in the current period. For example, in a long-term principal-agent relationship, a state may describe the amount of relationship specific capital or the current outside options of each party. In a dynamic oligopoly model, a state may describe the number of active firms, the production capacity of each firm, or demand and cost shocks that can be persistent over time.

In many relationships of economic interest, parties cannot only perform actions but also have the option to transfer money to each other or to a third party. Repeated games with monetary transfers and risk-neutral players have been widely studied in the literature. Examples include studies of employment relations by Levin (2002, 2003) and Malcomson and MacLeod (1989), partnerships and team production by Doornik (2006) and Rayo (2007), prisoner dilemma games by Fong and Surti (2009), international trade agreements by Klimenko, Ramey and Watson (2008) and cartels by Harrington and Skrzypacz (2007, 2011).[1] Levin (2003) shows for repeated principal agent games with transfers that one can restrict attention to a simple class of stationary equilibria in order to implement every public perfect equilibrium payoff. Goldlücke and Kranz (2012) derive a similar characterization for general repeated games with transfers.

This paper extends these results to stochastic games with voluntary transfers and imperfect monitoring of actions. For any given discount factor $\delta \in [0, 1)$, all public perfect equilibrium (PPE) payoffs can be implemented with a simple class of equilibria. Based on that result, algorithms are developed that allow to approximate or to exactly compute the set of PPE payoffs.

A simple equilibrium is described by an equilibrium phase and for each player a punishment phase. In the equilibrium phase the chosen action profile only depends on the current state, like in a Markov Perfect equilibrium. Voluntary transfers after the new state has been realized are used to smooth incentive constraints across players. Play moves to a player's punishment phase whenever that player refuses to make a required transfer. Punishments have a simple stick-and-carrot structure: one punishment action profile per player and state is defined. After the punished profile has been played and subsequent transfers are conducted, play moves back to the equilibrium phase. We show that for every discount factor there is an *optimal* simple equilibrium that implements in every state the highest joint continuation payoffs (i.e. the sum of payoffs across all players) in the equilibrium phase and in the punishment phases the lowest continuation payoff for the punished player that can be achieved by any simple equilibrium. By varying up-front payments in the very first period, one can implement every PPE payoff with such an optimal simple equilibrium.

[1] Miller and Watson (2011), Gjertsen et. al (2010), Fong and Surti (2009), Kranz and Ohlendorf (2009) and Baliga and Evans (2000) study renegotiation-proof equilibria in repeated games with transfers.

Based on that result, we develop algorithms for games with finite action spaces that allow to approximate or to exactly compute the set of pure strategy PPE payoffs and yield (optimal) simple equilibria to implement any payoff. To compute inner and outer approximations of the PPE payoff set, one can use decomposition methods, in which attention can be restricted to state-wise maximal joint continuation payoffs and minimal continuation payoffs for each player. Sufficiently fine approximations allow to reduce for each state the set of action profiles that can possibly be part of an optimal simple equilibrium. If these sets can be sufficiently reduced, a brute force method that solves a linear optimization for every combination of remaining action profiles allows to find an optimal simple equilibrium and to exactly compute the set of PPE payoffs.

If actions can be perfectly monitored, the characterization of optimal simple equilibria substantially simplifies. Decomposition steps just require to evaluate a simple formula for each state and action profile, while under imperfect monitoring a linear optimization problem has to be solved. Furthermore, we develop an alternative policy elimination algorithm that exactly computes the set of pure strategy subgame perfect equilibrium payoffs by repeatedly solving a single agent Markov decision problem for the equilibrium phase and a nested variation of a Markov decision problem for the punishment phases.

In general, the flexibility of discounted stochastic games comes at the price that solving them entails considerably more difficulties than solving infinitely repeated games. Finding just a single equilibrium of a stochastic game can be challenging, while an infinite repetition of the stage game Nash equilibrium is always an equilibrium in a repeated game. Complexities increase when one wants to determine the set of all equilibrium payoffs. For stochastic games without transfers and in the limit case as the discount factor converges towards 1, folk theorems have been established by Dutta (1995) for perfect monitoring and Fudenberg and Yamamoto (2010) and Hörner et. al. (2011) for imperfect monitoring of actions in irreducible stochastic games. For fixed discount factors Judd, Yeltekin and Conklin (2003) and Abreu and Sannikov (2011) have developed numerical methods, based on the seminal recursive techniques by Abreu, Pearce and Stacchetti (1990, henceforth APS), to approximate the equilibrium payoff sets for repeated games with public correlation and perfect monitoring. In principle, these methods can be extended to general stochastic games (see, e.g. Sleet and Yeltekin, 2003), but it is still an open question how tractable such extensions will be in terms of computational requirements, delivering guidance for closed-form solutions, and the ability to deal with imperfect monitoring. This paper shows that in stochastic games with monetary transfers, one can very effectively handle these issues.

Applied literature that studies stochastic games typically restricts attention to Markov perfect equilibria (MPE) in which actions only condition on the current state.[2] Voluntary transfers that do not change the state would then never be

[2]Examples include studies of learning-by-doing by Benkard (2004) and Besanko et. al. (2010), advertisement dynamics by Doraszelski and Markovich (2007), consumer learning by

conducted. Focusing on MPE has advantages, e.g. strategies have a simple structure and there exist quick algorithms to find a MPE. However, there are also drawbacks.

One issue is that the set of MPE payoffs can be very sensitive to the definition of the state space. For example, a repeated game has by definition just one state, so only an infinite repetition of the same stage game Nash equilibrium can be a MPE. Yet, if one specifies a state to be described by the action profile of the previous period (which may have some small influence on the current period's payoff function), also collusive grim-trigger strategies can be implemented as a MPE.

Another issue is that there are no effective algorithms to compute all MPE payoffs of stochastic game, even if one just considers pure strategies.[3] Existing algorithms, e.g. Pakes & McGuire (1994, 2001), are very effective in finding one MPE, but except for special games there is no guarantee that it is unique. Besanko et. al. (2010) illustrate the multiplicity problem and show how the homotopy method can be used to find multiple MPE. Still there is no guarantee, however, that all (pure) MPE are found.

For those reasons, effective methods to compute the set of all PPE payoffs and an implementation with a simple class of strategy profiles generally seem quite useful in order to complement the analysis of MPE.

While monetary transfers may not be feasible in all social interactions, the possibility of transfers is plausible in many problems of economic interest. Even for illegal collusion, transfer schemes are in line with the evidence from several actual cartel agreements. For example, the citric acid and lysine cartels required members that exceeded their sales quota in some period to purchase the product from their competitors in the next period; transfers were implemented via sales between firms. Harrington and Skrzypacz (2011) describe transfer schemes used by cartels in more detail and provide further examples.[4] Risk-neutrality is also often a sensible approximation, in particular if players are countries or firms or if payments of the stochastic games are small in comparison to expected lifetime income and individuals have access to well functioning financial markets. Even in contexts in which transfers or risk-neutrality may be considered strong

Ching (2010), capacity expansion by Besanko and Doraszelski (2004), or network externalities by Markovich and Moenius (2009).

[3]For a game with finite action spaces, one could always use a brute-force method that checks for every pure strategy Markov strategy profile whether it constitutes a MPE. Yet, the number of Markov strategy profiles increases very fast: is given by $\prod_{x \in X} |A(x)|$, where $|A(x)|$ is the number of strategy profiles in state x. This renders a brute force method practically infeasible except for very small stochastic games.

[4]Further examples of cartels with transfers schemes include the choline chloride, organic peroxides, sodium gluconate, sorbates, vitamins, and zinc phosphate cartels. Interesting detailed descriptions can also be obtained from older cases, in which cartel members more openly documented their collusive agreements. An example is the Supreme Court decision Addyston Pipe & Steel Co. v. U. S., 175 U.S. 211 (1899). It describes the details of a bid-rigging cartel in which a firm that won a contract had to make payment to the other cartel members.

assumptions, our results can be useful since the set of implementable PPE payoffs with transfers provides an upper bound on payoffs that can be implemented by equilibria without transfers or under risk-aversion.

The structure of this paper is as follows. Section 2 describes the model and defines simple strategy profiles. Section 3 first provides an intuitive overview of how transfers facilitate the analysis. It is then shown that every PPE can be implemented with simple equilibria. Section 4 describes algorithms that allow to approximate or exactly compute the set of pure strategy PPE payoffs. Section 5 shows how the results simplify for games with perfect monitoring and develops an alternative algorithm that exploits these simplifications. Section 6 illustrates with examples how numerical or analytical solutions can be obtained with the developed methods. All proofs are relegated to an Appendix.

2 Model and Simple Strategy Profiles

2.1 Model

We consider an n player stochastic game of the following form. There are infinitely many periods and future payoffs are discounted with a common discount factor $\delta \in [0, 1)$. There is a finite set of states X and $x_0 \in X$ denotes the initial state. A period is comprised of two stages: a transfer stage and an action stage. There is no discounting between stages.

In the transfer stage, every player simultaneously chooses a non-negative vector of transfers to all other players.[5] To have a compact strategy space, we assume that a player's transfers cannot exceed some finite upper bound. Yet, we assume that this upper bound is large enough to be never binding given the constraint that transfers must be voluntary. Players also have the option to transfer money to a non-involved third party, which has the same effect as burning money.[6] Transfers are perfectly monitored.

In the action stage, players simultaneously choose actions. In state $x \in X$, player i can choose a pure action a_i from a finite or compact action set $A_i(x)$. The set of pure action profiles in state x is denoted by $A(x) = A_1(x) \times ... \times A_n(x)$.

After actions have been conducted, a signal y from a finite signal space Y and a new state $x' \in X$ are drawn by nature and commonly observed by all players.

[5] To have a compact strategy space, we assume that a player's transfers cannot exceed an upper bound of $\frac{1}{1-\delta} \sum_{i=1}^{n} \left[\max_{x \in X, a \in A(x)} \pi_i(a, x) - \min_{x \in X, a \in A(x)} \pi_i(a, x)\right]$ where $\pi_i(a, x)$ are expected stage game payoffs defined below. That bound is large enough to be never binding given the incentive constraints of voluntary transfers.

[6] An extension to the case without money burning is possible if one allows for a public correlation device. Instead of burning money, players will coordinate with positive probability a continuation equilibrium that minimizes the sum of continuation payoffs. In a similar fashion as in Goldlücke and Kranz's (2012) analysis for repeated games, one can provide a characterization with an extended class of simple equilibria.

We denote by $\phi(y, x'|x, a)$ the probability that signal y and state x' are drawn; it depends only on the current state x and the chosen action profile a. Player i's stage game payoff is denoted by $\hat{\pi}_i(a_i, y, x)$ and depends on the signal y, player i's action a_i and the initial state x. We denote by $\pi_i(a, x)$ player i's expected stage game payoff in state x if action profile a is played. If the action space in state x is compact then stage game payoffs and the probability distribution of signals and new states shall be continuous in the action profile a.

We assume that players are risk-neutral and that payoffs are additively separable in the stage game payoff and money. This means that the expected payoff of player i in a period with state x, in which she makes a net transfer of p_i and action profile a has been played, is given by $\pi_i(a, x) - p_i$.

When referring to (continuation) payoffs of the stochastic game, we mean expected average discounted continuation payoffs, i.e. the expected sum of continuation payoffs multiplied by $(1-\delta)$. A payoff function $u : X \to \mathbb{R}^n$ maps every state into a vector of payoffs for each player. We generally use upper case letters to denote joint payoffs of all players, e.g.

$$U = \sum_{i=1}^{n} u_i.$$

We study the payoff sets of pure strategy equilibria and for finite action spaces we also consider the case that players can mix over actions. If equilibria with mixed actions are considered, $\mathcal{A}(x)$ shall denote the set of mixed action profiles at the action stage in state x otherwise $\mathcal{A}(x) = A(x)$ shall denote the set of pure action profiles. For a mixed action profile $\alpha \in \mathcal{A}(x)$, we denote by $\pi_i(\alpha, x)$ player i's expected stage game payoff taking expectations over mixing probabilities and signal realizations.

A public history of the stochastic game describes the sequence of all states, public signals and monetary transfers that have occurred before a given point in time. A public strategy σ_i of player i in the stochastic game maps every public history that ends before the action stage into a possibly mixed action $\alpha_i \in \mathcal{A}_i(x)$, and every public history that ends before a payment stage into a vector of monetary transfers. A public perfect equilibrium is a profile of public strategies that constitutes mutual best replies after every history. We restrict attention to public perfect equilibria.

A vector $\boldsymbol{\alpha}$ that assigns an action profile $\boldsymbol{\alpha}(x) \in \mathcal{A}(x)$ to every state $x \in X$ is called a policy and $\boldsymbol{\mathcal{A}} = \times_{x \in X} \mathcal{A}(x)$ denotes the set of all policies. For briefness sake, we abbreviate an action profile $\boldsymbol{\alpha}(x)$ by the policy $\boldsymbol{\alpha}$ if it is clear which action profile is selected, e.g. $\pi(\boldsymbol{\alpha}, x) \equiv \pi(\boldsymbol{\alpha}(x), x)$.

2.2 Simple strategy profiles

We now describe the structure of simple strategy profiles. In a simple strategy profile, it will never be the case that a player at the same time makes and receives

transfers. We therefore describe transfers by the net payments that players make.[7]

A simple strategy profile is characterized by $n+2$ phases. Play starts in the up-front transfer phase, in which players are required to make up-front transfers described by net payments p^0. Afterwards, play can be in one of $n+1$ phases, which we index by $k \in \mathcal{K} = \{e, 1, 2, ..., n\}$. We call the phase $k = e$ the equilibrium phase and $k = i \in \{1, ..., n\}$ the punishment phase of player i.

