Graduate School of Decision Sciences

Universität Konstanz

GSDS Working Paper No. 2015-04

Discounted Stochastic Games with Voluntary Transfers

Susanne Goldlücke | Sebastian Kranz

November 2015

Graduate School of Decision Sciences

All processes within our society are based on decisions – whether they are individual or collective decisions. Understanding how these decisions are made will provide the tools with which we can address the root causes of social science issues.

The GSDS offers an open and communicative academic environment for doctoral researchers who deal with issues of decision making and their application to important social science problems. It combines the perspectives of the various social science disciplines for a comprehensive understanding of human decision behavior and its economic and political consequences.

The GSDS primarily focuses on economics, political science and psychology, but also encompasses the complementary disciplines computer science, sociology and statistics. The GSDS is structured around four interdisciplinary research areas: (A) Behavioural Decision Making, (B) Intertemporal Choice and Markets, (C) Political Decisions and Institutions and (D) Information Processing and Statistical Analysis.

GSDS – Graduate School of Decision Sciences
University of Konstanz
Box 146
78457 Konstanz

Phone: +49 (0)7531 88 3633
Fax: +49 (0)7531 88 5193
E-mail: gsds.office@uni-konstanz.de

-gsds.uni-konstanz.de

ISSN: 2365-4120
November 2015

Discounted Stochastic Games with Voluntary Transfers*

Susanne Goldlücke† and Sebastian Kranz‡

September 2015

Abstract

This paper studies discounted stochastic games with perfect or imperfect public monitoring and the opportunity to conduct voluntary monetary transfers. This generalization of repeated games with transfers is ideally suited to study relational contracting in applications that allow for long-term investments, and also allows to study collusive industry dynamics. 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: dynamic games, relational contracting, monetary transfers, computation, imperfect public monitoring, public perfect equilibria

*Support by the German Research Foundation (DFG) through SFB-TR 15 for both authors and an individual research grant for the second author is gratefully acknowledged. Sebastian Kranz would like to thank the Cowles Foundation in Yale, where part of this work was conducted, for the stimulating research environment. Further thanks go to 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.

†Department of Economics, University of Konstanz. Email: susanne.goldluecke@uni-konstanz.de.

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

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 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 particular in the relational contracting literature. Examples include studies of employment relations by Malcomson and MacLeod (1989) and Levin (2002, 2003), 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 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. Despite the wide range of applications, repeated games are nevertheless considerably limited, because they cannot account for actions that have technological long run effects, like e.g. investment decisions.

This paper extends these results to stochastic games with voluntary transfers and imperfect monitoring of actions. For any given discount factor, all public perfect equilibrium (PPE) payoffs can be implemented with a class of simple 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 regime and for each player a punishment regime. The action profile that is played in the equilibrium regime only depends on the current state, as in a stationary Markov perfect equilibrium. Transfers depend on the current state and signal and also on the previous state. Play moves to a punishment regime whenever a player refuses to make a required transfer. Punishments have a simple stick-and-carrot structure: one punishment

[1] Baliga and Evans (2000), Fong and Surti (2009), Gjertsen et. al (2010), Miller and Watson (2011), and Goldlücke and Kranz (2013) study renegotiation-proof equilibria in repeated games with transfers.

action profile per player and state is defined. After the punishment profile has been played and subsequently required transfers are conducted, play moves back to the equilibrium regime. We show that there exists an optimal simple equilibrium, with largest joint equilibrium payoff and harshest punishments, such that all PPE payoffs can be implemented by varying the up-front payments of this equilibrium.

Repeated games have a special structure, in which the current action profile does not affect the set of continuation payoffs. This means that the harshest punishment that can be imposed on a deviating player is independent of the form of a deviation. For repeated games with transfers, this fact allows to compress all information of the continuation payoff set that is relevant to determine whether and how an action profile can be used, into a single number (Goldlücke and Kranz, 2012). In stochastic games, complications arise because different deviations can cause different state transitions. An optimal deviation is a dynamic problem and optimal punishment schemes must account for this. As a consequence, key results of the analysis of repeated games with transfers no longer apply, and different algorithms are needed.

For stochastic games with perfect monitoring and finite action spaces, we develop in Section 4 a fast algorithm to exactly compute the set of pure strategy subgame perfect equilibrium payoffs. To find the action profiles and transfers of the equilibrium regime we iteratively solve a single agent Markov decision problem. In each iteration the set of possible action profiles that can be played in equilibrium can be reduced. A key element is a fast method to find in each iteration the optimal punishment policies: it quickly solves the nested dynamic optimization problem of finding for a given punishment policy the optimal deviations in an inner loop and the corresponding optimal punishment policy in an outer loop.

To solve stochastic games with imperfect public monitoring, we develop methods that are more closely related to the methods by Judd, Yeltekin and Conklin (2003) and Abreu and Sannikov (2014), that were developed to approximate the payoff set of repeated games with perfect monitoring and public correlation.[2] They are based on the recursive techniques developed by Abreu, Pearce and Stacchetti (1990, henceforth APS) for repeated games. Our methods, developed in Section 5, allow to compute arbitrary fine inner and outer approximations of the PPE payoff set. 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 quickly reduced, it may even become tractable to then apply a brute force method, which solves a linear optimization problem for every combination of remaining action profiles, to exactly characterize optimal equilibria and the PPE payoff set.

Our characterization with simple equilibria not only allows numerical solution methods, but also helps to find closed form solutions in stochastic games. Section 6 illustrates this with two relational contracting examples. In the first example,

[2]Judd and Yeltekin (2011) and Sleet and Yeltekin (2015) extend these methods to approximate equilibrium payoff sets in stochastic games with perfect monitoring and public correlation.

an agent can exert effort to produce a durable good for a principal. It is illustrated how under unobservable effort levels, grim-trigger punishments completely fail to induce positive effort for any discount factor while optimal punishments that use a costly punishment technology can sustain positive effort levels. In the second example, an agent can invest to increase the value of his outside option. It illustrates how the set of equilibrium payoffs can be non-monotonic in the discount factor.

While the relational contracting literature on repeated games usually focuses on efficient SPE or PPE, applied industrial organization literature that studies stochastic games often restricts attention to Markov perfect equilibria (MPE) in which actions only condition on the current state.[3] Focusing on MPE has advantages, since strategies have a simple structure and there exist quick algorithms to find a MPE. Finding optimal collusive SPE or PPE payoffs is usually a much more complex task.[4]

However, there are also drawbacks of restricting attention to MPE. One issue is that the set of MPE payoffs can be very sensitive to the definition of the state space. For example, in the special case of a repeated game (a stochastic game with a single state), only stage game Nash equilibria can be played in an MPE. If the state-space of the repeated game is augmented by defining the current state to be the previous period's action profile as state, collusive strategies may now be supported as MPE. In contrast, the set of SPE payoffs (under perfect monitoring) is not changed by such an technological irrelevant augmentation of the state space. Another issue is that there are no effective algorithms to compute all MPE payoffs of stochastic game, even if one just considers pure strategies.[5] Existing algorithms, e.g. Pakes & McGuire (1994, 2001), are very effective in finding an 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. There is, however, still no guarantee 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 seem quite useful in order to complement the analysis of MPE.

While monetary transfers may not be feasible in all social interactions, the possi-

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

[4]Characterizing the SPE or PPE payoff set can be challenging even in the limit case of the discount factor converging towards 1. While by Dutta (1995) established a folk theorem for perfect monitoring, folk theorems for imperfect public monitoring have been derived much more recently by Fudenberg and Yamamoto (2010) and Hörner et. al. (2011) and with restriction to irreducible stochastic games.

[5]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.

bility of transfers is plausible in many problems of economic interest. Monetary transfers are a standard assumption in the already mentioned literature on relational contracting, even though attention has been usually restricted to repeated games. But 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. Even in contexts in which transfers may be considered strong 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.

The structure of this paper is as follows. Section 2 describes the model. In Section 3, simple equilibria are defined and it is shown that every PPE can be implemented with an optimal simple equilibrium. Section 4 develops an exact policy elimination algorithm for games with perfect monitoring. We illustrate the algorithm by numerically characterizing optimal collusive equilibria in a Cournot model with renewable, storable resources. We have implemented the policy elimination algorithm for stochastic games with perfect monitoring in the open source R package *dyngame*. Installation instructions are available on its Github page: https://github.com/skranz/dyngame. Section 5 highlights the links with the recursive structure of APS and we describe decomposition methods for our setting that allow to approximate the PPE payoff set for games with imperfect public monitoring. Finally, Section 6 studies relational contracting examples and shows how the methods allow closed-form analytical characterizations. The appendix contains remaining proofs.

2 The game

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, with $x_0 \in X$ denoting the initial state. A period is comprised of two stages: a transfer stage and an action stage without discounting between stages.

