University of Innsbruck

Promoting coordination in summary-statistic games

Dominik Erharter

Working Papers in Economics and Statistics

2013-28

Research Platform
eeecon
Empirical and Experimental Economics
University of Innsbruck
http://eeecon.uibk.ac.at/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact Address:
University of Innsbruck
Department of Public Finance
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 7171
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
http://eeecon.uibk.ac.at/wopec/

For a list of recent papers see the backpages of this paper.

Promoting coordination in summary-statistic games*

Dominik Erharter †

October 18, 2013

Abstract

This paper studies how external incentives can help agents to coordinate in summary-statistic games. Agents follow a myopic best-reply rule and face a trade-off between efficiency and strategic uncertainty. A principal can help agents to coordinate on the Pareto optimal equilibrium by monitoring an appropriate number of agents. The optimal monitoring policy is 'minimally-invasive' – for every strategy profile of the agents, the principal either monitors just enough agents to make high effort a best-reply or does not monitor at all. Furthermore, given the principal's payoffs are supermodular and increasing at an increasing rate, the optimal monitoring policy is monotone in the number of agents who choose high effort.

Keywords: adaptive learning; Markov decision process; coordination failure; order-statistic game;
JEL Classification: C73; C62; D86;

*This paper is a revised version of Chapter 4 of my dissertation at the University of Innsbruck. I thank Michael Greinecker, Rudolf Kerschbamer, Karl Schlag and Markus Walzl for very helpful comments and discussions. Funding by the Vice Rector for Research at the University of Innsbruck is gratefully acknowledged.

†University of Innsbruck, Department of Economics, SOWI Building, Universitätsstrasse 15, 6020 Innsbruck, Austria. Email: dominik.erharter@uibk.ac.at. Phone: +43 (0)512 507-7374. Fax: + 43 (0)512 507-2980.

1 Introduction

Many coordination problems feature a trade-off between Pareto optimality and strategic uncertainty. Consider workers in assembly who get a bonus for producing high quality. If a certain number of workers shirks, an entire batch is spoiled and others' effort is wasted. Hence, everyone exerts low effort. To make high effort worthwhile, there have to be enough other workers who choose high effort as well. Temporary incentives might help to generate the critical mass needed to make high effort a best-reply. For example, a principal might be able to monitor individual effort at a cost and to punish shirkers. Whether this is worthwhile depends on many factors, including the cost of monitoring and the stability of high effort equilibria.

This paper studies the optimization problem of a principal who can induce agents to coordinate on high effort by costly monitoring. This setting can be translated to many other coordination problems. For example, a government might help consumers to coordinate on efficient levels of production (Bryant, 1983) and a profession might help its members to coordinate on high ethical standards.[1] In the proposed model, agents play a sequence of summary-statistic games. A summary statistic game is a game where agents' payoff depends only on their own action and a summary statistic of all agents' actions.[2] Agents can exert low (L) or high (H) effort. The summary statistic can for example be the minimum or the median of all agents' effort, or the number of agents who exert high effort. Payoffs are such that the game has at most two pure-strategy Nash equilibria, one where all agents coordinate on low effort (All-L) and one where they coordinate on high effort (All-H). A principal observes the number of agents who choose high effort. By monitoring agents, the principal can change the agents' incentives in a way that is equivalent to a change of the relevant summary statistic. The threat of monitoring makes it more risky for agents to choose low effort and hence more attractive to choose high effort.

To capture the dynamic nature of transitions between equilibria, I assume that agents follow a myopic best-reply rule. Models of best-reply learning (Kandori et al., 1993; Young, 1993) assume that agents optimize, but base their beliefs only on the recent past. Furthermore, agents disregard their own impact on aggregate behavior, update their strategies infrequently and are prone to make errors. Compared to learning models where agents do not optimize, this is a very mild form of bounded-rationality. Early economic experiments have shown that behavior in summary-statistic games is driven to a large extent

[1] Schelling (1978) and Robles (1997) provide many other examples for coordination problems where the approach of this paper might help to reach efficient equilibria.

[2] Many games in the literature satisfy this property. In order-statistic games, agents' payoff depends only on their own action and an order statistic of all agents' actions. Examples include the minimum effort game Van Huyck et al. (1990) and the median effort game Van Huyck et al. (1991). Crawford (1995)'s and Robles (1997)'s summary-statistic payoff technologies allow payoffs to depend on order statistic or to convex combinations of order statistics. In the n-person stag hunt game by Kim (1996) and the total-effort game in Varian (2004), the relevant summary statistic is the number of agents who choose high effort. Following Devetag and Ortmann (2007), the term 'stag hunt game' is usually reserved for order-statistic games with two players and two actions. Aggregative games (Alos-Ferrer and Ania, 2005; Jensen, 2006, 2010) are summary-statistic games with continuous action spaces.

by concerns for strategic uncertainty and by historic precedent. Following Robles (1997) and Young (2001), best-reply learning is in line with these findings. The present model goes beyond best-reply learning in that it assumes that agents base beliefs on previous play even in a non-stationary environment with changing summary-statistics. Several experimental studies suggest that this is a plausible assumption. Van Huyck et al. (1991) let participants play sequences of median effort games with different payoff tables. They find that the median reached in later games is at least as high as in the initial game. Weber (2006) gradually increases the number of participants who are playing a minimum effort game and finds that this is a way to achieve coordination in large groups. Devetag (2005) finds that it is more likely that subjects coordinate on high effort in the minimum effort game if they first played a sequence of maximum effort games. Most relevant for this paper, Brandts and Cooper (2006) study whether temporary incentives can help subjects to coordinate on efficient equilibria. Participants play a sequence of minimal effort games and quickly converge to the lowest effort level. After some time, additional financial incentives are introduced and the authors observe a transition to the highest effort level. When they remove these incentives in the third part of the experiment, participants continue to coordinate on the highest effort level.