A simple strategy profile specifies for each phase $k \in \mathcal{K}$ and state x an action profile $\boldsymbol{\alpha}^k(x) \in \mathcal{A}(x)$. We refer to $\boldsymbol{\alpha}^e$ as the equilibrium phase policy and to $\boldsymbol{\alpha}^i$ as the punishment policy for player i. From period 2 onwards, required net transfers are described by net payments $p^k(y, x', x)$ that depend on the current phase k, the realized signal y, the realized state x' and the previous state x. The vectors of all policies $(\boldsymbol{\alpha}^k)_{k\in\mathcal{K}}$ and all payment functions $(p^k(.))_{k\in K}$ are called action plan and payment plan, respectively.

The transitions between phases are simple. If no player unilaterally deviates from a required transfer, play transits to the equilibrium phase: $k = e$. If player i unilaterally deviates from a required transfer, play transits to the punishment phase of player i, i.e. $k = i$. In all other situations the phase does not change. A simple equilibrium is a simple strategy profile that constitutes a public perfect equilibrium of the stochastic game.

3 Characterization with simple equilibria

This section first provides some intuition and then derives the main result that all PPE payoffs can be implemented with simple equilibria. It is helpful to think of three ways in which monetary transfers simplify the analysis:

1. Upfront transfers in the very first period allow flexible distribution of the joint equilibrium payoffs.

2. Transfers in later periods allow to balance incentive constraints between players.

[7] Any vector of net payments p can be mapped into a matrix of gross transfers $\tilde{p}_{ij}$ from i to j as follows. Denote by $I_P = \{i|p_i > 0\}$ the set of net payers and by $I_R = \{i|p_i \leq 0\} \cup \{0\}$ the set of net receivers including the sink for burned money indexed by 0. For any receiver $j \in I_R$, we denote by

$$s_j = \frac{|p_j|}{\sum_{j\in I_R} |p_j|}$$

the share she receives from the total amount that is transferred or burned and assume that each net payer distributes her gross transfers according to these proportions

$$\tilde{p}_{ij} = \begin{cases} s_j p_i & \text{if } i \in I_P \text{ and } j \in I_R \\ 0 & \text{otherwise.} \end{cases}$$

Note that the sum of net payments is always non-negative, since players can only perform non-negative gross transfers.

3. The payment of fines allows to settle punishments within one period.

3.1 Distributing with upfront transfers

Consider Figure 1. The shaded area shall illustrate for a two player stochastic game with fixed discount factor all payoffs of PPE that do not use upfront transfers. The set is assumed to be compact. The point $\bar{u}$ is the equilibrium payoff with the highest sum of payoffs for both players.

If one could impose enforceable upfront transfers without any liquidity constraints, the set of Pareto-optimal payoffs would be simply given by a line with slope -1 through this point. If upfront transfers must be incentive compatible, their maximum size is bounded by the harshest punishment that can be credibly imposed on a player that deviates from a required transfers. The harshest credible punishment for player $i = 1, 2$ is given by the continuation equilibrium after the first transfer stage that has the lowest payoff for player i. The idea to punish any deviation with the worst continuation equilibrium for the deviator is the crux of Abreu's (1988) optimal penal codes.

Points w^1 and w^2 in Figure 1 illustrate these worst equilibria for each player and $\bar{v}$ is the point where each coordinate $i = 1, 2$ describes the worst payoff of player i. The Pareto frontier of subgame perfect equilibrium payoffs with voluntary upfront transfers is given by the shown line segment through point $\bar{u}$ with slope -1 that is bounded by the lowest equilibrium payoff $\bar{v}_1$ of player 1 at the left and the lowest equilibrium payoff $\bar{v}_2$ of player 2 at the bottom. If we allow for money burning in upfront transfers, any point in the depicted triangle can be implemented in an incentive compatible way. That intuition naturally extends to n player games.

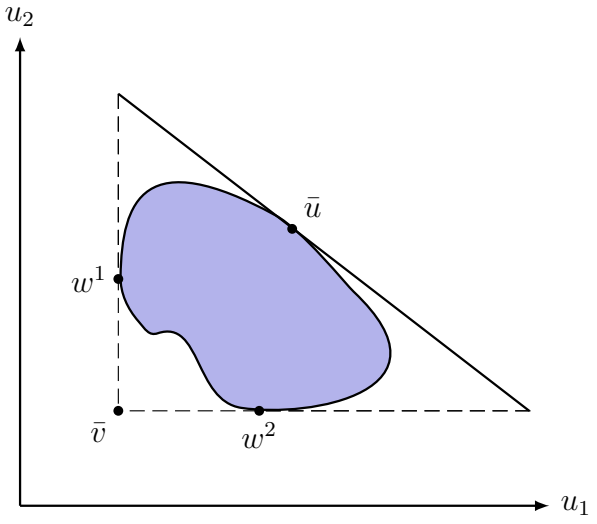

Figure 1: Distributing with upfront transfers

Proposition 1. *Assume that across all PPE that do not use transfers in the first period there exists a highest joint payoff* $\bar{U}$ *and for every player* $i = 1, ..., n$ *a lowest payoff* $\bar{v}_i$ *. Then the set of PPE payoffs with transfers in the first period is the simplex*

$$\{u \in \mathbb{R}^n | \sum_{i=1}^{n} u_i \leq \bar{U} \text{ and } u_i \geq \bar{v}_i\}.$$

That highest joint payoffs $\bar{U}$ and lowest payoffs $\bar{v}_i$ always exist is formally shown in Theorem 1 and not very surprising given the compactness result for the payoff sets of repeated games by APS. The set of PPE is thus defined by just $n+1$ real numbers: the highest joint PPE payoff $\bar{U}$ and the lowest PPE payoffs $\bar{v}_i$ for every player $i = 1, ..., n$.

3.2 Balancing incentive constraints

We now illustrate how transfers in later periods can be used to balance incentive constraints between players. Consider an infinitely repeated asymmetric prisoner's dilemma game described by the following payoff matrix:

	C	D
C	4,2	-3,6
D	5,-1	0,1

The goal shall be to implement mutual cooperation (C, C) in every period on the equilibrium path. Since the stage game Nash equilibrium yields the min-max payoff for both players, grim trigger punishments constitute optimal penal codes: any deviation is punished by playing forever the stage game Nash equilibrium (D, D).

No transfers First consider the case that no transfers are conducted. Given grim-trigger punishments, player 1 and 2 have no incentive to deviate from cooperation on the equilibrium path whenever the following conditions are satisfied:

$$\begin{aligned} &\text{Player 1: } 4 \geq (1-\delta)5 && \Leftrightarrow \delta \geq 0.2, \\ &\text{Player 2: } 2 \geq (1-\delta)6 \quad +\delta && \Leftrightarrow \delta \geq 0.8. \end{aligned}$$

The condition is tighter for player 2 than for player 1 for three reasons:

i) player 2 gets a lower payoff on the equilibrium path (2 vs 4),

ii) player 2 gains more in the period of defection (6 vs 5),

iii) player 2 is better off in each period of the punishment (1 vs 0).

Given such asymmetries, it is not necessarily optimal to repeat the same action profile in every period. For example, if the discount factor is $\delta = 0.7$, it is not possible to implement mutual cooperation in every period, but one can show that there is a SPE with non-stationary equilibrium path in which in every fourth period (C, D) is played instead of (C, C). Such a strategy profile relaxes the tight incentive constraint of player 2, by giving her a higher equilibrium path payoff. The incentive constraint for player 1 is tightened, but there is still sufficiently much slack left.

Note that even if players have access to a public correlation device, stationary equilibrium paths will not always be optimal.[8]

With transfers Assume now that (C, C) is played in every period and from period 2 onwards player 1 transfers an amount of $\frac{1.5}{\delta}$ to player 2 in each period on the equilibrium path. Using the one shot deviation property, it suffices to check that no player has an incentive for a one shot deviation from the actions or the transfers. Player 1 has no incentive to deviate from the transfers on the equilibrium path if and only if[9]

$$(1-\delta) * 1.5 \leq \delta * (4 - 1.5) \Leftrightarrow \delta \geq 0.375$$

and there is no profitable one shot deviation from the cooperative actions if and only if

$$\begin{aligned} &\text{Player 1: } 4 - 1.5 \geq (1-\delta)5 && \Leftrightarrow \delta \geq 0.5, \\ &\text{Player 2: } 2 + 1.5 \geq (1-\delta)6 \quad +\delta && \Leftrightarrow \delta \geq 0.5. \end{aligned}$$

The incentive constraints between the players are now perfectly balanced. Indeed, if we sum both players' incentive constraints

$$\text{Joint: } 4 + 2 \geq (1-\delta)(5+6) + \delta(0+1) \Leftrightarrow \delta \geq 0.5.$$

[8]For an example, consider the following stage game:

	A	B
A	0,0	-1,3
B	3,-1	0,0

Using a public correlation device to mix with equal probability between the profiles (A, B) and (B, A) on the equilibrium path and punishing deviations with infinite repetition of the stage game Nash equilibrium (B, B) constitutes a SPE whenever $\delta \geq \frac{1}{2}$. One can easily show that no other action profile with a stationary equilibrium path can sustain positive expected payoffs for any discount factor below $\frac{1}{2}$. Yet, a non-stationary equilibrium path $\{(A, B), (B, A), (A, B), ...\}$ that deterministically alternates between (A, B) and (B, A) can be implemented for every $\delta \geq \frac{1}{3}$. The reason is that when the profile (A, B) shall be played, only player 1 has an incentive to deviate. It is thus beneficial to give her a higher continuation payoff than player 2 and the reverse holds true if (B, A) shall be played. Unlike the stationary path, the non-stationary path has the feature that the player who currently has higher incentives to deviate gets a higher continuation payoff. Applying the results below, one can moreover establish that for $\delta < \frac{1}{3}$, one cannot implement any joint payoff above 0, even if one would allow for monetary transfers.

[9]To derive the condition, it is useful to think of transfers taking place at the end of the current period but discount them by δ. Indeed, one could introduce an additional transfer stage at the end of period (assuming the new state would be already known in that stage) and show that the set of PPE payoffs would not change.

we find the same critical discount factor as for the individual constraints. Intuitively, our formal results below show that in general stochastic games incentive constraints can always be perfectly balanced. This result is crucial for being able to restrict attention to simple equilibria and also facilitates computation of optimal equilibria within the class of simple equilibria.

3.3 Intuition for fines and stick-and-carrot punishments

If transfers are not possible, optimally deterring a player from deviations can become a very complicated problem. Basically, if players observe a deviation or an imperfect signal that is very likely under a profitable deviation, they have to coordinate on future actions that yield a sufficiently low payoff for the deviator. The punishments must themselves be stable against deviations and have to take into account how states can change on the desired path of play or after any deviation. Under imperfect monitoring, suspicious signals can also arise on the equilibrium path, which means "punishments" in Pareto optimal equilibria must entail as low efficiency losses as possible.

The benefits of transfers for simplifying optimal punishments are easiest seen for the case of pure strategy equilibria under perfect monitoring. Instead of conducting harmful punishment actions, one can always give the deviator the possibility to pay a fine that is as costly as if the punishment actions were conducted. If the fine is paid, one can move back to efficient equilibrium path. Punishment actions must only be conducted if a deviator fails to pay a fine. After one period of punishment actions, one can again give the punished player the chance to move back to efficient equilibrium path play if she pays a fine that will be as costly as the remaining punishment. This is the key intuition for why optimal penal codes can be characterized with stick-and-carrot type punishments with a single punishment action profile per player and state.[10]

If monitoring is imperfect or mixed strategies are used, deviations from prescribed actions may not be perfectly detected so that there is no clear notion of a fine. Still one can impose higher payments under signals that are relatively more likely under profitable deviations than on the equilibrium path.

There can be signals, like a project-failure in a team production setting, that indicate that some player has deviated but are not informative about which player deviated. In such cases it can be necessary to punish with a jointly inefficient continuation equilibrium. In our framework, such joint inefficiencies can be implemented via money burning.[11]

[10] That transfers can balance incentive constraints among several punishing players is also relevant for the result that stick-and-carrot punishments always suffice.

[11] Alternatively, if players would have a public correlation device, one could coordinate with some probability to a continuation equilibrium with low joint continuation payoffs in a similar fashion as Goldlücke and Kranz (2012) describe for repeated games.

3.4 Complications compared to repeated games with transfers

The overview above correctly suggests that many insights from repeated games with transfers carry over to the analysis of stochastic games with transfers. Nevertheless, substantial complications are introduced by allowing for multiple states.

In a repeated game, the current action profile does not affect the set of continuation payoffs. This means that the harshest punishment that can in principle be imposed on a deviating player is independent of the form of a deviation. The analysis of Goldlücke and Kranz (2012) heavily exploits this fact. It allows to compress all the relevant information of the continuation payoff set, which is crucial to determine whether and how effective an action profile can be used, into a single number that is given by the difference of joint equilibrium phase payoffs and the sum of all players' punishment payoffs multiplied by a factor that adjusts for discounting. One can formulate the relevant incentive constraints for simple equilibria by separating a joint measure of all player's current incentives to deviate from that number characterizing the continuation payoff set. This result is extremely useful for the formal characterization and allows to develop very fast algorithms to compute the PPE or SPE payoff set.