In the transfer stage, every player simultaneously chooses a non-negative vector of transfers to all other players.[6] Players also have the option to transfer money to a non-involved third party, which has the same effect as burning money. All

[6]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. This bound is large enough to be never binding given the incentive constraints of voluntary transfers.

transfers are perfectly monitored, there is no limited liability, and transfers do not affect the state transitions.

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

After actions have been taken, 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. We denote by $q(y, x'|x, a)$ the probability that signal y and state x' are drawn, depending on the current state x and the chosen action profile a. Player i's stage game payoff is denoted by $\hat{\pi}_i(x, a_i, y)$ and depends only on what is observable to this player: the signal y, the player's own action a_i, and the current state x. We denote by $\pi_i(x, a)$ player i's expected stage game payoff in state x if action profile a is played. If the action space in state x is not finite, we assume in addition that stage game payoffs and the probability distribution of signals and new states are 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 in which the state is x, action profile a is played, and $i's$ net transfer is given by p_i, is equal to $\pi_i(x, a) - p_i$.

For the case of a finite stage game we also consider behavior strategies and let $\mathcal{A}(x)$ denote the set of mixed action profiles at the action stage in state x. If the game is not finite, we restrict attention to pure strategy equilibria and let $\mathcal{A}(x) = A(x)$ denote the set of pure action profiles. For a mixed action profile $\alpha \in \mathcal{A}(x)$, we denote by $\pi_i(x, \alpha)$ player i's expected stage game payoff, taking expectations over mixing probabilities and signal realizations. A vector α that assigns an action profile $\alpha(x) \in \mathcal{A}(x)$ to every state $x \in X$ is also called a policy, and $\mathcal{A} = \times_{x \in X} \mathcal{A}(x)$ denotes the set of all policies.[7] For briefness sake, we often suppress the dependence on x and write $\pi(x, \alpha)$ instead of $\pi(x, \alpha(x))$. Moreover, we often use capital letters to denote the joint payoff of all players, e.g.

$$\Pi(x, \alpha) = \sum_{i=1}^{n} \pi_i(x, \alpha). \tag{1}$$

When referring to payoffs of the stochastic game, we mean expected average discounted payoffs, i.e., the discounted sum of expected payoffs multiplied by $(1-\delta)$.

A public history of the stochastic game is a sequence of all states, monetary transfers and public signals 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 in period t into a possibly mixed action in $\mathcal{A}_i(x_t)$, and every public history that ends before a payment stage into a vector of monetary transfers. A profile of public strategies for each player determines a probability

[7] Whether α denotes a single action profile or a whole policy depends on the context.

distribution over the outcomes of the game. Expected payoffs from a strategy profile σ are denoted by

$$u_i(x_0, \sigma) = (1-\delta) \sum_{t=0}^{\infty} \delta^t E_{x_0,\sigma}[\pi_i(x_t, \alpha_t) - p_{t,i}]. \tag{2}$$

A public perfect equilibrium (PPE) is a profile of public strategies that constitute mutual best replies after every public history. We restrict attention to public perfect equilibria. We denote by $\mathcal{U}(x_0)$ the set of PPE payoffs with initial state x_0, and by $\mathcal{U}^0(x_0)$ the set of payoffs of PPE without up-front transfers. These sets depend on the discount factor, but since the discount factor is fixed, we do not make this dependence explicit.

We show in Section 5 how the recursive methods of APS can be translated to this stochastic game with monetary transfers. Following the steps of APS, one can show the following compactness result, which we already state here to simplify the subsequent discussion.

Proposition 1. *The set $\mathcal{U}(x_0)$ of PPE payoffs in our discounted stochastic game with monetary transfers is compact.*

Proof. Follows directly from Lemma 2 in Section 5. □

3 Characterization with simple equilibria

This section first defines simple strategy profiles and characterizes PPE in simple strategies. To convey the intuition behind our results, it is explained in what ways monetary transfers simplify the analysis. First, up-front transfers in the first period allow the players to flexibly distribute the total equilibrium payoff. Similarly, variation in transfers can be used in every period to substitute for variation in continuation payoffs. This intuition is used to show that simple equilibria suffice to describe the PPE payoff set. Second, transfers can balance incentive constraints between players in asymmetric situations and third, payment of fines allows to settle punishments within one period.

3.1 Simple strategy profiles

A simple strategy profile is characterized by $n+2$ regimes. Play starts in the up-front transfer regime, in which players are required to make up-front transfers described by net payments p^0.[8] Afterward, play can be in one of $n+1$ regimes,

[8] In a simple strategy profile, no player makes and receives positive transfers at the same time. Any vector of net payments p can be mapped into a $n \times (n+1)$-matrix of gross transfers $\tilde{p}_{ij}$ (= payment from i to j) as follows. Denote by $I_P = \{i | p_i > 0\}$ the set of net payers and by

which we index by $k \in \mathcal{K} = \{e, 1, 2, ..., n\}$. We call the regime $k = e$ the equilibrium regime and $k = i \in \{1, ..., n\}$ the punishment regime of player i.

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

The equilibrium and punishment regimes follow the logic of Abreu (1988), exploiting that transfers are perfectly monitored so that any deviation from a transfer can be punished in the same way. If no player unilaterally deviates from a required transfer, play moves to the equilibrium regime ($k = e$). If player i unilaterally deviates from a required transfer, play moves to the punishment regime of player i ($k = i$). In all other situations the regime does not change. A simple equilibrium is a simple strategy profile that constitutes a public perfect equilibrium of the stochastic game.

For a given simple strategy profile, we denote expected continuation payoffs in the equilibrium regime and the punishment regime by u^e and u^i, respectively. For all $k \in \mathcal{K}$ and each player i, these payoffs are given by[9]

$$u_i^k(x) = (1-\delta)\pi_i(x, \alpha^k) + \delta E[-(1-\delta)p_i^k(x, y, x') + u_i^e(x')|x, \alpha^k]. \tag{3}$$

We call $U(x) = \sum_{i=1}^n u_i^e(x)$ the joint equilibrium payoff and $v_i(x) = u_i^i(x)$ the punishment payoff of player i.

We use the one-shot deviation property to establish equilibrium conditions for simple strategies without up-front transfers. In state x, player i has no profitable one-shot deviation from any pure action a_i in the support of $\alpha_i^k(x)$ if and only if the following *action constraints* are satisfied for all $\hat{a}_i \in A_i(x)$:

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

$I_R = \{i|p_i \leq 0\} \cup \{0\}$ 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

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

[9]For $k = e$, the payoff u_i^e is defined implicitly by this equation, which has a unique solution.

Moreover, player i should have no incentive to deviate from required payments after the action stage. Hence we need for all regimes $k \in \mathcal{K}$, states x, x' and signals y that the following *payment constraints* hold:

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

Finally, the *budget constraints* must hold that require that the sum of payments is non-negative:

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

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

3.2 Distributing with up-front transfers

The effect of introducing up-front transfers is illustrated in Figure 1. Assume

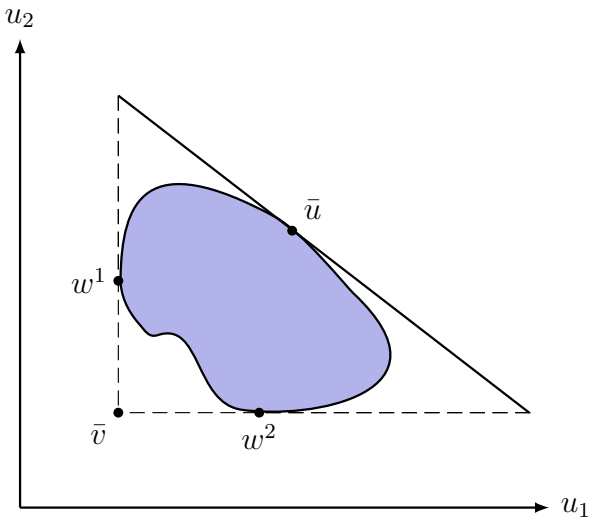

Figure 1: Distributing with up-front transfers

the shaded area is the PPE payoff set in a two player stochastic game with fixed discount factor without up-front transfers. The point $\bar{u}$ is the equilibrium payoff with the highest sum of payoffs for both players. If one could impose any up-front transfer, the set of Pareto optimal payoffs would be simply given by a line with slope -1 through this point. If up-front 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 points w^1 and w^2 in Figure 1 illustrate these worst continuation payoffs after the first transfer stage for each player, with $\bar{v}_i$ denoting the worst payoff of player i. The Pareto frontier of PPE payoffs with voluntary up-front transfers is given by the line segment through point $\bar{u}$ with slope -1 that is bounded by the lowest equilibrium payoff

$\bar{v}_1$ of player 1 and the lowest equilibrium payoff $\bar{v}_2$ of player 2. If we allow for money burning in the up-front transfers, any point in the depicted triangle can be implemented in an incentive compatible way.

This intuition naturally extends to n player games. We denote by

$$\bar{U}(x_0) = \max_{u \in \mathcal{U}(x_0)} \sum_{i=1}^{n} u_i \tag{4}$$

the maximum over joint PPE payoffs, and by

$$\bar{v}_i(x_0) = \min_{u \in \mathcal{U}(x_0)} u_i \tag{5}$$

the minimum over all possible PPE payoffs of player $i = 1, ..., n$. Note that these values would be the same if only PPE without up-front transfers, i.e, only payoff vectors in $\mathcal{U}^0(x_0)$, were considered instead.

Proposition 2. *The set of PPE payoffs is equal to the simplex*

$$\mathcal{U}(x_0) = \{u \in \mathbb{R}^n | \sum_{i=1}^{n} u_i \leq \bar{U}(x_0) \text{ and } u_i \geq \bar{v}_i(x_0)\}.$$

3.3 Optimal simple equilibria can implement all PPE payoffs

We now show that every PPE payoff can be implemented with a simple equilibrium. Assume that for all initial states a PPE exists. Since the set of PPE payoffs is compact for each initial state x, we can take the PPE $\sigma^e(x)$ with the largest total payoff $\bar{U}(x)$, and the PPE $\sigma^i(x)$ with the lowest possible payoff $\bar{v}_i(x)$ for player i among all PPE without up-front transfers. For all $k \in \mathcal{K}$, we define $\alpha^k(x)$ as the action profile that is played in the first period of $\sigma^k(x)$, and $w^k(x)(y, x')$ as the continuation payoffs in the second period when the realized signal in the first period is y and the game transits to state x'. We denote the equilibrium payoffs of $\sigma^k(x)$ in the game without up-front transfers by

$$\bar{u}_i^k(x) = (1-\delta)\pi_i(x, \alpha^k) + \delta E[w_i^k(x)(x', y)|x, \alpha^k]. \tag{6}$$

Then $w^k(x)$ enforces $\alpha^k(x)$, meaning that for all $a_i, \hat{a}_i \in A_i(x)$ with $\alpha_i(a_i) > 0$ it holds that

$$(1-\delta)\pi_i(x, a_i, \alpha_{-i}^k) + \delta E[w_i^k(x)|x, a_i, \alpha_{-i}^k)] \geq (1-\delta)\pi_i(x, \hat{a}_i, \alpha_{-i}^k) + \delta E[w_i^k(x)|x, \hat{a}_i, \alpha_{-i}^k]. \tag{7}$$

The vector of policies $(\alpha^k)_{k \in K}$ will be the action plan for the simple strategy profile that we are going to define. We define the payments in state x' following signal y and previous state x such that we achieve the continuation payoffs that enforce $\alpha^k(x)$. Hence, we define payments $p^k(x, y, x')$ such that

$$w^k(x)(x', y) = \bar{u}^e(x') - (1-\delta)p^k(x, y, x'). \tag{8}$$

It is straightforward to verify that the so defined simple strategy profile is indeed a PPE: Since continuation payoffs u_i^k in the simple strategy profile are equal to the payoffs $\bar{u}_i^k$ in the original equilibria, the action constraints (AC-k) are satisfied for all $k \in \mathcal{K}$. The payments in the payment plan are incentive compatible because player i at least weakly prefers the continuation payoff $w^k(x)(x', y)$ to $\bar{v}_i(x')$. Moreover, the sum of payments is non-negative since

$$\bar{U}(x') \geq \sum_{i=1}^{n} w_i^k(x)(x', y). \tag{9}$$

Hence, (PC-k) and (BC-k) are satisfied as well and we have shown the following result.

Theorem 1. *Assume a PPE exists. Then an optimal simple equilibrium exists such that by varying its up-front transfers in an incentive compatible way, every PPE payoff can be implemented.*

The goal of the following two subsections is to provide some easier intuition for why and how monetary transfers allow to restrict attention to simple equilibria.

3.4 Intuition: Stationarity on equilibrium path by balancing incentive constraints

A crucial factor why action profiles on the equilibrium path can be stationary (only depending on the state x) is that monetary transfers allow to balance incentive constraints among players. We want to illustrate this point with a simply 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 a 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.

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. Player 1 has no incentive to deviate from the transfers on the equilibrium path if and only if[10]

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

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

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

we find the same critical discount factor as for the individual constraints.

This intuition generalizes to stochastic games. Section 4 illustrates the incentive constraints with optimal balancing of payments for the case of perfect monitoring.

[10] 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.

3.5 Intuition: Settlement of punishments in one period

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 taken as a sign of a 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, such punishments arise on the equilibrium path following signals that indicate a deviation, and thus efficiency losses must be as low as possible in Pareto optimal equilibria.

The benefits of transfers for simplifying optimal punishments are easiest seen for the case of punishing an observable deviation. 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 play. Punishment actions only have to 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 punishments with a single punishment action profile per player and state.

Despite this simplification, an optimal punishment policy must consider all states and take into account the dynamic nature of a punished player's best reply. The nature of this nested dynamic problem can be seen most clearly in the perfect monitoring case in Section 4, which develops a fast method to find optimal punishments policies.

3.6 A brute force algorithm to find an optimal simple equilibrium

We have shown in Subsection 3.1. that a simple equilibrium with action plan $(\alpha^k)_{k\in\mathcal{K}}$ exists if the set of payment plans that satisfy conditions (AC-k), (PC-k) and (BC-k) is nonempty. Moreover, this set is compact. 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 or a lower punishment payoff for some state x and some player i.

Proposition 3. *There exists a simple equilibrium with an action plan $(\alpha^k)_{k\in\mathcal{K}}$ if and only if there exists a payment plan $(\bar{p}^k)_{k\in\mathcal{K}}$ that solves the following linear*

program

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

The plan $(\bar{p}^k)_{k\in\mathcal{K}}$ is an optimal payment plan for $(\alpha^k)_{k\in\mathcal{K}}$.

Proof. The proof is straightforward and therefore omitted. □

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 big weakness of this brute-force method is that it becomes computationally infeasible, except for very small 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. Unfortunately, the joint optimization problem of action plan and payment plan is non-convex, so that one cannot rely on efficient general purpose methods of convex optimization problems that guarantee a global optimum. For mixed strategy equilibria, there is the additional complication that the number of action constraints depends on the support of the mixed action profiles that shall be implemented.

4 Solving Games with Perfect Monitoring

In this section, we develop efficient methods to find an optimal simple equilibrium and to exactly compute the set of PPE payoffs in games with perfect monitoring and a finite action space.

4.1 Characterization for a given action plan

Consider a pure equilibrium regime policy a^e that specifies an action profile for each state x. An optimal payment plan under perfect monitoring involves no money burning. Therefore the joint equilibrium path payoffs U are given as the

solution to the following linear system of equations:[11]

$$U(x) = (1-\delta)\Pi(x, a^e) + \delta E[U(x')|x, a^e] \text{ for all } x \in X. \tag{10}$$

Now consider a pure punishment policy a^i against player i. After a deviation, a punished player i will be made exactly indifferent between paying the fines that settle the punishment within one period, or to refuse any payments and play against other players who follow this punishment policy in all future. Player i's punishment payoffs v_i given a punishment policy a^i will therefore be given as the solution to the following Bellman equation

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

It follows from the contraction mapping theorem that there exists a unique payoff vector v_i that solves this Bellman equation. This optimization problem for finding player i's dynamic best reply payoff is a discounted Markov decision process. One can compute v_i, for example with the policy iteration algorithm.[12] 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 $\tilde{a}_i(x)$ for each state x, which solves

$$\tilde{a}_i(x) \in \arg\max_{a_i \in A_i(x)} \{(1-\delta)\left(\pi_i(a_i, a^i_{-i}, x)\right) + \delta E[v_i(x')|x, a_i, a^i_{-i}]\}. \tag{12}$$

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

$$v_i(x) = (1-\delta)\pi_i(\tilde{a}_i, a^i_{-i}, x) + \delta E[v_i(x')|x, \tilde{a}_i, a^i_{-i}]. \tag{13}$$

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).

The following result is key for solving games with perfect monitoring.

Theorem 2. *Assume there is perfect monitoring. Under an optimal payment plan given a pure action plan $(a^k)_{k\in\mathcal{K}}$, joint equilibrium payoffs U solve (10) and for each player i the punishment payoffs v_i solve (11). There exists a simple equilibrium with action plan $(a^k)_{k\in\mathcal{K}}$ if and only if for all $x \in X$ these payoffs satisfy*

$$U(x) \geq \sum_{i=1}^{n} v_i(x), \tag{14}$$

[11] The condition has a unique solution since the transition matrix has eigenvalues with absolute value no larger than 1. The solution is given by $U = (1-\delta)(\mathrm{I} - \delta Q(a^e))^{-1}\Pi(a^e)$, where $Q(a^e)$ is the transition matrix given that players follow the policy a^e.