In stochastic games, different deviations can cause different state transitions, i.e. the scope for punishment is a function of the exact nature of the deviation. While obviously we have to distinguish the continuation payoff sets for the different states, a resulting additional complication is that the relevant information contained in each continuation payoff set can no more be compressed into a single number. We have to keep track separately of each player's punishment payoffs and it is not possible anymore to separate incentive constraints into a measure of all players' joint current incentives to deviate and joint measures characterizing future payoffs. Fortunately, however, the results below show that it is still possible to characterize the PPE payoff set with a relatively simple class of equilibria. Furthermore, the algorithms developed in this paper to compute the equilibrium payoff sets are still quite quick for a moderate number of states and action profiles.

3.5 Characterization

Let $W(x)$ denote the set of PPE continuation payoffs at the action stage in state x. Consider a PPE in which after some history a (possibly mixed) action profile $\alpha \in \mathcal{A}(x)$ is played in state x. Let $\tau(x', y)$ denote the subsequent vector of gross transfers if subsequently state $x' \in X$ and signal $y \in Y$ realize and let $w(x', y, \tau') \in W(x)$ be the vector of continuation payoffs at the subsequent action stage if actual transfers τ' have been conducted. Being part of a PPE these values satisfy the following conditions. No player i has an incentive to deviation from any pure action a_i in the support of α_i, i.e.for all $i \in \{1, ...n\}$, all $a_i \in \text{supp}(\alpha_i)$

and all $\hat{a}_i \in A_i(x)$

$$\begin{array}{rcl} (1-\delta)\pi_i(a_i, \alpha_{-i}, x) & + & \delta E[w_i(x', y, \tau) - (1-\delta)p_i^\tau(x', y)|x, a_i, \alpha_{-i}] \quad \geq \\ (1-\delta)\pi_i(\hat{a}_i, \alpha_{-i}, x) & + & \delta E[w_i(x', y, \tau) - (1-\delta)p_i^\tau(x', y)|x, \hat{a}_i, \alpha_{-i}]. \end{array} \qquad \text{(AC')}$$

No player i has incentives to deviate from any required transfers

$$-(1-\delta)\sum_{j\neq i}\max(\tau_i(x', y), 0) + w_i(x', y, \tau) \geq -(1-\delta)\sum_{j\neq i}\max(\tau'_{ij}, 0) + w_i(x', y, (\tau'_i, \tau_{-i})). \qquad \text{(TC')}$$

We say (w, τ) implements an action profile α if for all states x, we have $w(x) \in W(x)$ and (AC) and (TC) are satisfied.

The possibility of transfers allows to use a schedule of transfers and continuation equilibria an equivalent pair to (w, τ) with a simple structure that implements α. For each player i, let v^i be a payoff function that satisfies for each player i and for every state x' that $v^i(x') \in W_i(x)$ and $v_i^i(x)$ is at least as small as the lowest continuation payoff of player i if player i transfers nothing in state x':

$$\tilde{v}_i^i(x') \leq \min_{y\in Y} w_i(x', y, (0, \tau_{-i})).$$

We will commonly refer to $v_i(x')$ as a punishment payoff for player i in state x'. Let u be a payoff function such that for every state x', i) we have $u(x') \in W(x)$, ii) $u_i(x') \geq v_i^i(x')$, and iii)

$$\tilde{U}(x') \equiv \sum_{i=1}^n \tilde{u}_i(x') \geq \max_y \sum_{i=1}^n w_i(x', y, \tau).$$

Let

$$\tilde{p}(x', y) = p^\tau(x', y) + \frac{\tilde{u}(x') - w(x', y, \tau)}{1-\delta}.$$

and $\tilde{\tau}$ a profile of gross transfers that implements the net payments $\tilde{p}(x', y)$ in

such a fashion that no player simultaneously receives and makes transfers.[12] Let

$$\tilde{w}(x', y, \tau') = \begin{cases} \tilde{u}(x') & \text{if } \tau' = \tilde{\tau} \\ \tilde{v}^i(x') & \text{if } \tau'_i \neq \tilde{\tau}_i \text{ and } \tau'_{-i} = \tilde{\tau}_{-i} \\ \tilde{u}(x') & \text{otherwise.} \end{cases}$$

If no player deviates from the required transfers (or multilateral deviations take place), continuation payoffs will only depend on the state x' and be given by $\tilde{u}(x')$, if player i unilaterally deviates from the required transfers, continuation payoffs will be given by the associated punishment payoffs $v^i(x')$.

Lemma 1. *If α can be implemented with a payment functions p and continuation payoff functions w and v then α can also be implemented if for every state x' there exists payoff vectors $\tilde{u}(x')$ such that*

$$\tilde{U}(x') \equiv \sum_{i=1}^{n} \tilde{u}_i(x') \geq \max_y \sum_{i=1}^{n} w_i(x', y),$$

and for each player i there is a payoff $\tilde{v}_i(x')$ with $\tilde{v}_i(x') \leq v_i(x')$ and $\tilde{v}_i(x') \leq \tilde{u}_i(x')$. To implement α, we set continuation payoffs if no player unilaterally deviates from payments to $\tilde{w}(x', y) = \tilde{u}(x')$ and required net payments to

$$\tilde{p}(x, y) = p_i(x', y) + \frac{\tilde{u}_i(x') - w_i(x', y)}{1 - \delta}.$$

If players' actions can only be imperfectly monitored, it is sometimes only possible to implement an action profile α for given u and v if after some signals money is burned. We denote by

$$\hat{U}(x, \alpha, u, v) = \max_p ((1-\delta)\Pi(\alpha, x) + \delta E[U(x', y) - (1-\delta)\sum_{i=1}^{n} p_i(x', y)|x, \alpha])$$
$$\text{s.t.(AC),(PC),(BC)} \qquad \text{(LP-e)}$$

[12] Any vector of net payments p can be mapped into a matrix of gross transfers $\tilde{\tau}_{ij}$ from i to j as follows. Denote by $I_P = \{i|p_i > 0\}$ the set of net payers and by $I_R = \{i|p_i \leq 0\} \cup \{0\}$ the set of net receivers including the sink for burned money indexed by 0. For any receiver $j \in I_R$, we denote by

$$s_j = \frac{|p_j|}{\sum_{j \in I_R} |p_j|}$$

the share she receives from the total amount that is transferred or burned and assume that each net payer distributes her gross transfers according to these proportions

$$\tilde{p}_{ij} = \begin{cases} s_j p_i & \text{if } i \in I_P \text{ and } j \in I_R \\ 0 & \text{otherwise.} \end{cases}$$

Note that the sum of net payments is always non-negative, since players can only perform non-negative gross transfers.

Let $W(x)$ denote the set of PPE continuation payoffs before the action stage in state x and $W_i(x)$ the set of corresponding continuation payoffs of player i.We say a (possibly mixed) action profile $\alpha \in \mathcal{A}(x)$ can be implemented in state x if for every possible realization of subsequent state $x' \in X$ and signal $y \in Y$ there exists a payments $p(x', y')$ and subsequent continuation payoffs $w(x', y) \in W(x')$ for the case that no player unilaterally deviates from the required payment and continuation payoffs $v_i(x') \in W_i(x)$ for the case player i unilaterally deviates from a required payment that satisfy the following constraints.

The action constraints that no player i has an incentive to deviation from any pure action a_i in the support of α_i: For all $a_i \in \text{supp}(\alpha_i)$ and all $\hat{a}_i \in A_i(x)$

$$\begin{array}{lcll} (1-\delta)\pi_i(a_i, \alpha_{-i}, x) & + & \delta E[w_i(x', y) - (1-\delta)p_i(x', y)|x, a_i, \alpha_{-i}] & \geq \\ (1-\delta)\pi_i(\hat{a}_i, \alpha_{-i}, x) & + & \delta E[w_i(x', y) - (1-\delta)p_i(x', y)|x, \hat{a}_i, \alpha_{-i}]. & \text{(AC)} \end{array}$$

The payment constraint that no player i has incentives to deviate from required payments after the action stage

$$(1-\delta)p_i(x', y) \leq w_i(x', y) - v_i(x'). \tag{PC}$$

The budget constraints that the sum of payments are non-negative for all $x' \in X$ and $y \in Y$

$$\sum_{i=1}^{n} p_i(x', y) \geq 0. \tag{BC}$$

This sum of payments is simply the total amount of money that is burned.

Lemma 2. *Assume α can be implemented with a payment functions p and continuation payoff functions w and v then α can also be implemented if for every state x' there exists payoff vectors $\tilde{u}(x')$ such that*

$$\tilde{U}(x') \equiv \sum_{i=1}^{n} \tilde{u}_i(x') \geq \max_{y} \sum_{i=1}^{n} w_i(x', y),$$

and for each player i there is a payoff $\tilde{v}_i(x')$ with $\tilde{v}_i(x') \leq v_i(x')$ and $\tilde{v}_i(x') \leq \tilde{u}_i(x')$. To implement α, we set continuation payoffs if no player unilaterally deviates from payments to $\tilde{w}(x', y) = \tilde{u}(x')$ and required net payments to

$$\tilde{p}(x, y) = p_i(x', y) + \frac{\tilde{u}_i(x') - w_i(x', y)}{1-\delta}.$$

Lemma 4 states that it becomes easier to implement an action profile if the sum of continuation payoffs gets larger or the punishment payoffs of any player are reduced in any state or after any signal. The payments $\tilde{p}$ in Lemma 1 are chosen such that player i's expected continuation payoff

$$E[u_i(x', y) - (1-\delta)p_i(x', y)|x, a_i, \alpha_{-i}]$$

given the information available at the action stage are the same for $\tilde{u}$ and $\tilde{p}$ as for u and p, no matter which action profile is played. This means transforming the original PPE by replacing the payments p by $\tilde{p}$ and subsequent continuation payoffs u by $\tilde{u}$ does not change incentives for a one shot deviation at any prior point of time.

If players' actions can only be imperfectly monitored, it is sometimes only possible to implement an action profile α for given u and v if after some signals money is burned. We denote by

$$\hat{U}(x,\alpha,u,v) = \max_{p}((1-\delta)\Pi(\alpha,x) + \delta E[U(x',y) - (1-\delta)\sum_{i=1}^{n} p_i(x',y)|x,\alpha])$$
$$\text{s.t.(AC),(PC),(BC)} \qquad \text{(LP-e)}$$

the highest expected continuation payoff that can be achieved if an action profile α shall be implemented in state x given continuation and punishment payoffs u and v. For the punishment phases, we similarly denote by

$$\hat{v}_i(x,\alpha,u,v) = \min_{p}((1-\delta)\pi_i(\alpha,x) + \delta E[u_i(x',y) - (1-\delta)p_i(x',y)|x,\alpha])$$
$$\text{s.t.(AC),(PC),(BC)} \qquad \text{(LP-i)}$$

the minimum expected payoff that can be imposed on player i if an action profile α shall be implemented. (LP-e) and (LP-i) are just linear optimization problems.

Lemma 1 guarantees that if two payoff functions u and $\tilde{u}$ have the same joint payoffs U and satisfy $u, \tilde{u} \geq v$ then (LP-k), for each $k \in \{e, 1, ..., n\}$, has the same solution for u and $\tilde{u}$. With slight abuse of notation we will therefore write these solutions as functions of joint payoffs U, i.e. as $\hat{U}(x,\alpha,U,v)$ and $\hat{v}_i(x,\alpha,U,v)$, respectively. If joint continuation payoffs are below joint punishment payoffs in some state $x' \in X$, i.e. $U(x',y) < V(x')$ for some $y \in Y$, or no solution to (LP-k) exists, we set $\hat{U}(x,\alpha,U,v) = -\infty$ and $\hat{v}_i(x,\alpha,U,v) = \infty$, respectively. The next result is also direct consequence of Lemma 4.

If players' actions can only be imperfectly monitored, it is sometimes only possible to implement an action profile α for given u and v if after some signals money is burned. We denote by

$$\hat{U}(x,\alpha,u,v) = \max_{p}((1-\delta)\Pi(\alpha,x) + \delta E[U(x',y) - (1-\delta)\sum_{i=1}^{n} p_i(x',y)|x,\alpha])$$
$$\text{s.t.(AC),(PC),(BC)} \qquad \text{(LP-e)}$$

3.6 Characterization

The one shot deviation property establishes that a profile of public strategies is a PPE if and only if after no public history any player has a profitable one shot deviation. Consider the continuation play of a PPE after some history ending

before the action stage in state x. First a (possibly mixed) action profile $\alpha \in \mathcal{A}(x)$ is played and then, when state x' arises and signal y is observed, payments $p(x', y)$ are conducted. Expected continuation payoffs after the payment stage, in case no player deviates, shall be denoted by $u_i(x', y)$. Let $v_i(x')$ denote the infimum of player i's continuation payoffs if she deviates from a required payment $p_i(x', y)$ in state x', where the infimum is taken across all possible signals y and all possible deviations from the payment. We will call $v_i(x')$ player i's punishment payoff in state x'.

Player i has no incentive for a one shot deviation from any pure action a_i in the support of α_i if and only if the following action constraints are satisfied for all $a_i \in \text{supp}(\alpha_i)$ and all $\hat{a}_i \in A_i(x)$

$$\begin{array}{rcll} (1-\delta)\pi_i(a_i, \alpha_{-i}, x) & + & \delta E[u_i(x', y) - (1-\delta)p_i(x', y)|x, a_i, \alpha_{-i}] & \geq \\ (1-\delta)\pi_i(\hat{a}_i, \alpha_{-i}, x) & + & \delta E[u_i(x', y) - (1-\delta)p_i(x', y)|x, \hat{a}_i, \alpha_{-i}]. & \text{(AC)} \end{array}$$

The following payment constraint is a necessary condition that player i has no incentives to deviate from required payments after the action stage

$$(1-\delta)p_i(x', y) \leq u_i(x', y) - v_i(x'). \tag{PC}$$

Since there is no external funding, it must also be the case that the sum of payments are non-negative

$$\sum_{i=1}^{n} p_i(x', y) \geq 0. \tag{BC}$$

This sum of payments is simply the total amount of money that is burned.

We say an action profile $\alpha \in \mathcal{A}(x)$ is implemented in state x with a payment function p given continuation and punishment payoffs u and v if the constraints (AC),(PC) and (BC) are satisfied.

For given u and p , let us define

$$U(x'|u) = \max_y \sum_{i=1}^{n} u_i(x', y)$$

Lemma 3. *Assume α is implemented in state x with a payment function p given continuation and punishment payoffs u and v. Then given continuation payoffs $\tilde{u}$ that satisfy*

$$U(x'|\tilde{u}) \geq U(x'|u) \forall x' \in X$$

and punishment payoffs $\tilde{v}$ that satisfy the payment function

$$\tilde{p}_i(x', y) = p_i(x', y) + \frac{\tilde{u}_i(x', y) - u_i(x', y)}{1-\delta}$$

given continuation and punishment payoffs $\tilde{u}$ and $\tilde{v}$ that satisfy $v_i(x') \geq \tilde{v}_i(x')$, $\tilde{u}(y, x') \geq \tilde{v}_i(x')$ and

$$\sum_{i=1}^{n} \tilde{u}_i(x', y) \geq \sum_{i=1}^{n} u_i(x', y) \forall x', y.$$

nxc

Lemma 4. *If α is implemented in state x with a payment function p given continuation and punishment payoffs u and v then α is also implemented by the payment function*

$$\tilde{p}_i(x', y) = p_i(x', y) + \frac{\tilde{u}_i(x', y) - u_i(x', y)}{1 - \delta}$$

given continuation and punishment payoffs $\tilde{u}$ and $\tilde{v}$ that satisfy $v_i(x') \geq \tilde{v}_i(x')$, $\tilde{u}(y, x') \geq \tilde{v}_i(x')$ and

$$\sum_{i=1}^{n} \tilde{u}_i(x', y) \geq \sum_{i=1}^{n} u_i(x', y) \forall x', y.$$

Lemma 4 states that it becomes easier to implement an action profile if the sum of continuation payoffs gets larger or the punishment payoffs of any player are reduced in any state or after any signal. The payments $\tilde{p}$ in Lemma 1 are chosen such that player i's expected continuation payoff

$$E[u_i(x', y) - (1 - \delta)p_i(x', y)|x, a_i, \alpha_{-i}]$$

given the information available at the action stage are the same for $\tilde{u}$ and $\tilde{p}$ as for u and p, no matter which action profile is played. This means transforming the original PPE by replacing the payments p by $\tilde{p}$ and subsequent continuation payoffs u by $\tilde{u}$ does not change incentives for a one shot deviation at any prior point of time.

If players' actions can only be imperfectly monitored, it is sometimes only possible to implement an action profile α for given u and v if after some signals money is burned. We denote by

$$\hat{U}(x, \alpha, u, v) = \max_{p}((1 - \delta)\Pi(\alpha, x) + \delta E[U(x', y) - (1 - \delta)\sum_{i=1}^{n} p_i(x', y)|x, \alpha])$$
$$\text{s.t.(AC),(PC),(BC)} \qquad \text{(LP-e)}$$

the highest expected continuation payoff that can be achieved if an action profile α shall be implemented in state x given continuation and punishment payoffs u and v. For the punishment phases, we similarly denote by

$$\hat{v}_i(x, \alpha, u, v) = \min_{p}((1 - \delta)\pi_i(\alpha, x) + \delta E[u_i(x', y) - (1 - \delta)p_i(x', y)|x, \alpha])$$
$$\text{s.t.(AC),(PC),(BC)} \qquad \text{(LP-i)}$$

the minimum expected payoff that can be imposed on player i if an action profile α shall be implemented. (LP-e) and (LP-i) are just linear optimization problems.

Lemma 1 guarantees that if two payoff functions u and $\tilde{u}$ have the same joint payoffs U and satisfy $u, \tilde{u} \geq v$ then (LP-k), for each $k \in \{e, 1, ..., n\}$, has the same solution for u and $\tilde{u}$. With slight abuse of notation we will therefore write these

solutions as functions of joint payoffs U, i.e. as $\hat{U}(x,\alpha,U,v)$ and $\hat{v}_i(x,\alpha,U,v)$, respectively. If joint continuation payoffs are below joint punishment payoffs in some state $x' \in X$, i.e. $U(x',y) < V(x')$ for some $y \in Y$, or no solution to (LP-k) exists, we set $\hat{U}(x,\alpha,U,v) = -\infty$ and $\hat{v}_i(x,\alpha,U,v) = \infty$, respectively. The next result is also direct consequence of Lemma 4.

Lemma 5. *For all $i,j = 1,...,n$ and all $x', x \in X$ and $y \in Y$*

- *$\hat{U}(x,\alpha,U,v)$ is weakly increasing in $U(x')$ and weakly decreasing in $v_j(x')$,*
- *$\hat{v}_i(x,\alpha,U,v)$ is weakly decreasing in $U(x',y)$ and weakly increasing in $v_j(x')$.*

Lemma 5 states that higher joint continuation payoffs or lower punishment payoffs in any state x' allow to implement higher joint payoffs $\hat{U}(.)$ and lower punishment payoffs $\hat{v}_i(.)$. Reminiscent to the decomposition methods by APS, one can interpret $\hat{U}(x,\alpha,U,v)$ as the highest joint payoffs and $\hat{v}_i(x,\alpha,U,v)$ as the lowest payoff for player i that can be decomposed in state x with an action profile α given a continuation payoff whose highest joint payoffs for each state are given by U and lowest payoffs for each state and player by v. Lemma 5 loosely corresponds to the fact that in APS the set of payoffs that can be decomposed gets weakly larger if the set of continuation payoffs gets larger.

Let $\mathcal{A}(x,U,v) \subset \mathcal{A}(x)$ be the subset of action profiles that can be implemented in state x given U and v for some payment function. This means solutions to LP-e and LP-i exist if and only if $\alpha \in \mathcal{A}(x,U,v)$.

Lemma 6. *The set of implementable action profiles $\mathcal{A}(x,U,v)$ is compact and upper-hemi continuous in U and v. $\hat{U}(x,\alpha,U,v)$ and $\hat{v}_i(x,\alpha,U,v)$ are continuous in α for all $\alpha \in \mathcal{A}(x,U,v)$.*

We can now establish our key result that there exists optimal simple equilibria that can implement any PPE payoff.

Theorem 1. *Assume a PPE exists. Then an optimal simple equilibrium with an action plan $(\bar{\boldsymbol{\alpha}}^k)_{k\in\mathcal{K}}$ exists such that by varying its upfront transfers in an incentive compatible way, every PPE payoff can be implemented. The sets of PPE continuation payoffs for every state x are compact; their maximal joint continuation payoffs and minimal punishment payoffs satisfy*

$$\bar{U}(x) = \hat{U}(x,\bar{\boldsymbol{\alpha}}^e,\bar{U},\bar{v})\forall x,$$
$$\bar{v}_i(x) = \hat{v}_i(x,\bar{\boldsymbol{\alpha}}^i,\bar{U},\bar{v})\forall x,i.$$

4 Computing Payoff Sets and Optimal Simple Equilibria

Based on the previous results, this section describes different methods to exactly compute or to approximate the set of PPE payoffs and to find (optimal) simple equilibria to implement these payoffs.

4.1 Optimal payment plans and a brute force algorithm

For a given simple strategy profile, we denote expected continuation payoffs in the equilibrium phase and the punishment phase for player i by u^s and v_i^s, respectively. The equilibrium phase payoff are implicitly defined by

$$u_i^s(x) = (1-\delta)\pi_i(\boldsymbol{\alpha}^e, x) + \delta E[-(1-\delta)p_i^e(x', y, x) + u_i^s(x')|\boldsymbol{\alpha}^e, x]). \tag{1}$$

Player i's punishment payoffs are given by

$$v_i^s(x) = (1-\delta)\pi_i(\boldsymbol{\alpha}^i, x) + \delta E[-(1-\delta)p_i^i(x', y, x) + u_i^s(x')|\boldsymbol{\alpha}^i, x]). \tag{2}$$

Let (AC-k), (PC-k) and (BC-k) denote the action payment and budget constraints for policy $\boldsymbol{\alpha}^k$ and payment function $p^k(.)$ given continuation and punishment payoffs u^s and v^s.

We say a payment plan is *optimal* for a given action plan if all constraints (AC-k), (PC-k) and (BC-k) are satisfied and there is no other payment plan that satisfies these conditions and yields a higher joint payoff $U^s(x)$ or a lower punishment payoff $v_i^s(x)$ for some state x and some player i.

Proposition 2. *There exists a simple equilibrium with an action plan $(\boldsymbol{\alpha}^k)_{k\in\mathcal{K}}$ if and only if there exists a payment plan $(\bar{p}^k)_{k\in\mathcal{K}}$ that solves the following linear program*

$$(\bar{p}^k)_k \in \arg\max_{(p^k)_k} \sum_{x\in X}\sum_{i=1}^{n}(u_i^s(x) - v_i^s(x)) \qquad \text{(LP-OPP)}$$
$$s.t. \text{(AC-k)}, \text{(PC-k)}, \text{(BC-k)} \forall k = e, 1, ..., n$$

$(\bar{p}^k)_{k\in\mathcal{K}}$ is an optimal payment plan for $(\boldsymbol{\alpha}^k)_{k\in\mathcal{K}}$. A simple equilibrium with action plan $(\boldsymbol{\alpha}^k)_{k\in\mathcal{K}}$ and an optimal payment plan satisfies

$$\begin{aligned} U^s(x) &= \hat{U}(x, \boldsymbol{\alpha}^e, U^s, v^s), \\ v_i^s(x) &= \hat{v}_i(x, \boldsymbol{\alpha}^i, U^s, v^s). \end{aligned}$$

An optimal simple equilibrium has an optimal action plan and a corresponding optimal payment plan. Together with Theorem 1, this result directly leads to a brute force algorithm to characterize the set of pure strategy PPE payoffs given a finite action space: simply go through all possible action plans and solve (LP-OPP). An action plan with the largest solution will be optimal. Similarly, one can obtain a lower bound on the set of mixed strategy PPE payoffs, by solving (LP-OPP) for all mixing probabilities from some finite grid. Despite an infinite number of mixed action plans, the optimization problem for each mixed action plan is finite because only deviations to pure actions have to be checked.

The weakness of this method is that it can become computationally infeasible, already for moderately sized action and state spaces. That is because the number

of possible action plans grows very quickly in the number of states and actions per state and player.

For particular applications there will exist more efficient methods to jointly optimize over payment and action plans than a brute force search over all action plans. In general, however, the joint optimization problem is non-convex, as e.g. joint equilibrium phase payoffs U^s are not jointly concave in the actions and payments. One can therefore not in general rely on efficient methods for convex optimization problems that guarantee a global optimum. For mixed strategy equilibria, there is the additional complication that number of action constraints depends on the support of the mixed action profiles that shall be implemented.

4.2 Decomposition Methods for Outer and Inner Approximations

In this subsection we illustrate how the methods for repeated games of APS and Judd, Yeltekin and Conklin (2003, henceforth JYC) can be translated to our framework to get an algorithm that allows outer and inner approximations of the equilibrium payoff set.

Let $D : \mathbb{R}^{(n+1)|X|} \to \mathbb{R}^{(n+1)|X|}$ be a decomposition operator that maps a collection (U, v) of joint equilibrium and punishment payoffs into a new collection of of such payoffs (U', v') that satisfy for each state $x \in X$:

$$U'(x) = \max_{\alpha \in \mathcal{A}(x)} \hat{U}(x, \alpha, U, v), \tag{3}$$

$$v_i'(x) = \min_{\alpha \in \mathcal{A}(x)} \hat{v}_i(x, \alpha, U, v). \tag{4}$$

This means D computes the largest joint equilibrium payoff and lowest punishment payoffs that can be decomposed with any action profiles $\alpha \in \mathcal{A}(x)$. For any integer m, we denote by D^m the operator that m times applies D.

Proposition 3. *Let U^0 and v^0 be payoffs satisfying $U^0(x) \geq \bar{U}(x)$ and $v_i^0(x) \leq \bar{v}_i(x)$ for all $x \in X$ and all $i = 1, ..., n$. Then the resulting payoffs after m decomposition steps, i.e. $D^m(U^0, v^0)$, converge to $\bar{U}$ (from above) and $\bar{v}$ (from below) as $m \to \infty$.*

Repeatedly applying the decomposition operator D yields in every round a tighter outer approximation for $\bar{U}$ and $\bar{v}$ and of the corresponding payoff set of PPE equilibria.

A tighter outer approximation is obtained more quickly if the initial values U^0 and v^0 are closer to $\bar{U}$ and $\bar{v}$. For games with imperfect monitoring, good initial values U^0 and v^0 will be the optimal joint equilibrium and punishment payoffs of a perfect monitoring version of the game, which can be solved much faster using methods that will be described in Section 5.