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

and $(a^k)_k$ *satisfies for all* $k \in \mathcal{K}$ *and* $x \in X$

$$(1-\delta)\Pi(x,a^k) + \delta E[U|x,a^k] \geq \sum_{i=1}^{n} \max_{\hat{a}_i \in A_i(x)} (1-\delta)\pi_i(x,\hat{a}_i,a^k_{-i}) + \delta E[v_i|x,\hat{a}_i,a^k_{-i}]. \tag{15}$$

4.2 Finding optimal action plans

Note from inequality (15) that it is easier to implement any action profile $a^k(x)$ if -ceteris paribus- joint payoffs $U(x)$ increase in some state or punishment payoffs $v_i(x)$ decrease for some player in some state. Therefore the action plan of an optimal simple equilibrium maximizes $U(x)$ and minimizes $v_i(x)$ for each state and player across all action profiles that satisfy the conditions (14) and (15) in Theorem 2.

We now develop an iterative algorithm to find such an optimal action plan. In every iteration of the algorithm there is a candidate set of action profiles $\hat{A}(x) \subset A(x)$ which have not yet been ruled out as being possibly played in some simple equilibrium. $\hat{A} = \prod_{x \in X} \hat{A}(x)$ shall denote the corresponding set of policies.

Optimal equilibrium regime policy

Let $U(.,a^e)$ denote the solution of (10) for equilibrium regime policy a^e. We denote by

$$U(x;\hat{A}) = \max_{a^e \in \hat{A}} U(x,a^e) \tag{16}$$

the maximum joint payoff that can be implemented in state x using equilibrium regime policies from $\hat{A}$. Like the problem (11) of finding a dynamic best reply against a given punishment policy the problem of computing $U(.;\hat{A})$ is a finite discounted Markov decision process. A solution always exists and it can be efficiently solved using policy iteration.

Optimal punishment policies

Let $v_i(.,a^i)$ be the resulting punishment payoffs, which solves the Bellman equation (11), given a policy a^i against player i. For the punishment regimes, we define by

$$v_i(x;\hat{A}) = \min_{a^i \in \hat{A}} v_i(x,a^i) \tag{17}$$

player i's minimum punishment payoff in state x across all punishment policies in $\hat{A}$. Let $\bar{a}^i(\hat{A})$ be the optimal punishment policy that solves this problem. Computing $v_i(x;\hat{A})$ and $\bar{a}^i(\hat{A})$ is a nested dynamic optimization problem. We need to find that dynamic punishment policy that minimizes player i's dynamic best-reply payoff against this punishment policy. While a brute force method that

tries out all possible punishment policies is theoretically possible, it is usually computationally infeasible in practice since already for moderately sized games (like our example in Subsection 4.3 below) the set of candidate policies can be incredibly large.

A crucial building block for finding an optimal simple equilibrium is Algorithm 1 below, that solves this nested dynamic problem by searching among possible candidate punishment policies a^i in a monotone fashion.

We denote by

$$c_i(x, a, v_i) = \max_{\hat{a}_i \in A_i(x)} ((1-\delta)\pi_i(x, \hat{a}_i, a_{-i}) + \delta E[v_i(x')|x, \hat{a}_i, a_{-i}]) \tag{18}$$

player i's best-reply payoff of a static version of the game in state x in which action profile a shall be played and continuation payoffs in the next period are given by fixed numerical vector v_i.

Algorithm 1. *Nested policy iteration to find an optimal punishment policy* $\bar{a}^i(\hat{A})$

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

$$a^{r+1}(x) \in \arg\min_{a \in \hat{A}(x)} c_i(x, a, v_i(., a^r)) \tag{19}$$

3. *Stop if* a^r *itself solves step 2. Otherwise increment the round* r *and go back to step 1.*

Note that in step 2, we update the punishment policy by minimizing state-by-state the best reply payoffs $c_i(x, a, v_i(., a^r))$ for the fixed punishment payoff $v_i(., a^r)$ derived in the previous step. This operation can be performed very quickly. Remarkably, this simple static update rule for the punishment policy suffices for the punishment payoffs $v_i(., a^r)$ to monotonically decrease in every round r.

Proposition 4. *Algorithm 1 always terminates in a finite number of periods, yielding an optimal punishment policy* $a^i(\hat{A})$*. The punishment payoffs decrease in every round (except for the last round):*

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

The proof in the appendix exploits monotonicity properties of the contraction mapping operator that is used to solve the Markov decision process in step 1. In the examples we computed, the algorithm typically finds an optimal punishment policy by examining a very small fraction of all possible policies.[13] While one can construct examples in which the algorithm has to check every possible policy in $\hat{A}$, the monotonicity results suggest that the algorithm typically stops after a few rounds.

The outer loop

The procedure allows us to compute for every set of considered action profiles $\hat{A}$ the highest joint payoffs $U(.,\hat{A})$ and lowest punishment payoffs $v_i(.,\hat{A})$ that can be implemented if all action profiles in $\hat{A}$ would be enforceable in a PPE. Following similar steps as in the proof of Theorem 2, one can easily show that given a simple equilibrium with equilibrium regime payoffs $U(.,\hat{A})$ and punishment payoffs $v_i(.,\hat{A})$ exists, an action profile $a(x)$ can be played in a PPE starting in state x, if and only if the following condition on joint payoffs is satisfied.

$$(1-\delta)\Pi(a,x)+\delta E[U(x',\hat{A})|a,x] \geq \sum_{i=1}^{n} \max_{\hat{a}_i \in A_i(x)} (1-\delta)\pi_i(x,\hat{a}_i,a_{-i})+\delta E[v_i(x',\hat{A})|\hat{a}_i,a_{-i}(x),x]. \quad (20)$$

If we start with the set of all action profiles $\hat{A}=A$, we know that all action profiles that do not satisfy this condition can never be played in a PPE. We can remove those action profiles from the set $\hat{A}$. If the optimal policies $\hat{a}^k(\hat{A})$ have remained in the set, they form an optimal simple equilibrium, otherwise we must repeat this procedure with the smaller set of action profiles until this condition is satisfied.

Algorithm 2. *Policy elimination algorithm to find optimal action plans*

0. *Let $j=0$ and initially consider all policies as candidates: $\hat{A}^j=A$.*
1. *Compute $U^e(.;\hat{A}^j)$ and a corresponding optimal equilibrium regime policy $\hat{a}^e(\hat{A}^j)$.*
2. *For every player i compute $v_i(.;\hat{A}^r)$ and a corresponding optimal punishment policy $\hat{a}^i(\hat{A}^r)$*

[13] For an example, consider the Cournot game described in Subsection 4.3 below. It has 21*21=441 states and, depending on the state, a player has between 0 to 20 different stage game actions. If we punish player 1, the number of potentially relevant pure strategy punishment policies a brute force algorithm has to search is given by the number of pure Markov strategies of player 2. Here, each player has $\prod_{m_1=0}^{20}\prod_{m_2=0}^{20} m_1 = (20!)^{21}$ different pure Markov strategies. This is an incredible large number and renders a brute-force approach infeasible. Yet, in no iteration of the outer loop, does Algorithm 1 need more than just 4 rounds to find an optimal punishment policy.

3. *For every state x, let $\hat{A}^{j+1}(x)$ be the set of all action profiles that satisfy condition (20) using $U^e(.;\hat{A}^j)$ and $v_i(.;A^j)$ as equilibrium regime and punishment payoffs.*

4. *Stop if the optimal policies $\hat{a}^k(\hat{A}^j)$ are contained in $\hat{A}^{j+1}$. They then constitute an optimal action plan. Also stop if for some state x the set $\hat{A}^{j+1}$ 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 $(\bar{a}^k)_{k\in\mathcal{K}}$ 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 $\hat{A}^j$ gets strictly smaller with each round. Therefore $U(x;\hat{A}^j)$ weakly decreases and $v_i(x;\hat{A}^j)$ weakly increases each iteration. Condition (20) is easier satisfied for higher values of $U(x;\hat{A}^j)$ and for lower values of $v_i(x;\hat{A}^j)$. 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 $\hat{a}^k(\hat{A}^j)$ all survive Step 3, it follows from Proposition 2 that a simple equilibrium with these policies exists. That they constitute an optimal action plan simply follows again from the fact that $U(x;\hat{A}^j)$ weakly decreases and $v_i(x;\hat{A}^j)$ weakly increases each round. That the algorithm terminates in a finite number of rounds is a consequence of the finite action space and the fact that the set of possible policies $\hat{A}^j$ gets strictly smaller each round.