To obtain bounds on the approximation error, it is also necessary to obtain inner approximations of the equilibrium payoff sets. To find an inner approximation for the payoff set of a repeated game, JYC suggest to shrink the outer approximation of the payoff set by a small amount, say 2%-3% and to apply the decomposition operator on the shrunken set. If the decomposition operator increases the shrunken set then the decomposed set forms an inner approximation of the equilibrium payoff set.

A similar approach can be used in our framework. One reduces the outer approximations of $\bar{U}$ and increases the outer approximations of $\bar{v}$ by a small amount and then applies the decomposition operator D on these shrunken values. If the decomposition increases all joint equilibrium payoffs and reduces all punishment payoffs, we have found an inner approximation. For each decomposition step, we get a corresponding action plan consisting of the optimizers of (3) and (4). Propoposition 4 shows that for this action plan the linear program (LP-OPP) always has a solution. We obtain from that solution a simple equilibrium and an even tighter inner approximation.

Proposition 4. *There exists a simple equilibrium with an action plan* $(\boldsymbol{\alpha}^k)_{k\in\mathcal{K}}$ *if and only if there exists joint equilibrium and punishment payoffs* U *and* v *such that*

$$\begin{aligned} \hat{U}(x, \boldsymbol{\alpha}^e, U, v) &\geq U(x) \forall x \in X, && (5)\\ \hat{v}_i(x, \boldsymbol{\alpha}^i, U, v) &\leq v_i(x) \forall x \in X, i = 1, ..., n. && (6) \end{aligned}$$

An alternative method to search for an inner approximation is to run (LP-OPP) for the action plans that result from the decomposition steps of the outer approximation. If a solution exists, it also forms an inner approximation.

Inner and outer approximations allow to reduce for every state and phase the set of action profiles that can possible be part of an optimal action plan. Let (U^{in}, v^{in}) and (U^{out}, v^{out}) describe the inner and outer approximations. Consider a state x and an action profile $\alpha \in \mathcal{A}(x)$. If α cannot be implemented given U^{out} and v^{out}, there does not exist any PPE in which α is played and we can dismiss it. If α can be implemented, but

$$\hat{U}(x, \alpha, U^{out}, v^{out}) < U^{in}(x)$$

then α will not be played in the equilibrium phase in state x of an optimal equilibrium, since even with the outer approximations of U and v it can decompose a lower joint payoff than the current inner approximation. Similarly, if

$$\hat{v}_i(x, \alpha, U^{out}, v^{out}) > v_i^{in}(x)$$

then α will not be an optimal punishment profile for player i in state x.

Hence, finer inner and outer approximations speed up the computation of new approximations since a smaller set of action profiles has to be considered. Moreover, once the number of candidate action profiles has been sufficiently reduced, it can become tractable to compute the exact payoff set by applying the brute force method from Subsection 4.1 on the remaining action plans.

5 Perfect monitoring

5.1 Decomposition

In this section, attention is restricted to equilibria in pure strategies in games with perfect monitoring, i.e. players commonly observe all past action profiles. The following proposition shows how the problems (LP-k) drastically simplify in this case.

Proposition 5. *Assume monitoring is perfect, a is a pure strategy profile, and $U(x') \geq V(x') \forall x' \in X$. Then*

1. *all solutions to (LP-e) satisfy*

$$\hat{U}(a, x, U, v) = (1-\delta)\Pi(a, x) + \delta E[U(x')|a, x], \tag{7}$$

2. *all solutions to (LP-i) satisfy for all $i \in \{1, ..., n\}$*

$$\hat{v}_i(a, x, U, v) = \max_{\hat{a}_i \in A_i(x)} \{(1-\delta)\pi_i(\hat{a}_i, a_{-i}, x) + \delta E[v_i(x')|\hat{a}_i, a, x]\}, \tag{8}$$

3. *a solution to (LP-k) for given a, x, U and v exists if and only if for all $k \in \mathcal{K}$*

$$(1-\delta)\Pi(a, x) + \delta E[U(x')|a, x] \geq \sum_{i=1}^{n} \max_{\hat{a}_i \in A_i(x)} \{(1-\delta)\pi_i(\hat{a}_i, a_{-i}, x) + \delta E[v_i(x')|\hat{a}_i, a_{-i}, x]\}. \tag{9}$$

These results are quite intuitive. Since deviations can be perfectly observed, there is no need to burn money on the equilibrium path. Equation (7) simply describes the joint continuation payoffs in the absence of money burning. Furthermore, perfect monitoring allows in punishment phases, to always reduce the punished player's payoff to his best reply payoff given that continuation payoffs are given by v. These best-reply payoffs are given by (8). Condition (9) is the sum of the resulting action constraints across all players. That this condition is sufficient is due to the fact that payments can be used to perfectly balance incentives to deviate across players in the way Section 3.2 has exemplified.

Proposition 5 allows a quick implementation of the decomposition steps to find inner and outer approximations described in Section 4. For a decomposition step one just has to evaluate conditions (9) and (7) or (8) for the candidate set of possibly optimal action profiles; no linear optimization problem has to be solved.

5.2 Simple Equilibria with Optimal Payment Plans

We now show, how for a given action plan one can compute joint equilibrium payoffs and punishment payoffs under an optimal payment plan. Assume a simple equilibrium exists for an action plan $(\boldsymbol{a}^k)_{k \in \mathcal{K}}$. Recall from Proposition 2 that

$$\hat{U}(\boldsymbol{a}^e(x), x, U^s, v^s) = U^s(x).$$

Together with (9) , we find that U^s can be easily computed by solving the following system of linear equations:

$$U^s = (1-\delta)\Pi(\boldsymbol{a}^e) + \Omega(\boldsymbol{a}^e)U^s \tag{10}$$

where $\Omega(\boldsymbol{a})$ shall denote the transition matrix between states given that players follow the policy $\boldsymbol{a}$.

For the punishment states, Propositions 2 and 5 imply that punishment payoffs must satisfy the following Bellman equation:

$$v_i^s(x) = \max_{\hat{a}_i \in A_i(x)} \{(1-\delta)\left(\pi_i(\hat{a}_i, \boldsymbol{a}_{-i}^i, x)\right) + \delta E[v_i^s(x')|x, \hat{a}_i, \boldsymbol{a}_{-i}^i]\}. \tag{11}$$

It follows from the contraction mapping theorem that there exists a unique payoff vector v_i^s that solves this Bellman equation. The solution corresponds to player i's payoffs in case she refuses to make any payments and plays a best reply in every period assuming that other players follow the policy $\boldsymbol{a}_{-i}^i$ in all future.

Finding player i's punishment payoffs constitutes a single agent dynamic optimization problem, more precisely, a discounted Markov decision process. One can compute v_i^s, for example, with the policy iteration algorithm.[13] It consists of a policy improvement step and a value determination step. The policy improvement step calculates for some punishment payoffs v_i an optimal best-reply action $\tilde{\boldsymbol{a}}_i(x)$ for each state x, which solves

$$\tilde{\boldsymbol{a}}_i(x) \in \arg\max_{a_i \in A_i(x)} \{(1-\delta)\left(\pi_i(a_i, \boldsymbol{a}_{-i}^i, x)\right) + \delta E[v_i(x')|x, a_i, \boldsymbol{a}_{-i}^i]\}.$$

The value determination step calculates the corresponding payoffs of player i by solving the system of linear equations

$$v_i = (1-\delta)\pi_i(\tilde{\boldsymbol{a}}_i, \boldsymbol{a}_{-i}^i) + \delta\Omega(\tilde{\boldsymbol{a}}_i, \boldsymbol{a}_{-i}^i)v_i. \tag{12}$$

Starting with some arbitrary payoff function v_i, the policy iteration algorithm alternates between policy step and value iteration step until the payoffs do not change anymore, in which case they will satisfy (11).

Together with Propositions 2 and 5 these observations lead to the following result:

Corollary 1. *Under perfect monitoring, the joint equilibrium payoffs U^s and player i's punishment payoffs v_i^s in a simple equilibrium with (pure) action plan $(\boldsymbol{a}^k)_{k\in\mathcal{K}}$ and an optimal payment plan are given by the solutions of (10) and (11), respectively. A simple equilibrium with action plan $(\boldsymbol{a}^k)_{k\in\mathcal{K}}$ exists if and only if for every state x and every phase $k \in \{e, 1, ..., n\}$*

$$(1-\delta)\Pi(\boldsymbol{a}^k, x) + \delta E[U^s(x)|\boldsymbol{a}^k, x] \geq \sum_{i=1}^{n} \max_{\hat{a}_i \in A_i(x)} \{(1-\delta)\pi_i(\hat{a}_i, \boldsymbol{a}_{-i}^k, x) + \delta E[v_i^s(x')|\hat{a}_i, \boldsymbol{a}_{-i}^k(x), x]\}. \tag{13}$$

[13] For details on policy iteration, convergence speed and alternative computation methods to solve Markov Decision Processes, see e.g. Puterman (1994).

When applying the methods described in Section 4, Corollary 1 is useful for computing inner approximations and to find an optimal simple equilibrium once the candidate set of action plans has been sufficiently reduced.

5.3 A Policy Elimination Algorithm

We now develop a quick policy elimination algorithm that exactly computes the set of pure strategy SPE payoffs in stochastic games with perfect monitoring and a finite action space.

In every round of the algorithm there is a candidate set of action profiles $\hat{\boldsymbol{A}}(x) \subset A(x)$ which have not yet been ruled out as being possible played in some simple equilibrium. $\hat{\boldsymbol{A}} = \times_{x \in X} \hat{\boldsymbol{A}}(x)$ shall denote the corresponding set of policies. Let $U^s(.|\boldsymbol{a}^e)$ denote the solution of (10) for equilibrium phase policy $\boldsymbol{a}^e$ and $v_i^s(.|\boldsymbol{a}^i)$ the solution of (11) under punishment policy $\boldsymbol{a}^i$. We denote by

$$U^s(x|\hat{\boldsymbol{A}}) = \max_{\boldsymbol{a}^e \in \hat{A}} U^s(x|\boldsymbol{a}^e) \tag{14}$$

the maximum joint payoff that can be implemented in state x using equilibrium phase policies from $\hat{A}$. The problem of computing $U^e(.|\hat{\boldsymbol{A}})$ is a finite discounted Markov decision process (MDP). Standard results for MDP establish that there always exists a policy $\hat{\boldsymbol{a}}^e(\hat{\boldsymbol{A}}) \in \hat{\boldsymbol{A}}$ that solves (14) simultaneously in all states. One can compute $U^e(.|\hat{\boldsymbol{A}})$ with a policy iteration algorithm, for which the value determination step is given by (10).

For the punishment phases, we define by

$$v_i^s(x|\hat{\boldsymbol{A}}) = \min_{\boldsymbol{a}^i \in \hat{A}} v_i^s(x|\boldsymbol{a}^i) \tag{15}$$

player i's minimal punishment payoff in state x across all punishment policies in $\hat{A}$. Computing $v_i^s(x|\hat{A})$ constitutes a nested dynamic optimization problem: one has to compute player i's best-reply policy against each considered candidate punishment policy. The analysis of this problem is relegated to Appendix A. It is shown that there always exists a punishment policy $\hat{\boldsymbol{a}}^i(\hat{\boldsymbol{A}}) \in \hat{\boldsymbol{A}}$ that solves (15) simultaneously for all states $x \in X$ and a nested policy iteration method is developed that strictly improves punishment policies in each step and allows to quickly compute $v_i(.|\hat{\boldsymbol{A}})$.

The policy elimination algorithm works as follows:

Algorithm. *Policy elimination algorithm to find optimal action plans*

0. *Let $r = 0$ and initially consider all policies as candidates: $\hat{\boldsymbol{A}}^0 = A$*
1. *Compute $U^s(.|\hat{\boldsymbol{A}}^r)$ and a corresponding optimal equilibrium phase policy $\hat{\boldsymbol{a}}^e(\hat{\boldsymbol{A}}^r)$*

2. *For every player i compute $v_i^s(.|\hat{\boldsymbol{A}}^r)$ and a corresponding optimal punishment policy $\hat{\boldsymbol{a}}^i(\hat{\boldsymbol{A}}^r)$*

3. *For every state x, let $\hat{\boldsymbol{A}}^{r+1}(x)$ be the set of all action profiles that satisfy condition (9) from Proposition (5) using $U^s(.|\hat{\boldsymbol{A}}^r)$ and $v_i^s(.|\hat{\boldsymbol{A}}^r)$ as equilibrium phase and punishment payoffs.*

4. *Stop if the optimal policies $\hat{\boldsymbol{a}}^k(\hat{\boldsymbol{A}}^r)$ are contained in $\hat{\boldsymbol{A}}^{r+1}$. They then constitute an optimal action plan. Also stop if for some state x the set $\hat{\boldsymbol{A}}^{r+1}$ is empty. Then no SPE in pure strategies exists. Increment the round r and repeat Steps 1-3 until one of the stopping conditions is satisfied.*

The policy elimination algorithm always stops in a finite number of rounds. It either finds an optimal action plan $(\bar{\boldsymbol{a}}^k)_{k\in\mathcal{K}}$ or yields the result that no SPE in pure strategies exists. Given our previous results, it is straightforward that this algorithm works.