4.3 Example: Quantity competition with stochastic reserves

As numerical example, consider a stochastic game variation of the example Cournot used to motivate his famous model of quantity competition. There are two producers of mineral water, who have finite water reserves in their reservoirs. A state is two dimensional $x = (x_1, x_2)$, where 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 $\bar{x}$. We solve this game with the following parameters: 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]

[14]To replicate the example, follow the instructions on the Github page of our R package dyngame: https://github.com/skranz/dyngame. This package has implemented the policy elimination algorithm described above. This example with 21*21=441 states is solved with 8 iterations of the outer loop, and takes less than a minute on an average notebook bought in 2013.

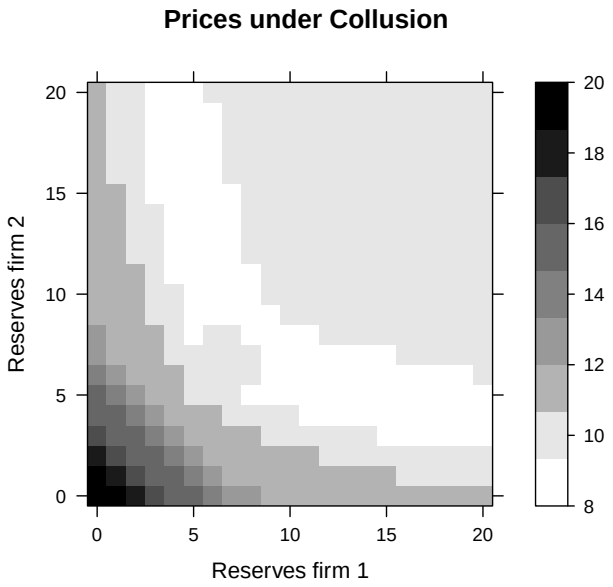

Figure 2: Optimal collusive prices as function of firms' reserves. Brighter areas correspond to lower prices.

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. This seems intuitive, since firms are able to supply more with larger reserves. Yet, moving to the upper right corner we see that equilibrium prices are not monotonically decreasing in the reserves: once reserves become sufficiently large, prices increase again. An intuitive reason for this effect 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 sell large quantities in the next period.

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

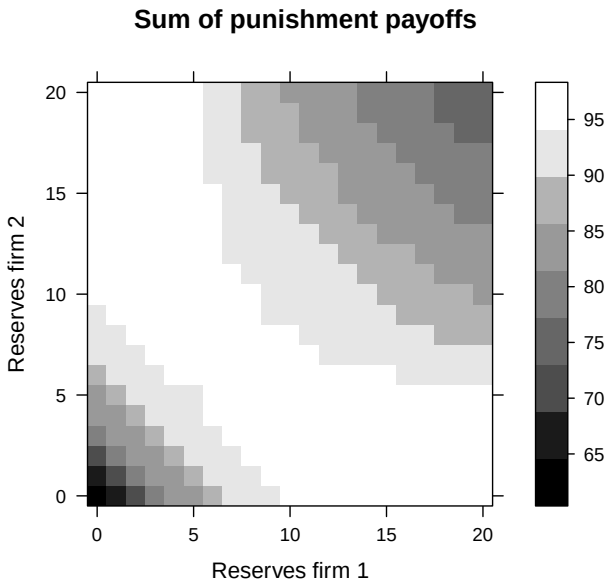

Figure 3: Sum of punishment payoffs $\bar{v}_1(x) + \bar{v}_2(x)$. Darker areas correspond to lower punishment payoffs.

5 Decomposition methods

To deal with imperfect public monitoring, we first reformulate the recursive APS approach for our class of games. We then adapt the implementation of Judd, Yeltekin, and Conklin (2003) to our framework with transfers and imperfect public monitoring to develop decompositions methods that allow to approximate the set of PPE payoffs.

5.1 APS

The recursive methods of APS directly transfer to stochastic games (see e.g. Judd and Yeltekin (2011) for such an extension). In the following, we adapt the terminology to our case of a stochastic game with transfers.

For any collection of (continuation payoff) sets $\mathcal{W} = \prod_{x \in X} \mathcal{W}(x)$ with $\mathcal{W}(x) \subset \mathbb{R}^n$, we say that an action profile $\alpha \in \mathcal{A}(x)$ is **enforceable** on $\mathcal{W}$ in state x if there exists a function

$$w : Y \times X \to \bigcup_{\hat{x} \in X} \mathcal{W}(\hat{x}) \text{ with } w(y, x') \in \mathcal{W}(x') \text{ for all } y,$$

such that α is a Nash equilibrium of the static game with action set $A(x)$ and payoffs

$$(1 - \delta)\pi_i(x, a) + \delta E[w_i(y, x')|x, a].$$

The function w **enforces** α. Note that the payoff functions in the static game are continuous. We say that a payoff vector v is **decomposable** on $\mathcal{W}$ in state x if there exist $\alpha \in \mathcal{A}(x)$ and w such that α is enforced by w on $\mathcal{W}$ in state x and

$$v_i = (1-\delta)\pi_i(x,\alpha) + \delta E[w_i(y,x')|x,\alpha].$$

We define an operator B that maps a collection of continuation payoff sets $\mathcal{W}$ into a collection of sets of decomposable payoffs:

$$B(\mathcal{W}) = \prod_{x\in X}\{v \in \mathbb{R}^n; v \text{ is decomposable on } \mathcal{W} \text{ in state } x\}.$$

We have illustrated in Section 3.2 how the possibility of upfront transfers transforms the payoff set into a simplex. To account for upfront transfers (but not yet assuming compactness of the payoff set), we define a set operator T that maps a subset $\mathcal{W} \subset \mathbb{R}^n$ to

$$T(\mathcal{W}) = \left\{u \in \mathbb{R}^n | \sum_{i=1}^n u_i \leq \sum_{i=1}^n w_i \text{ and } u_i \geq w_i^i \text{ for some } w, w^1, ..., w^n \in \mathcal{W}\right\}. \tag{21}$$

The possibility of up-front transfers is incorporated by defining

$$D(\mathcal{W}) = \times_{x\in X} T(B(\mathcal{W})_x),$$

A set $\mathcal{W}$ is called **self-generating** if $\mathcal{W} \subset D(\mathcal{W})$.

Lemma 1. *The results of APS apply:*

(i) The operator D is monotone: If $\mathcal{W} \subset \mathcal{W}'$ then $D(\mathcal{W}) \subset D(\mathcal{W}')$.

(ii) If $\mathcal{W}$ is compact, then $D(\mathcal{W})$ is compact.

(iii) The set of PPE payoffs $\mathcal{U} = \prod_{x_0\in X}\mathcal{U}(x_0)$ is a fixed point of D.

(iv) Any bounded self-generating set is a subset of $\mathcal{U}$.

Proof. Closely following the arguments in APS yields the results $(i)-(iv)$ for the operator B. The results for D follow since T is monotone, preserves compactness, and has $\mathcal{U}(x_0)$ as a fixed point. Moreover, applying T to a subset of $\mathcal{U}(x_0)$ yields again a subset of $\mathcal{U}(x_0)$. □

Note that when we apply the operator D to a compact set, the result is a collection of n-simplices, which are spanned by $n+1$ vectors of the form $(u_1, ..., u_n)$ with $u_i = v_i$ for all but at most one j, and $u_j = U - \sum_{i\neq j} v_i$, for some $v_1, ..., v_n, U$. To represent such a simplex, one therefore needs only $n+1$ numbers, and if we iteratively apply the operator D, we obtain decreasing sequences of such simplices.

To ensure that $\mathcal{U}$ is a subset of the sets in this sequence, we start with vectors U^0 and v^0 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 $i = 1, ..., n$. For example, we can start with the maximum feasible payoff

$$U^0(x) = \max_s \sum_{i=1}^{n} u_i(x, s),$$

where we know from Dutta (1995) that maximization over the set of pure Markov strategies suffices, and transfers play no role. Similarly, v^0 can be the min-max payoffs

$$v_i^0(x) = \inf_{\sigma_{-i}} \sup_{\sigma_i} u_i(x, \sigma),$$

where again transfers play no role. When we iteratively apply D to the set

$$F = \prod_{x \in X} \{u \in \mathbb{R}^n : \sum_{i=1}^{n} u_i \leq U^0(x) \text{ and } u_i \geq v^0(x)\},$$

then $\mathcal{U} \subset D^m(F)$ for all m. This follows from monotonicity of the operator D and the fact that $\mathcal{U}$ is a fixed point of D. The sequence $D^m(F)$ converges against $\mathcal{U}$ in the Hausdorff-metric.

Lemma 2. *The set $\bigcap_{m=1}^{\infty} D^m(F)$ is equal to $\mathcal{U}$.*

Proof. The proof is omitted as it directly follows from APS. □

As the intersection of compact sets, the set of PPE payoffs $\mathcal{U}$ must be compact as well.

5.2 Decomposition methods for outer and inner approximations

For any $(U, v) \in \mathbb{R}^{(n+1)|X|}$ and action profile $\alpha \in \mathcal{A}(x)$, let $\mathcal{W}(x, \alpha, U, v)$ be the set of all w that enforce α in state x on the set of continuation payoffs

$$\prod_{x' \in X} \{u \in \mathbb{R}^n : \sum_{i=1}^{n} u_i \leq U(x') \text{ and } u_i \geq v(x')\}.$$

We can rewrite our decomposition operator D as a map $\hat{D} : \mathbb{R}^{(n+1)|X|} \to \mathbb{R}^{(n+1)|X|}$, which maps a collection (U, v) of maximum total payoffs and minimum payoffs into a new collection of such payoffs (U', v') such that the following conditions hold:

- For each state $x \in X$

$$U'(x) = \max_{\alpha \in \mathcal{A}(x)} \hat{U}(x, \alpha, U, v) \tag{22}$$

where $\hat{U}(x,\alpha,U,v)$ is defined by $\hat{U}(x,\alpha,U,v) = -\infty$ if the set $\mathcal{W}(x,\alpha,U,v)$ is empty and else by

$$\hat{U}(x,\alpha,U,v) = \max_{w\in\mathcal{W}(x,\alpha,U,v)} (1-\delta)\Pi(x,\alpha) + \delta E[W(y,x')|x,\alpha]. \quad (23)$$

- For each state $x \in X$ and $i \in \{1,...,n\}$

$$v_i'(x) = \min_{\alpha\in\mathcal{A}(x)} \hat{v}_i(x,\alpha,U,v) \quad (24)$$

where $\hat{v}_i(x,\alpha,U,v)$ is defined by $\hat{v}_i(x,\alpha,U,v) = \infty$ if the set $\mathcal{W}(x,\alpha,U,v)$ is empty and else by

$$\hat{v}_i(x,\alpha,U,v) = \min_{w\in\mathcal{W}(x,\alpha,U,v)} (1-\delta)\pi_i(x,\alpha) + \delta E[w_i(y,x')|x,\alpha]. \quad (25)$$

Note that the condition that $w \in \mathcal{W}(x,\alpha,U,v)$ means that $w_i(y,x') \geq v_i$ for all $i = 1,...,n$, and

$$W(y,x') = \sum_{i=1}^{n} w_i(y,x') \leq U,$$

and

$$(1-\delta)\pi_i(x,a_i,\alpha_{-i}) + \delta E[w_i|x,a_i,\alpha_{-i})] \geq (1-\delta)\pi_i(x,\hat{a}_i,\alpha_{-i}) + \delta E[w_i|x,\hat{a}_i,\alpha_{-i}^k]$$

for all $a_i, \hat{a}_i \in A_i(x)$ with $\alpha(a_i) > 0$. Therefore, the optimizations over $\mathcal{W}$ are just linear optimization problems.

We can directly write the results for the operator D obtained in Lemma 1 in terms of this new operator $\hat{D}$. Since $\mathcal{U}$ is compact, it is described by the largest total payoffs and lowest possible payoffs $(\bar{U},\bar{v})$. Hence $(\bar{U},\bar{v})$ is a fixed point of $\hat{D}$, which means that

$$\bar{U}(x) = \hat{U}(x,\bar{\alpha}^e(x),\bar{U},\bar{v}) \text{ for all } x, \quad (26)$$

$$\bar{v}_i(x) = \hat{v}_i(x,\bar{\alpha}^i(x),\bar{U},\bar{v}) \text{ for all } x,i \quad (27)$$

for some action plan $(\bar{\alpha}^k)_k$. This action plan is the action plan of an optimal simple equilibrium which according to Theorem 1 describes the PPE payoff set. Since $\mathcal{U}$ is the largest fixed point of D, a converse also holds: Among all action plans $(\alpha^k)_k$ and values (U,v) that satisfy equations (26) and (27), the action plan that maximizes $\sum_{x\in X}(U(x) - \sum_{i=1}^{n} v_i(x))$ must be an action plan of an optimal simple equilibrium, with the corresponding (U,v) describing the PPE payoff set.

There is also a connection between simple equilibria and the compact self-generating sets of D. There exists a simple equilibrium with an action plan $(\alpha^k)_{k\in\mathcal{K}}$ if and only if there exist U and v such that

$$\hat{U}(x,\alpha^e,U,v) \geq U(x) \text{ for all } x \in X, \quad (28)$$

$$\hat{v}_i(x,\alpha^i,U,v) \leq v_i(x) \text{ for all } x \in X, i = 1,...,n. \quad (29)$$

Finally, we know from Lemma 2 that if we start with a set $\mathcal{F}$ that contains $\mathcal{U}$, the sequence $(D^m(\mathcal{F}))_m$ that is obtained by iteratively applying the operator D converges to $\mathcal{U}$ in the Hausdorff metric as $m \to \infty$. Expressed in terms of the operator $\hat{D}$, this result means: Starting with values $U^0(x) \geq \bar{U}(x)$ and $v_i^0(x) \leq \bar{v}_i(x)$ for all $x \in X$, the sequence $(\hat{D}^m(U^0, v^0))_m$ converges to $\bar{U}$ (from above) and $\bar{v}$ (from below). Repeatedly applying the operator $\hat{D}$ yields in every round a tighter outer approximation for $\bar{U}$ and $\bar{v}$, and hence for the PPE payoff set.

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 the methods from Section 4.

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, Judd, Yeltekin, and Conklin (2003) 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 $\hat{D}$ on these adjusted 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 (22) and (24). 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.

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 regime the set of action profiles that can possibly 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 $\mathcal{W}(x, \alpha, U^{out}, v^{out})$ is empty, then there does not exist any PPE in which α is played and we can dismiss it. If α can be enforced by some $w \in \mathcal{W}(x, \alpha, U^{out}, v^{out})$, but

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

then α will not be played in the equilibrium regime in state x of an optimal equilibrium, since even with the outer approximations of U and v it can only 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, if the number of candidate action profiles can be sufficiently reduced, it may become tractable to compute the exact payoff set by applying the brute force method from Subsection 3.6 on the remaining action plans.

6 Principal-agent examples

The following two examples illustrate how our results can be used to easily obtain closed form solutions in two examples of principal-agent relationships that are described by stochastic games.

6.1 A principal-agent game with a durable good

In our first example, a principal (player 1) can employ an agent (player 2) to produce a single durable good for her. If the product has been successfully produced, the state of the world will be given by x_1, 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

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

We assume $(1-\delta)c < \delta$, 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 is $\bar{v}_2 = 0$, and the principal's punishment payoffs are $\bar{v}_1(x_0) = 0$ and $\bar{v}_1(x_1) = 1$. Using Theorem 2, we can conclude that effort e can be implemented if and only if $U(x_0,e) \geq e\delta$, i.e., if

$$(1-\delta)\, c \leq \delta^2(1-e). \tag{30}$$

Condition (30) 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 $\delta \in [0,1)$. The intuition is simple. 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 will not conduct any transfers. The principal can only reward the agent for positive effort in the case that the agent has exerted high effort but the project has not been successful, which happens with probability $(1-e)$. Thus, the agent cannot be reimbursed for full effort, but there is a positive chance to get reimbursed for partial effort.

Imperfect monitoring and costly punishment 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. The reason is that 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 exert 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 for sufficiently large discount factors.

We now introduce the possibility of costly punishment. Assume 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 and the state becomes again x_0, while for $d = 0$ the product remains intact. The agent incurs costs for destructive efforts of size kd with $k \geq 0$.