Unless the algorithm stops in the current round, Step 3 always eliminates some candidate policies, i.e. the set of candidate policies $\hat{\boldsymbol{A}}^r$ gets strictly smaller with each round. Therefore $U^s(x|\hat{\boldsymbol{A}}^r)$ weakly decreases and $v_i^s(x|\hat{\boldsymbol{A}}^r)$ weakly increases each round. Condition (9) is easier satisfied for higher values of $U^s(x|\hat{\boldsymbol{A}}^r)$ and for lower values of $v^s(x|\hat{\boldsymbol{A}}^r)$. Therefore, a necessary condition that an action profile is ever played in a simple equilibrium is that it survives Step 3. Conversely, if the polices $\hat{\boldsymbol{a}}^k(\hat{\boldsymbol{A}}^r)$ all survive Step 3, it follows from Corollary 1 that a simple equilibrium with these policies exists. That they constitute an optimal action plan simply follows again from the fact that $U^s(x|\hat{\boldsymbol{A}}^r)$ weakly decreases and $v_i^s(x|\hat{\boldsymbol{A}}^r)$ weakly increases each round. That the algorithm terminates in a finite number of round is a consequence of the finite action space and the fact that the set of possible policies $\hat{\boldsymbol{A}}^r$ gets strictly smaller each round.

6 Examples

6.1 Quantity competition with stochastic reserves

As first example, we consider a stochastic game version of the example Cournot used to motivate his famous model of quantity competition. There are two producers of mineral water producers who have finite water reserves in their reservoirs. A state is two dimensional $x = (x_1, x_2)$ and x_i describes the amount of water currently stored in firm i's reservoir. In each period, each firm i simultaneously chooses an integer amount of water $a_i \in \{0, 1, 2, ..., x_i\}$ that it takes from its reservoir and sells on the market. Market prices are given by an inverse demand function $P(a_1, a_2)$. A firm's reserves can increase after each period by some random integer amount, up to a maximal reservoir capacity of $\bar{x}$.

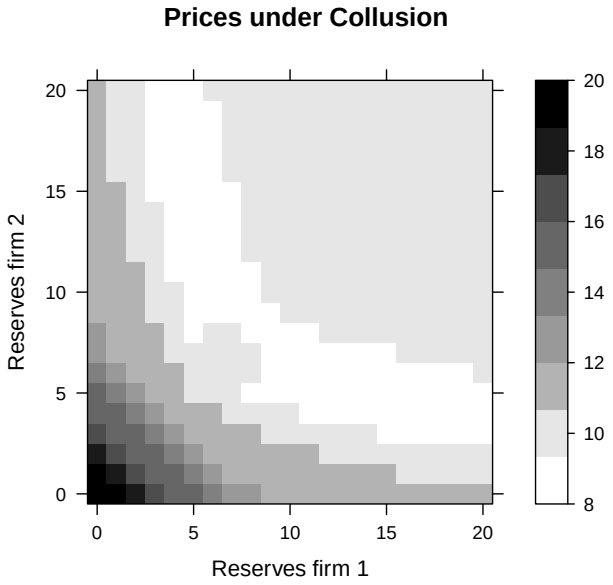

Figure 2: Optimal collusive prices as function of firms' reserves. Brighter areas correspond to lower prices.

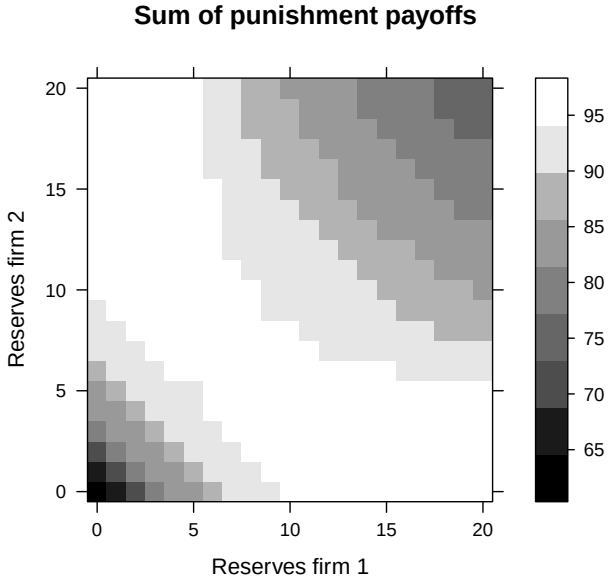

Figure 3: Sum of punishment payoffs $\bar{v}_1(x) + \bar{v}_2(x)$. Darker areas correspond to lower punishment payoffs.

The example is solved with an R implementation of the policy elimination algorithm described in Section 5.3. The following parameters are used: maximum capacity of each firm $\bar{x} = 20$, discount factor $\delta = \frac{2}{3}$, inverse demand function $P(a_1, a_2) = 20 - a_1 - a_2$, and reserves refill with equal probability by 3 or 4 units

each period.[14] Figure 2 illustrates the solution of the dynamic game by showing the market prices in an optimal collusive equilibrium as a function of the oil reserves of both firms.

Starting from the lower left corner, one sees that prices are initially reduced when firms' water reserves increase. Yet, the upper right corner illustrates that equilibrium prices are not monotonically decreasing in reserves: once reserves become sufficiently large, prices increase again. An intuitive reason for this observation is that once reserves grow large, it becomes easier to facilitate collusion as deviations from a collusive agreement can be punished more severely by a credible threat to selling large quantities in the next period.

Figure 3 fosters this intuition. It illustrates the sum of punishment payoffs $\bar{v}_1(x) + \bar{v}_2(x)$ that can be imposed on players as function of the current state. One sees how harsh punishments can be credibly implemented when reserves are large.

6.2 A Principal-Agent Game with a Durable Good

This example illustrates how our results can be used to easily obtain closed form solutions in a simple stochastic game that describes a principal agent relationship. An agent (player 1) can produce a single durable good for a principal (player 2). If the product has been successfully produced, the state of the world will be given by x_1 and otherwise it is x_0. In state x_0, the agent can choose production effort $e \in [0, 1]$ and the product will be successfully produced in the next period with probability e. The principal's stage game payoff is 1 in state x_1 and 0 in state x_0. The agent's stage game payoff is $-ce$ where $c > 0$ is an exogenous cost parameter. For the moment, we assume that once the product has been produced, the state stays x_1 forever.

Perfect monitoring We first consider the case of perfect monitoring. In the terminal state x_1 joint payoffs are given by $U(x_1) = 1$. The joint equilibrium payoff in state x_0 in a simple equilibrium with effort e satisfies

$$\begin{aligned} U^s(e, x_0) &= -(1-\delta)ce + \delta(e + (1-e)U^s(e, x_0)) \Leftrightarrow \\ U^s(e, x_0) &= \frac{\delta - (1-\delta)c}{\delta e + (1-\delta)} e. \end{aligned}$$

We assume $(1-\delta)k < \delta$, i.e. it is socially efficient that the agent exerts maximum effort. In an optimal simple equilibrium, the agent's punishment payoff in both states are 0 and the principal's punishment payoff's are $v_2^s(x_0) = 0$ and $v_2^s(x_1) = 1$. Using Corrollary (1) one then finds that effort e can be implemented if and only if

$$(1-\delta)\, c \leq \delta^2(1-e). \tag{16}$$

[14] Solving this example with 21*21=441 states takes less than a minute on an average notebook bought in 2009.

Condition (16) implies that positive effort can be induced under sufficiently large discount factors, while it is not possible to induce full effort $e = 1$ under any given discount factor $\delta \in [0, 1)$. On first thought, that result seems surprising since effort costs are linear. The intuition is simple, however. Once the product has been successfully built, the game is in the absorbing state x_1. Since payoffs in x_1 are fixed, the principal won't conduct any transfers. How can the principal credible reimburse the agent for positive effort? This is only possible with a transfer in the case that the agent has exerted high effort but the project has not been successful: the probability that the agent gets reimbursed for the required effort level is thus given by $(1 - e)$. Thus, the agent cannot be reimbursed for full effort, but there is a positive chance to get reimbursed for partial effort.

Imperfect Monitoring Consider now imperfect monitoring in the form that the principal can only observe the realized state. It is straightforward that then in every simple equilibrium the agent chooses zero effort and no transfers are conducted. That is because the principal cannot be induced to make any payments in state x_1 and at the same time any transfers by the principal in the state x_0 increase the agent's incentives not to conduct any effort. This observation illustrates how monitoring imperfections may be much more devastating in a stochastic game than in a repeated game: in a standard repeated principal agent games with a noisy public signal about the agent's effort choice, (approximately) socially optimal effort levels can always be implemented under sufficiently large discount factors.

Costly punishment Assume now that in state x_1 the agent can choose destructive effort $d \in \{0, 1\}$ where $d = 1$ has the consequence that the product is destroyed in the next round while for $d = 0$ the product remains intact. The agent incurs costs for destructive efforts of size kd with $k \geq 0$.

Since d can be induced from the state transitions, we can assume w.l.o.g. that d is perfectly observable. It follows from condition (9) in Proposition 5 that the agent can be induced to destroy the product if and only if

$$(1 - \delta)k \leq \delta U^s(e, x_0). \tag{17}$$

In an optimal simple equilibrium with $d = 1$, the principal's punishment payoffs satisfy

$$\begin{aligned} v_2^s(d = 1, x_0) &= 0 \\ v_2^s(d = 1, x_1) &= (1 - \delta). \end{aligned}$$

The maximum payment the principal can then be induced to make in state x_1 is thus given by

$$p_2^{max}(x_1) = \frac{1}{1 - \delta}(1 - (1 - \delta)) = \frac{\delta}{1 - \delta}.$$

Assume effort e is imperfectly monitored and only the realized state can be observed. Given $p_2^{max}(x_1)$, the agent can be induced to implement an effort level $e > 0$ if and only if

$$\begin{aligned} ce &\leq \delta e \frac{\delta}{1-\delta} \Leftrightarrow \\ (1-\delta)c &\leq \delta^2. \end{aligned} \tag{18}$$

Since $U^s(e, x_0)$ is strictly increasing in e, in every optimal simple equilibrium in which the agent chooses positive effort, we have maximal effort $e = 1$. The joint equilibrium phase payoff in state x_0 is then given by

$$U^s(e = 1, x_0) = -(1-\delta)c + \delta.$$

Inserting this payoff into condition (17), we find that high effort can be implemented if and only if

$$(1-\delta)(\delta c + k) \leq \delta^2 \tag{19}$$

and (18) hold. Hence, if the agent has the opportunity to exert costly effort to punish the principal after a successful project, full effort provision can be implemented under sufficiently large discount factors. Note that conditions (18) and (19) are also necessary and sufficient for the existence of a simple equilibrium with positive effort in a perfect monitoring variant of the principal agent game in which only binary effort choices $e \in \{0, 1\}$ are possible.

Optimal penal codes vs grim-trigger punishments The constructed simple equilibria use optimal penal codes in which the agent uses a punishment that is costly in the current period and that is only conducted because it is rewarded in the future. In repeated games, simple grim-trigger punishments that punish any deviation by an infinite reversion to a stage game Nash equilibrium are generally also able to implement cooperative actions given sufficiently large discount factors. In the current example, a natural variant of grim-trigger punishments would be to punish any deviation from required effort or transfers by reverting to the unique MPE of the stochastic game: $e = d = 0$ and no transfers. However, such grim-trigger punishments won't be able to implement any positive effort by the agent, since the principal could not be induced to make positive transfers in state x_1. The ineffectiveness of grim-trigger punishments in this simple example illustrates that for stochastic games it is particularly useful to have a simple characterization of equilibria with optimal penal codes.

References

Abreu, D., 1986. Extremal equilibria of oligopolistic supergames, Journal of Economic Theory, 39(1), pp.191–225.

Abreu, D., 1988. On the theory of infinitely repeated games with discounting. Econometrica, pp.383–396.

Abreu, D., Pearce, D. & Stacchetti, E., 1990. Toward a theory of discounted repeated games with imperfect monitoring. Econometrica, pp.1041–1063.

Abreu, D. & Sannikov, Y., 2011. An Algorithm for Two Player Repeated Games with Perfect Monitoring.

Baliga, S. & Evans, R., 2000. Renegotiation in repeated games with side-payments. Games and Economic Behavior, 33(2), pp.159–176.

Benkard, C.L., 2000. Learning and forgetting: The dynamics of aircraft production. American Economic Review, 90(4), pp.1034–1054.

Besanko, D. et al., 2010. Learning-by-doing, organizational forgetting, and industry dynamics. Econometrica, 78(2), pp.453–508.

Besanko, D. & Doraszelski, U., 2004. Capacity dynamics and endogenous asymmetries in firm size. RAND Journal of Economics, pp.23–49.

Ching, A.T., 2010. A dynamic oligopoly structural model for the prescription drug market after patent expiration. International Economic Review, 51(4), pp.1175–1207.

Doornik, K., 2006. Relational contracting in partnerships. Journal of Economics & Management Strategy, 15(2), pp.517–548.

Doraszelski, U. & Markovich, S., 2007. Advertising dynamics and competitive advantage. The RAND Journal of Economics, 38(3), pp.557–592.

Dutta, P.K., 1995. A folk theorem for stochastic games. Journal of Economic Theory, 66(1), pp.1–32.

Fong, Y. & Surti, J., 2009, On the Optimal Degree of Cooperation in the Repeated Prisoner's Dilemma with Side Payments, Games and Economic Behavior, 67(1), 277-291.

Fudenberg, D. & Yamamoto, Y., 2011. The folk theorem for irreducible stochastic games with imperfect public monitoring. Journal of Economic Theory.

Gjertsen H, Groves T., Miller D., Niesten E., Squires D. & Watson J., 2010. A Contract-Theoretic Model of Conservation Agreements, mimeo.

Goldlücke, S. & Kranz, Sebastian, 2012. Infinitely repeated games with public monitoring and monetary transfers. Journal of Economic Theory (forthcoming).

Harrington, J.E. & Skrzypacz, A., 2011. Private monitoring and communication in cartels: Explaining recent collusive practices. The American Economic Review, 101(6), pp.2425–2449.

Harrington Jr, J.E. & Skrzypacz, A., 2007. Collusion under monitoring of sales. The RAND Journal of Economics, 38(2), pp.314–331.

Hörner, J. T. Sugaya, S. Takahashi, & N. Vieille. 2011. Recursive methods in discounted stochastic games: An algorithm for $\delta \to 1$ and a folk theorem. Econometrica, 79(4), pp.1277–1318.

Judd, K.L., Yeltekin, S. & Conklin, J., 2003. Computing supergame equilibria. Econometrica, 71(4), pp.1239–1254.

Klimenko, M., Ramey, G. & Watson, J., 2008. Recurrent trade agreements and the value of external enforcement. Journal of International Economics, 74(2), pp.475–499.

Kranz, S. & Ohlendorf, S., 2009. Renegotiation-proof relational contracts with side payments. SFB/TR 15 Discussion Paper

Levin, J., 2002. Multilateral contracting and the employment relationship. The Quarterly Journal of Economics, 117(3), p.1075.

Levin, J., 2003. Relational incentive contracts. The American Economic Review, 93(3), pp.835–857.

Malcomson, J.M., 1999. Individual employment contracts. Handbook of labor economics, 3, pp.2291–2372.

Markovich, S. & Moenius, J., 2009. Winning while losing: Competition dynamics in the presence of indirect network effects. International Journal of Industrial Organization, 27(3), pp.346–357.

Miller, D.A., Watson, J. & others, 2011. A theory of disagreement in repeated games with bargaining. mimeo.

Pakes, A. & McGuire, P., 1994. Computing Markov-Perfect Nash Equilibria: Numerical Implications of a Dynamic Differentiated Product Model. The Rand Journal of Economics, pp.555–589.

Pakes, A. & McGuire, P., 2001. Stochastic algorithms, symmetric Markov perfect equilibrium, and the "curse"of dimensionality. Econometrica, 69(5), pp.1261–1281.

Puterman, M.L., 1994. Markov decision processes: Discrete stochastic dynamic programming, John Wiley & Sons, Inc. New York, NY, USA.

Rayo, L., 2007. Relational incentives and moral hazard in teams. The Review of Economic Studies, 74(3), p.937.

Sleet, C. & Yeltekin, S., 2003. On the computation of value correspondences. mimeo.

Appendix A: Computing Punishment Payoffs in Policy Elimination Algorithm

This appendix develops a quick algorithm to compute the punishment payoffs $v_i(x|\hat{\boldsymbol{A}})$ and punishment policies $\hat{\boldsymbol{a}}^i(\hat{\boldsymbol{A}})$ in Step 2 of the policy elimination algorithm described in Section 5.3. The problem cannot be reduced to a simple Markov

decision process, but we show that a nested policy iteration method exists that searches among possible candidate policies $\boldsymbol{a}^i$ in a monotone fashion.

We denote by

$$c_i(a,x,v) = \max_{\hat{a}_i \in A_i(x)} \{(1-\delta)\,(\pi_i(\hat{a}_i, a_{-i}|x)) + \delta E[v_i(x')|x, \hat{a}_i, a_{-i}]\}$$

player i's best-reply payoff given that in the current period in state x action profile a shall be played and continuation payoffs in the next period only depend on the realized state x' and are given by $v_i(x')$.

The following nested policy iteration algorithm yields an optimal punishment policy $\hat{\boldsymbol{a}}^i(\hat{\boldsymbol{A}})$.

Algorithm. *Nested policy iteration to find an optimal punishment policy*

0. *Set the round to* $r = 0$ *and start with some initial policy* $\boldsymbol{a}^0 \in \hat{\boldsymbol{A}}$
1. *Calculate player* i*'s punishment payoffs* $v_i(.|\boldsymbol{a}^r)$ *given punishment policy* $\boldsymbol{a}^r$ *by solving the corresponding Markov Decision Process*
2. *Let* $\boldsymbol{a}^{r+1}$ *be a policy that minimizes state by state player* i*'s best-reply payoff against action profile* $\boldsymbol{a}^{r+1}(x)$ *given continuation payoffs* $v_i(.|\boldsymbol{a}^r)$*, i.e.*

$$\boldsymbol{a}^{r+1}(x) \in \arg\min_{a \in \hat{\boldsymbol{A}}(x)} c_i(x, a, v_i(.|\boldsymbol{a}^r)) \tag{20}$$

3. *Stop if* $\boldsymbol{a}^r$ *itself solves step 3. Otherwise increment the round* r *and go back to step 2.*

In Step 2, we update the punishment policy by minimizing state-by-state the best reply payoffs $c_i(x, a, v_i(.|\boldsymbol{a}^r))$. This operation can be performed very quickly. The following result shows that this updating rule causes the punishment payoffs $v_i(.|\boldsymbol{a}^r)$ to monotonically decrease in every round.

Proposition 6. *Algorithm 2 always terminates in a finite number of periods yielding an optimal punishment policy* $\boldsymbol{a}^i(\hat{\boldsymbol{A}})$*. The punishment payoffs decrease in every round (except for the last round):*

$$v_i^s(x|\boldsymbol{a}^{r+1}) \leq v_i^s(x|\boldsymbol{a}^r) \text{ for all } x \in X \text{ and}$$
$$v_i^s(x|\boldsymbol{a}^{r+1}) < v_i^s(x|\boldsymbol{a}^r) \text{ for some } x \in X.$$

The proof heavily exploits monotonicity properties of the contraction mapping operator that is used to solve the Markov decision process in Step 2. In the examples we computed, the algorithm typically finds an optimal punishment policy by examining just a very small fraction of all possible policies. While one can construct examples in which the algorithm has to check every possible policy in $\hat{\boldsymbol{A}}$, the monotonicity results suggest that the algorithm typically stops after a few rounds.

Appendix B: Proofs

Proposition 1, Lemma 1 and Lemma 2 are straightforward and proofs are omitted.

Proof of Lemma 6: The first sentence follows from the fact that action constraints and payment constraints are weak inequalities and linear in U and v. The second sentence follows then directly from the Theorem of the Maximum.■

Proof of Theorem 1: Denote by $\bar{U}(x)$ the supremum of joint continuation payoffs across all public perfect continuation equilibria starting in state x. Similarly, denote by $\bar{v}_i(x)$ the infimum of player i's continuation payoffs across all PPE continuation payoffs starting in state x. Denote by $\bar{u}$ some payoff function that satisfies

$$\sum_{i=1}^{n} \bar{u}_i(x) = \bar{U}(x) \forall x, y$$

and

$$\bar{u}_i(x) \geq \bar{v}_i(x) \forall x, i$$

Let $\{\bar{U}(x,m)\}_{m=1}^{\infty}$ be a sequence of joint PPE continuation payoffs that converges to $\bar{U}(x)$. Let $\bar{\sigma}^e(x,m)$ be a PPE starting at the action stage in state x that implements the joint payoff $\bar{U}(x,m)$. Let $\bar{\alpha}^e(x,m)$ denote the first action profile that is played in $\bar{\sigma}^e(x,m)$. Let $\hat{p}^e(.|x,\bar{u},m)$ denote a payment function that solves (LP-e) for $\bar{\alpha}^e(x,m)$, $\bar{u}$ and $\bar{v}$. That such a payment function exists follows from Lemma 1. Since $\{\bar{\alpha}^e(x,m), \hat{p}^e(.|x,\bar{u},m)\}_m$ is an infinite sequence in a compact space, there must be a converging subsequence with a limit $(\bar{\boldsymbol{\alpha}}^e(x), \hat{p}^e(.|x,\bar{u}))$ where $\bar{\boldsymbol{\alpha}}^e(x) \in \mathcal{A}(x)$. All action, payment and budget constraints will be satisfied for this limit point, because these constraints are weak inequalities and continuous in actions and payments. It follows from the Theorem of the Maximum that $\hat{p}^e(.|x,\bar{u})$ solves (LP-e) for $\bar{\boldsymbol{\alpha}}^e(x)$ given $\bar{u}$ and $\bar{v}$; the corresponding value of (LP-e) satisfies

$$\hat{U}(x, \bar{\boldsymbol{\alpha}}^e(x), \bar{u}, \bar{v}) \geq \bar{U}(x) \tag{21}$$

In a similar fashion, we can show that for every player i and every state x there exists an action profile $\bar{\boldsymbol{\alpha}}^i(x) \in A(x)$ and a payment function $\hat{p}^i(.|x,\bar{u})$ that solves (LP-i) for $\bar{\boldsymbol{\alpha}}^i(x)$ given $\bar{u}$ and $\bar{v}$. The corresponding value of (LP-i) satisfies

$$\hat{v}_i(x, \bar{\boldsymbol{\alpha}}^i(x), \bar{u}, \bar{v}) \leq \bar{v}_i(x) \tag{22}$$

Let $\tilde{\sigma}$ be a simple strategy profile with action plan $(\bar{\boldsymbol{\alpha}}^k)_{k\in\mathcal{K}}$ as specified above and a payment plan $(\tilde{p}^k)_{k\in\mathcal{K}}$ that is specified below. Expected continuation payoffs at the action stage in state x and equilibrium phase of $\tilde{\sigma}$ are given by

$$\tilde{u}_i(x) \;=\; (1-\delta)\pi_i(\bar{\boldsymbol{\alpha}}^k, x) + \delta E[-(1-\delta)\tilde{p}_i^e(x', y, x) + \tilde{u}_i(x')|x, \bar{\boldsymbol{\alpha}}^e]) \tag{23}$$

The payments $\tilde{p}^e$ for the equilibrium phase shall be chosen such that $\tilde{u}_i$ and $\tilde{p}^e$ satisfy the transformation specified in Lemma 1 with respect to $\bar{u}$ and $\hat{p}^e$:

$$\tilde{p}_i^e(x', y, x) = \hat{p}_i^e(x', y|x, \bar{u}) + \frac{\tilde{u}_i(x') - \bar{u}_i(x')}{1-\delta} \tag{24}$$

Since $\tilde{u}_i$ is a linear function of the payments $\tilde{p}_i$, (24) describes a system of linear equations with as many as unknowns as equations, i.e. at least one solution exists. For the punishment phase of each player i, we similarly specify payments for each player $j = 1, ..., n$ by

$$\tilde{p}_j^i(x', y, x) = \hat{p}_j^i(x', y|x, \bar{u}) + \frac{\tilde{u}_j(x') - \bar{u}_j(x')}{1 - \delta} \tag{25}$$

With these payments, expected continuation payoffs of $\tilde{\sigma}$ after the action stage in state x and phase k are then the same as if continuation payoffs were $\bar{u}$ and payments were $\hat{p}^k$, i.e. we have

$$\tilde{U}(x) = \hat{U}(x, \bar{\boldsymbol{\alpha}}^e(x), \bar{U}, \bar{v}) \tag{26}$$

and

$$\tilde{v}_i(x) = \hat{v}_i(x, \bar{\boldsymbol{\alpha}}^i(x), \bar{u}, \bar{v}) \tag{27}$$

which implies that $\tilde{U}(x) \geq \bar{U}(x)$ and $\tilde{v}_i(x) \leq \bar{v}_i(x)$. It follows from Lemma 1 that all action, payment and budget constraints of $\tilde{\sigma}$ are satisfied, i.e. $\tilde{\sigma}$ constitutes a simple equilibrium. We thus must have

$$\tilde{U}(x) = \bar{U}(x) = \hat{U}(x, \bar{\boldsymbol{\alpha}}^e(x), \bar{U}, \bar{v})$$

and

$$\tilde{v}_i(x) = \hat{v}_i(x, \bar{\boldsymbol{\alpha}}^i(x), \bar{u}, \bar{v}) = \bar{v}_i(x)$$

It then follows from Proposition 1 that every PPE payoff can be implemented by varying the upfront transfers of $\tilde{\sigma}$ and that the payoff set is compact.■

Proof of Proposition 2: The set of payment plans that satisfy conditions (AC-k), (PC-k) and (BC-k) is compact, i.e. for every state x there exists a maximum level $\bar{U}^s(x|(\boldsymbol{\alpha}^k)_{k\in\mathcal{K}})$ of joint equilibrium phase payoffs and a minimum level $\bar{v}_i^s(x|(\boldsymbol{\alpha}^k)_{k\in\mathcal{K}})$ of player i's punishment payoffs that can be implemented with simple equilibria with payment plan $(\boldsymbol{\alpha}^k)_k$. The proof now proceeds in similar steps than the proof for Theorem 1. One can show that there always exists a payment plan that creates a simple equilibrium that has joint payoffs of $\bar{U}^s(x|(\boldsymbol{\alpha}^k)_{k\in\mathcal{K}})$ and punishment payoffs of $\bar{v}_i^s(x|(\boldsymbol{\alpha}^k)_{k\in\mathcal{K}})$ for all players at the same time in all states. By maximizing $\sum_{x\in X}\sum_{i=1}^n (u_i^s(x) - v_i^s(x))$, we select a payment plan that solves this problem.■

Proof of Proposition 3:The result follows directly from the monotonicity results in Lemma 5 and from the continuity results in Lemma 6.■

Proof of Proposition 4: We first show sufficiency. Let $\hat{p}^k(.|u, x)$ be the payment function that solves (LP-k) given v and some u satisfying $\sum u_i(x) = U(x)$ for all x. We now choose a payment plan $(p^k)_{k\in\mathcal{K}}$ such that a simple strategy profile with $(p^k)_{k\in\mathcal{K}}$ and action plan $(\boldsymbol{\alpha}^k)_{k\in\mathcal{K}}$ has equilibrium phase and punishment payoffs that satisfy $u_i^s(x) = u_i(x)$ and $v_i^s(x) = v_i(x)$. Such a payment plan is

given by a solution to the following system of linear equations that satisfies the transformation in Lemma 1

$$p_i^k(x', y, x) = \hat{p}_i^k(x', y|x, u) + \frac{u_i^s(x') - u_i(x')}{1-\delta} \forall i, x', x, y, k$$

It follows from Lemma 1 that $(p^k)_{k\in\mathcal{K}}$ and $(\alpha^k)_{k\in\mathcal{K}}$ form a simple equilibrium.