To find the optimal simple equilibrium, we consider the possible action profiles of the agent. If the optimal simple equilibrium has no destructive effort ($a_2^1(x_1) = 0$), it must be the same as in the previous case with zero production effort. If the optimal simple equilibrium has $a_2^1(x_1) = 1$, the principal's punishment payoffs are $\bar{v}_1(x_0) = 0$ and

$$\bar{v}_1(x_1) = (1 - \delta).$$

The agent's punishment payoff is still $\bar{v}_2 = 0$ in both states. To find the optimal simple equilibrium for the case that positive effort is possible, note first that maximum incentives for the agent are created by maximally rewarding the transition between the two states compared to staying in a state. According to condition (28), an action plan with equilibrium regime action $a_2^e(x_0) = e > 0$ and punishment regime action $a_2^1(x_0) = 1$ can be part of a simple equilibrium if and only if there exist values $U(x_0), U(x_1)$ with $-(1-\delta)ce+\delta(eU(x_1)+(1-e)U(x_0)) \geq U(x_0)$ and $1 \geq U(x_1)$ such that

$$-(1 - \delta)k + \delta U(x_0) \geq 0. \tag{31}$$

and

$$e \in \arg\max_{\hat{e}} -(1 - \delta)c\hat{e} + \delta\hat{e}(U(x_1) - (1 - \delta)) \tag{32}$$

This last condition shows that every optimal simple equilibrium in which the agent chooses positive effort must have maximal effort $e = 1$. Moreover, the conditions

are easiest to satisfy if $U(x_1) = 1$ and $U(x_0) = \delta - (1-\delta)c$, which means that destroying output is not optimal on the equilibrium path. Overall, it follows that high effort can be implemented if and only if

$$(1-\delta)(\delta c + k) \leq \delta^2 \tag{33}$$

and

$$(1-\delta)c \leq \delta^2. \tag{34}$$

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.

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 Nash reversion strategies 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 analog to Nash reversion 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 a punishment cannot achieve any positive effort by the agent, since the principal will never make positive transfers in state x_1. The ineffectiveness of reversion to a MPE as a punishment in this simple example illustrates that for stochastic games it is particularly useful to have a simple characterization of equilibria with optimal penal codes.

6.2 A principal-agent game with an outside option

As our last example, we consider a principal-agent game in which the agent can devote effort to two different tasks: He can exert production effort in the relationship with the principal, and/or exert search effort to work towards an outside alternative.[15] This example illustrates that the presence of transfers does not imply that the set of PPE payoffs is increasing in the discount factor. We will see that when the agent can invest into his outside option, his punishment payoff is increasing in the discount factor and consequently the set of PPE payoffs can be smaller for larger discount factors.

The game between the principal (player 1) and the agent (player 2) is as follows. If the game is in the initial state x_0, principal and agent first decide whether they take their outside option, which yields 0 for both. If both decide against the outside option, the agent can choose unobservable productive effort $e \in [0,1]$ and search effort $s \in [0,1]$. The cost of effort to the agent is equal to $c(e,s) = (e+s)^2$.

[15] The set-up is reminiscent of Herbold (2014), who analyzes on the job search. In our simple example, however, it is never optimal to have the agent spend some effort in the current relationship and some on search.

With probability e, the principal receives a return $y \geq 2$.[16] With probability s, the game moves to a state x_1, in which the game is the same as in x_0 except that taking the outside option would now yield 1 to the agent. We assume that the agent can search independently of the principal: If one of the players decides to take the outside option in state x_0, the agent can choose search effort $s \in [0,1]$ at cost $c(0,s)$ to increase the probability s of a state transition.[17]

The principal's min-max payoff is $\bar{v}_1 = 0$ in both states. The agent's min-max payoff in state x_1 is given by $\bar{v}_2(x_1) = 1$, while in state x_0 is given by

$$\bar{v}_2(x_0) = \max_s \frac{\delta - (1-\delta)s}{\delta s + 1 - \delta} s,$$

which can be calculated to equal

$$\bar{v}_2(x_0) = \frac{2 - 4\delta + 3\delta^2 - 2(1-\delta)\sqrt{1 - 2\delta + 2\delta^2}}{\delta^2}.$$

The punishment payoff $\bar{v}_2(x_0)$ is increasing in δ, since the same search effort creates a larger surplus when δ is larger. All these min-max payoffs are achieved by MPE.

To characterize the set of PPE payoffs we need to determine the largest surplus that can be generated in a simple equilibrium. Note first that any effort level that can be implemented in a simple equilibrium in state x_1 can also be implemented in state x_0. Since we are only interested in the simple equilibrium that generates the largest possible surplus in state x_0, it suffices to consider simple equilibria in which agent and principal would take the outside option in state x_1, yielding payoff vector $(0,1)$ in state x_1. Since the agent's marginal return to effort is constant, we only need to consider simple strategy profiles in which the agent either concentrates on creating surplus in the relationship ($s(x_0) = 0$) or outside of the relationship ($e(x_0) = 0$).[18]

The maximum feasible joint surplus is achieved by work effort $e^{FB} = 1$ and search effort $s^{FB} = 0$, yielding a surplus of $U^0(x_0) = y - 1$. We first ask for conditions on the discount factor δ such that $U^0(x_0)$ can be attained in a simple equilibrium. In this case, the algorithm that is outlined in Section 5.2 would terminate after the first step. If a high return y is rewarded with maximum continuation payoff $U^0(x_0)$, and a low return 0 with minimum continuation payoff $\bar{v}_2(x_0)$, the return to production effort if $s = 0$ is equal to $\frac{\delta}{1-\delta}(U^0(x_0) - \bar{v}_2(x_0))$, while the return to search effort if $e = 0$ is equal to $\frac{\delta}{1-\delta}(1 - \bar{v}_2(x_0))$. Hence, first best effort is

[16] The assumption $y \geq 2$ guarantees that cooperation is efficient. It follows from the analysis below that for $y \leq 2$, no cooperation at all is possible.

[17] Note that this assumption implies that the discount factor in this example cannot be interpreted as a survival rate of the relationship.

[18] To see this formally, note that for any continuation payoffs given by w, the agent maximizes $-c(e+s)(1-\delta) + \delta(s + (1-s)ew(y,0) + (1-s)(1-e)w(0,0))$. The Hesse matrix has principal determinants equal to $-c''(e+s)(1-\delta)$ and $(c''(e+s)(1-\delta))^2 - (c''(e+s)(1-\delta) + \delta(w(y,0) - w_0(0,0)))^2 \leq 0$, hence there is no interior maximum.

enforceable with continuation payoffs between $\bar{v}_2(x_0)$ and $U^0(x_0)$ in state x_0 if and only if the following two conditions are satisfied:

$$y \geq 2$$

and

$$\frac{\delta}{1-\delta}(y - 1 - \bar{v}_2(x_0)) \geq c'(1) = 2. \tag{35}$$

The first condition is always satisfied. Evaluating condition (35), one can show that it is never satisfied for $y = 2$, but that for $y > 2$ there is a cut-off $\bar{\delta}$ such that it holds for larger δ. If $\delta \geq \bar{\delta}$, the set of PPE payoffs is given by

$$\mathcal{U}(x_0) = \{(u_1, u_2) \in \mathbb{R}^2;\ u_1 + u_2 = y - 1, u_1 \geq 0, u_2 \geq \bar{v}_2(x_0)\}.$$

In this range of discount factors, the payoff set is shrinking in δ, since the agent needs to receive a larger share of the surplus as the discount factor increases.

For $\delta < \bar{\delta}$, the largest effort level is given by a fixed point equation (corresponding to equation (26)). An effort level $e > 0$ can be implemented with continuation payoffs between $\bar{v}_2(x_0)$ and U if $\frac{\delta}{1-\delta}(U - \bar{v}_2(x_0)) = 2e$ and $U \geq 1$. The largest possible effort level in a simple equilibrium is therefore given by the largest solution to

$$e = \frac{\delta}{2(1-\delta)}(e(y - e) - \bar{v}_2(x_0))$$

that also satisfies $e(y - e) \geq 1$. If no solution exists, no cooperation is possible and $\mathcal{U}(x_0)$ only contains the payoff vector $(0, \bar{v}_2(x_0))$.

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, 56(2), pp.383–396.

Abreu, D., Pearce, D. & Stacchetti, E., 1990. Toward a theory of discounted repeated games with imperfect monitoring. Econometrica, 58(5), pp.1041–1063.

Abreu, D. & Sannikov, Y., 2014. An Algorithm for Two Player Repeated Games with Perfect Monitoring. Theoretical Economics, 9, 313-338.

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, 35(1), 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, 146, pp.1664-1683.

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, 147(3), pp.1191-1221.

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

Herbold, D., 2015. A Repeated Principal-Agent Model with On-the-Job Search. University of Frankfurt, mimeo.

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.

Judd, K.L. & Yeltekin, S. 2011. Computing equilibria of dynamic games, working paper, Tepper School of Business.

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.

Goldlücke, S. & Kranz, S., 2013. Renegotiation-proof relational contracts. Games and Economic Behavior, 80, pp. 157-178.

Herbold, D., 2014. The Agency Costs of On-the-Job Search, working paper, Goethe-University Frankfurt.

Levin, J., 2002. Multilateral contracting and the employment relationship. The Quarterly Journal of Economics, 117(3), pp.1075-1103.

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., 2013. A theory of disagreement in repeated games with bargaining. Econometrica, 81(6), pp.2303–2350.

Pakes, A. & McGuire, P., 1994. Computing Markov-Perfect Nash Equilibria: Numerical Implications of a Dynamic Differentiated Product Model. The RAND Journal of Economics, 25(4), 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.937-963.