Necessity is straightforward. If there exists a simple equilibrium with action plan $(\boldsymbol{\alpha}^k)_{k\in\mathcal{K}}$ then (LP-OPP) must have a solution. If U and v are the payoffs corresponding to that solution, conditions (5) and (6) obviously hold true because the separate problem (LP-e) and (LP-i) impose fewer constraints than (LP-OPP). ■

Proof of Proposition 5:

1. To prove the first result, we show that if there is a solution to (LP-e) then there is a solution without money burning. Consider the punishment phase i and the case that no player has deviated in the previous action stage. If a player $j \neq i$ burns money, she can instead simply not burn it: not burning, relaxes j's action constraints while not changing i's payoff. If player i burns money, she can instead simply transfer the same amount to player j. Both arguments also work the equilibrium phase. If some player i is asked to burn money after she deviated in the previous period, she can simply transfer the same amount to another player $j \neq i$. Since for j the deviation of i is a zero probability event, any such transfers don't change j's action constraints.

2. We now prove the second result. Throughout the remaining proof, we assume that u is the payoff vector with $\sum_{i=1}^n u_i = U$ and $u_i \geq v_i \forall i$ that is used as argument in problem (LP-k). We denote by

$$p_j^{max}(x) = \frac{1}{1-\delta}(u_j(x) - v_j(x))$$

the maximum payment that satisfies player j's payment constraint in state x. Since deviations are perfectly observed, player i's incentives to deviate from an action profile are minimized if whenever she deviates and state x' realizes she has to pay $p_i^{max}(x')$. Given such maximum fines, player i's action constraint in state x and for problem $(LP-k)$ simplifies to

$$(1-\delta)\left(\pi_i(\boldsymbol{a}^k, x) - \delta E[p_i^k|\boldsymbol{a}^k, x]\right) + \delta E[u|\boldsymbol{a}^k, x] \geq \max_{a_i \in A_i(x)} \left((1-\delta)\pi_i(a_i, \boldsymbol{a}_{-i}^k, x) + \delta E[v_i(x')|a_i, \boldsymbol{a}_{-i}^k, x]\right) \quad (28)$$

We now note that a player's action constraints are weakly stricter than her payment constraints in the following sense. If payments satisfy $p_j(x') = p_j^{max}(x')$ for all resulting states $x' \in X$ then the action constraint (28) becomes

$$(1-\delta)\pi_i(a^k, x) + \delta E[v_i(x')|a^k, x] \geq \max_{a_i \in A_i(x)} \left((1-\delta)\pi_i(a_i, a_{-i}^k, x) + \delta E[v_i(x')|a_i, a_{-i}^k, x]\right) \quad (29)$$

Obviously, condition (29) either binds exactly or is violated. In problem (LP-i) it is optimal to choose expected payments for player i as high as possible given payment constraints and action constraints. Optimal payments for player i will thus always be such that player $i's$ action constraints are exactly binding; continuation payoffs are thus given by (8).

3. Condition (9) is simply the sum of players' action constraints (28) given maximum fines. Hence, (9) is necessary for the existence of solution to (LP-k).

We now show sufficiency. Obviously, for problem (LP-k) there always exist payments such for all but one freely selected player $j \neq k$ all action and payment constraints are satisfied. Furthermore, since action constraints are weakly stricter than payment constraints, we also know that such payments exist under which the action constraints of all players $i \neq j$ are exactly binding. If no money burning is used, expected payments of player j in state x satisfy:

$$\begin{aligned} -(1-\delta)\delta E[p_j(x')|a^k,x] \quad &= \\ (1-\delta)\left(\Pi(a^k,x)-\pi_j(a^k,x)\right)+\delta(U(x)-u_j(x)) \\ -\sum_{i\neq j}\max_{\hat{a}_i\in A_i(x)}\{(1-\delta)\pi_i(\hat{a}_i,a^k_{-i},x)+\delta E[v_i(x')|\hat{a}_i,a^k_{-i},x]\} \end{aligned} \tag{30}$$

Plugging in these expected transfers into the action constraint for player j for state x yields condition (9). This means (9) implies that there exist payments without money burning that satisfy the action constraints for all players in all states and phases and all payment constraints of all players $i \neq j$. We now show that there also exists payments that additionally satisfy the payment constraints for player j. If no money burning is used, the sum of payment constraints across all players is always satisfied, since

$$\sum_{i=1}^{n} p_i(x') = 0 \leq \frac{1}{1-\delta}\sum_{i=1}^{n}(u_i(x')-v_i(x')) = \sum_{i=1}^{n} p_i^{max}(x'). \tag{31}$$

Assume that given payments satisfy all action constraints and all payment constraints except for the payment constraints for player j in some realized state x'. (31) then guarantees that there is a set of players J such that the transfers $p_j(x')$ of players $j \in J$ can be sufficiently increased and transferred to player i so that the payment constraints of all players $\{i\} \cup J$ are satisfied. Furthermore, our previous result that for every player i action constraints are stricter than the payment constraints, implies that there is a set of realized states $\hat{X} \subset X$ such that player i's transfers $p_i(\hat{x}')$ for all $\hat{x}'$ can be increased and given to players $j \in J$ such that with this compensation expected transfers $E[p_j(x')|a^k,x]$ do not change for any player $j \in J \cup \{i\}$.■

Proof of Proposition 6: Let $C_i^{\boldsymbol{a}}$ be a operator mapping the set of punishment payoffs in itself defined by

$$C_i^{\boldsymbol{a}}(v_i)[x] = c_i(x,\boldsymbol{a}(x),v_i)$$

It can be easily verified that $C_i^{\boldsymbol{a}}$ is a contraction-mapping operator. It follows from the contraction-mapping theorem that player i's best-reply payoffs are given by the unique fixed point of $C_i^{\boldsymbol{a}}$, which we denote by $v_i(\boldsymbol{a})$. This means

$$v_i(\boldsymbol{a}) = C_i^{\boldsymbol{a}}(v_i(\boldsymbol{a})) \tag{32}$$

It is a well known result that the operator $C_i^{\boldsymbol{a}}$ is monotone. This means

$$v_i \leq \tilde{v}_i \Rightarrow C_i^{\boldsymbol{a}}(v_i) \leq C_i^{\boldsymbol{a}}(\tilde{v}_i) \tag{33}$$

where $v_i \leq \tilde{v}_i$ is defined as $v_i(x) \leq \tilde{v}_i(x) \forall x \in X$. We denote by $[C_i^{\boldsymbol{a}}]^k$ the operator that applies k times $C_i^{\boldsymbol{a}}$ and define its limit by

$$[C_i^{\boldsymbol{a}}]^\infty = \lim_{k\to\infty} [C_i^{\boldsymbol{a}}]^k.$$

The contraction mapping theorem implies that $[C_i^{\boldsymbol{a}}]^\infty$ is well defined and transforms every payoff function v into the fixed point of $C_i^{\boldsymbol{a}}$, i.e.

$$[C_i^{\boldsymbol{a}}]^\infty(v) = v(\boldsymbol{a}) \tag{34}$$

Furthermore, it follows from monotonicity of $C_i^{\boldsymbol{a}}$ that

$$C_i^{\boldsymbol{a}}(v_i) \leq v_i \Rightarrow [C_i^{\boldsymbol{a}}]^\infty(v_i) \leq v_i \tag{35}$$

and

$$C_i^{\boldsymbol{a}}(v_i) < v_i \Rightarrow [C_i^{\boldsymbol{a}}]^\infty(v_i) < v_i \tag{36}$$

where two payoff functions u_i and $\tilde{u}_i$ satisfy $u_i < \tilde{u}_i$ if $u_i \leq \tilde{u}_i$ and $u_i \neq \tilde{u}_i$.

We now show that for any two policies $\boldsymbol{a}$ and $\tilde{\boldsymbol{a}}$ the following monotonicity results hold

$$\begin{aligned}
C_i^{\boldsymbol{a}}(v(\boldsymbol{a})) = C_i^{\tilde{\boldsymbol{a}}}(v(\boldsymbol{a})) &\Rightarrow v(\boldsymbol{a}) = v(\tilde{\boldsymbol{a}}) && (37)\\
C_i^{\boldsymbol{a}}(v(\boldsymbol{a})) > C_i^{\tilde{\boldsymbol{a}}}(v(\boldsymbol{a})) &\Rightarrow v(\boldsymbol{a}) > v(\tilde{\boldsymbol{a}}) && (38)\\
v(\boldsymbol{a}) \not\leq v(\tilde{\boldsymbol{a}}) &\Rightarrow C_i^{\boldsymbol{a}}(v(\boldsymbol{a})) \not\leq C_i^{\tilde{\boldsymbol{a}}}(v(\boldsymbol{a})) && (39)
\end{aligned}$$

We exemplify the proof for (38). It follows in this order from from (32), the left part of (38), (35) and (34) that

$$v(\boldsymbol{a}) = C_i^{\boldsymbol{a}}(v(\boldsymbol{a})) > C_i^{\tilde{\boldsymbol{a}}}(v(\boldsymbol{a})) \geq \left(C_i^{\tilde{\boldsymbol{a}}}\right)^\infty (v(\boldsymbol{a})) = v(\tilde{\boldsymbol{a}}).$$

(37) and can be proven similarly. To prove (39), assume that there is some $\tilde{\boldsymbol{a}}$ with $C_i^{\boldsymbol{a}}(v) \leq C_i^{\tilde{\boldsymbol{a}}}(v)$ but $\tilde{v} \not\geq v$. We find

$$v = C_i^{\boldsymbol{a}}(v) \leq C_i^{\tilde{\boldsymbol{a}}}(v) \leq \left(C_i^{\tilde{\boldsymbol{a}}}\right)^\infty (v) = \tilde{v}$$

which contradicts the assumption $\tilde{v} \not\geq v$.

Intuitively, these monotonicity properties of the cheating payoff operator are crucial for why the algorithm works. If one wants to find out whether a policy $\tilde{\boldsymbol{a}}$ can yield lower punishment payoffs for player i than a policy $\boldsymbol{a}$, one does not have to solve player i's Markov decision process under policy $\tilde{\boldsymbol{a}}$. It suffices to check whether for some state x the cheating payoffs given policy $\tilde{\boldsymbol{a}}$ and punishment payoffs $v(\boldsymbol{a})$ are lower than $v(\boldsymbol{a})(x)$. If this is not the case for any admissible policy $\tilde{\boldsymbol{a}}$ then a policy $\boldsymbol{a}$ is an optimal punishment policy, in the sense that it minimizes player i's punishment payoffs in every state.

The fixed point condition (32) of the value determination step and the policy improvement step (20) imply that $v^r = C_i^{\boldsymbol{a}^r}(v^r) \geq C_i^{\boldsymbol{a}^{r+1}}(v^r)$. We first establish that if

$$v^r = C_i^{\boldsymbol{a}^r}(v^r) = C_i^{\boldsymbol{a}^{r+1}}(v^r). \tag{40}$$

then we have $v_i^r = \hat{v}_i$. For a proof by contradiction, assume that condition holds for some r but that there exists a policy $\hat{\boldsymbol{a}}$ such that $v(\boldsymbol{a}^r) \not\leq v(\hat{\boldsymbol{a}})$, i.e. $\hat{\boldsymbol{a}}$ leads in at least some state x to a strictly lower best-reply payoff for player i than $\boldsymbol{a}^r$. By (39) this would imply $C_i^{\boldsymbol{a}^r}(v^r) \not< C_i^{\tilde{\boldsymbol{a}}}(v^r)$. This means that $\hat{\boldsymbol{a}}$ must also be a solution to the policy improvement step and since (40) holds, we then must have

$$C_i^{\boldsymbol{a}^r}(v^r) = C_i^{\hat{\boldsymbol{a}}}(v^r)$$

However, (37) then implies that $v(\boldsymbol{a}^r) = v(\hat{\boldsymbol{a}})$, which contradicts the assumption $v(\boldsymbol{a}^r) \not\leq v(\hat{\boldsymbol{a}})$. Thus if the algorithm stops in a round R, we indeed have $v^R = \hat{v}_i$.

If the algorithm does not stop in round r, it must be the case that $v^r = C_i^{\boldsymbol{a}^r}(v^r) > C_i^{\boldsymbol{a}^{r+1}}(v^r)$. (38) then directly implies the monotonicity result $v^r > v^{r+1}$. The algorithm always stops in a finite number of rounds since the number of policies is finite and there a no cycles because of the monotonicity result.■