Rockafellar, R.T., 1970. Convex Analysis. Princeton University Press.

Sleet, C. & Yeltekin, S., 2015. On the computation of value correspondences. Dynamic Games and Applications.

Appendix: Remaining Proofs

Proof of Theorem 2:

For each state $x \in X$ and regime $k \in \mathcal{K}$, condition (15) allows to choose a distribution $u_i^k(x)$, $i = 1, ..., n$, of the surplus such that

$$\sum_{i=1}^{n} u_i^k(x) = (1-\delta)\Pi(x, a^k) + \delta E[U|x, a^k] \tag{36}$$

and

$$u_i^k(x) \geq \max_{\hat{a}_i}(1-\delta)\pi_i(x, \hat{a}_i, a_{-i}^k) + \delta E[v_i|x, \hat{a}_i, a_{-i}^k], \tag{37}$$

holding with equality for $i = k$. A simple strategy profile with transfers $p_i^k(x, a^k, x')$ achieves this distribution of payoffs if the expected transfers $\bar{t}_i^k(x) = (1-\delta)E[p_i^k(x, a^k, x')|x, a^k]$ satisfy

$$\delta \bar{t}_i^k(x) = (1-\delta)\pi_i(x, a^k) + \delta E[u_i^e|x, a^k] - u_i^k(x).$$

If we define $\bar{t}_i^k(x)$ by this condition, it holds that $\sum_{i=1}^n \bar{t}_i^k(x) = 0$. Moreover, it follows from condition (37) that

$$E[u_i^e - v_i|x, a^k] \geq \bar{t}_i^k(x).$$

The intuition behind this is that it is more difficult to induce an action and a subsequent expected payment afterward than to induce an expected payment. We still need to show that for each $k \in \mathcal{K}$ and state x there exist payments $t_i(x') = (1-\delta)p_i^k(x, a^k, x')$ for each state x' such that the following three conditions hold:

$$t_i(x') \leq u_i^e(x') - v_i(x') \tag{38}$$

$$\sum_{i=1}^n t_i(x') = 0 \tag{39}$$

$$\sum_{x'} q(x'|x)t_i(x') = \bar{t}_i^k(x) \tag{40}$$

We use Theorem 22.1 in Rockafellar's "Convex Analysis" to show that such payments exist. This theorem says that the existence of a vector with entries $t_i(x')$, $i = 1, .., n$, $x' \in X$, that satisfies the above three conditions is equivalent to the non-existence of real numbers $\lambda_i(x') \geq 0$, $\mu(x')$, and η_i, $i = 1, .., n$, $x' \in X$, that satisfy the following two conditions:

$$\lambda_i(x') + \mu(x') + \eta_i q(x'|x) = 0 \text{ for all } i, x' \tag{41}$$

$$\sum_{i,x'} \lambda_i(x')(u_i^e(x') - v_i(x')) + \sum_{i=1}^n \eta_i \bar{t}_i^k(x) < 0. \tag{42}$$

We assume to the contrary that such a solution to (41) and (42) exists. These two conditions imply that

$$-\sum_{x'} \mu(x')(U(x') - \sum_{i=1}^n v_i(x')) + \sum_{i=1}^n \eta_i(\bar{t}_i^k(x) - E[u_i^e - v_i|x]) < 0.$$

Let $\tilde{x}$ be a state with $\frac{\mu(\tilde{x})}{q(\tilde{x}|x)} \leq \frac{\mu(x')}{q(x'|x)}$ for all $x' \in X$. Since condition (41) holds for all $x' \in X$, it also holds for $x' = \tilde{x}$, i.e., $\eta_i = -\frac{\lambda_i(\tilde{x}) + \mu(\tilde{x})}{q(\tilde{x}|x)}$. Hence, it follows that

$$\sum_{x'} (\frac{\mu(\tilde{x})q(x'|x)}{q(\tilde{x}|x)} - \mu(x'))(U(x') - \sum_{i=1}^n v_i(x')) + \sum_{i=1}^n \frac{\lambda_i(\tilde{x})}{q(\tilde{x}|x)}(E[u_i^e - v_i|x] - \bar{t}_i^k(x)) < 0.$$

This implies

$$\sum_{x'} (\frac{\mu(\tilde{x})q(x'|x)}{q(\tilde{x}|x)} - \mu(x'))(U(x') - \sum_{i=1}^n v_i(x')) < 0.$$

By definition of $\tilde{x}$ and because of condition (14), the expression on the left-hand-side must be non-negative. Hence, we arrived at a contradiction, which means that the system given by (38), (39), and (40) must have a solution and we can define payments $(1-\delta)p_i^k(x,a^k,x') = t_i(x')$.

It remains to define the payments following a unilateral deviation. For any combination of states x,x' and signal y with $y_i \neq a_i^k(x)$ and $y_{-i} = a_{-i}^k(x)$ we choose payments

$$(1-\delta)p_i^k(x,y,x') = u_i^e(x') - v_i(x'), \tag{43}$$

such that continuation payoffs after a deviation in the action stage are indeed given by v_i. Payments for players other than i can be defined such that

$$(1-\delta)p_j^k(x,y,x') \leq u_j^e(x') - v_j(x')$$

and

$$\sum_{j=1}^{n} p_j^k(x,y,x') = 0,$$

using condition (14).

Now we have to show that the so defined simple strategy profile is indeed a PPE. The budget and payment constraints are satisfied by definition. The relevant action constraints take the form

$$u_i^k(x) \geq \max_{\hat{a}_i \in A_i(x)} ((1-\delta)\pi_i(x,\hat{a}_i,a_{-i}^k) + \delta E[v_i|x,\hat{a}_i,a_{-i}^k]),$$

and are therefore also satisfied (see inequality 37).

Proof of Proposition 4: For a given policy a, let C_i^a be an operator mapping the set of punishment payoffs in itself defined by

$$C_i^a(v_i)[x] = c_i(x,a(x),v_i)$$

It can be easily verified that C_i^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^a, which we denote by $v_i(a)$. This means

$$v_i(a) = C_i^a(v_i(a)) \tag{44}$$

It is a well known result that the operator C_i^a is monotone:

$$v_i \leq \tilde{v}_i \Rightarrow C_i^a(v_i) \leq C_i^a(\tilde{v}_i) \tag{45}$$

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 $[C_i^a]^k$ the operator that applies k times C_i^a and define its limit by

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

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

$$[C_i^a]^\infty(v) = v(a) \tag{46}$$

Furthermore, it follows from monotonicity of C_i^a that

$$C_i^a(v_i) \leq v_i \Rightarrow [C_i^a]^\infty(v_i) \leq v_i \tag{47}$$

and

$$C_i^a(v_i) < v_i \Rightarrow [C_i^a]^\infty(v_i) < v_i \tag{48}$$

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 a and $\tilde{a}$ the following monotonicity results hold

$$C_i^a(v(a)) = C_i^{\tilde{a}}(v(a)) \quad \Rightarrow \quad v(a) = v(\tilde{a}) \tag{49}$$

$$C_i^a(v(a)) > C_i^{\tilde{a}}(v(a)) \quad \Rightarrow \quad v(a) > v(\tilde{a}) \tag{50}$$

$$v(a) \not\leq v(\tilde{a}) \quad \Rightarrow \quad C_i^a(v(a)) \not\leq C_i^{\tilde{a}}(v(a)) \tag{51}$$

We exemplify the proof for (50). It follows from (44), the left part of (50), (47) and (46) that

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

(49) and can be proven similarly. To prove (51), assume that there is some $\tilde{a}$ with $C_i^a(v) \leq C_i^{\tilde{a}}(v)$ but $\tilde{v} \not\geq v$. We find

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

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

The fixed point condition (44) of the value determination step and the policy improvement step (19) imply that $v^r = C_i^{a^r}(v^r) \geq C_i^{a^{r+1}}(v^r)$. We first establish that if

$$v^r = C_i^{a^r}(v^r) = C_i^{a^{r+1}}(v^r). \tag{52}$$

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 $\hat{a}$ such that $v(a^r) \not\leq v(\hat{a})$, i.e. $\hat{a}$ leads

in at least some state x to a strictly lower best-reply payoff for player i than a^r. By (51) this would imply $C_i^{a^r}(v^r) \not< C_i^{\tilde{a}}(v^r)$. This means that $\hat{a}$ must also be a solution to the policy improvement step and since (52) holds, we then must have

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

However, (49) then implies that $v(a^r) = v(\hat{a})$, which contradicts the assumption $v(a^r) \not\leq v(\hat{a})$. 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 $v^r = C_i^{a^r}(v^r) > C_i^{a^{r+1}}(v^r)$. (50) 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.■