The use of best-reply learning allows to represent the principal's monitoring task as Markov decision process. The state $k \in \{0, 1, ..., n\}$ of this process is the number of agents who exert high effort and the principal's action $a \in \{0, 1, ..., n\}$ is the number of agents the principal monitors. Note that All-L corresponds to $k = 0$ and All-H corresponds to $k = n$. The principal's payoff increases in state k and decreases in action a.

Compared to a monitoring regime that would make high effort a dominant strategy once and for all, the optimal monitoring policy is '*minimally-invasive*'. That is, the principal either monitors just enough agents to make high effort a best-reply or does not monitor at all. As more and more agents exert high effort, strategic uncertainty decreases and less monitoring is needed. If most of the agents exert high effort, no monitoring is needed to make high effort a best-reply. Thus, there can be long stretches of time where the principal does not have to monitor at all.

If the principal's payoffs are supermodular and increase in k at an increasing rate, the optimal policy is monotone in k in the sense that if it is optimal for the principal to monitor *sufficiently* many agents to make high effort a best-reply in state k, making high effort a best-reply has also to be optimal in state $k + 1$.[3] Typically, such a policy leads to a quick convergence of play to All-L or All-H. To the contrary, a non-monotone policy can trap agents in intermediate states – that is, states of dis-coordination – for a long time.

The paper proceeds as follows. In section 2, the model framework is presented. Section 2.1 outlines a learning process without monitoring, section 2.2 introduces monitoring and section 2.3 discusses the relation to summary-statistic games. Section 3, characterizes the principal's optimization problem and shows that optimal monitoring policies are minimally-invasive and outlines sufficient conditions for the monotonicity of optimal policies. Section

[3]However, as the sufficient monitoring level decreases in k, the the actual monitoring level can still higher in k than in $k + 1$.

4 discusses the long-run implications of monotone and non-monotone monitoring policies. Section 5 discusses the robustness of the obtained results and concludes. Appendix A gives a brief overview to Markov chain theory and contains proofs from section 4, appendix B covers Markov decision processes and proofs for section 3.

2 Model

2.1 The basic best-reply dynamic

There is a population of n agents who repeatedly play a stage game according to a myopic best-reply rule. The stage game is a summary-statistic game where agents can either exert low effort (L) or high effort (H). Let $k \in \{0, 1, ..., n\}$ be the number of agents who choose high effort. Then All-L corresponds to $k = 0$ and All-H corresponds to $k = n$. Let $\gamma \in \{0, 1, ..., n-1\}$ be a threshold that represents the number of agents needed to make H a myopic best-reply. Then agents consider the best-reply rule[4]

$$Br(k) = \begin{cases} H & \text{if} \quad k > \gamma \\ L & \text{if} \quad k \leq \gamma \end{cases}, \tag{1}$$

stating that the agent should choose H if the fraction of agents choosing H is above γ and to choose L otherwise.

The best-reply dynamic evolves in discrete time $T = \{0, 1, ..., \infty\}$. In the initial period ($t = 0$), a strategy profile is randomly drawn from the set of all strategy profiles. In each subsequent period, one agent receives the chance to revise her strategy. Each agent is drawn with probability $1/n$. The agent drawn in period $t + 1$ observes k_t, the total number of agents who played high effort in the previous period and changes her strategy according to best-reply rule $Br(k_t)$ with probability $(1 - \epsilon)$ and by throwing a fair coin with probability $\epsilon \geq 0$. Afterwards, every agent exerts effort according to her strategy and stage payoffs are realized.

The evolution of the best-reply dynamic constitutes a random process. As agents' behavior depends only on the immediate past and not the entire history, this random process is a Markov chain that can be represented by a pair (K, P), where $K = (0, 1, ..., n-1)$ is the state space and $P : K \times K \to [0, 1]$ is a transition rule that specifies the probabilities of moving from one state to another at any period in time. Let $(K, P_{\gamma,\epsilon})$ be the Markov chain where the agent follows her best-reply rule with probability $(1 - \epsilon)$ and makes a mistake with probability $\epsilon \geq 0$. The Markov chain is called unperturbed if the mistake probability is zero and perturbed if there is a positive mistake probability. Transition probabilities are

[4]Section 2.3 discusses how this best-reply rule can be derived from agents' payoffs.

given by

$$p_{kj} = \begin{cases} (1-k/n)\, q_k & \text{if} \quad j = k+1 \\ (k/n)\, q_k + (1-k/n)\,(1-q_k) & \text{if} \quad j = k \\ (k/n)\,(1-q_k) & \text{if} \quad j = k-1 \end{cases}, \text{ where} \tag{2}$$

$$q_k = \begin{cases} 1-\epsilon/2 & \text{if} \quad k > \gamma \\ \epsilon/2 & \text{if} \quad k \leq \gamma \end{cases}, \tag{3}$$

is the probability that the agent chooses H given k. The probability of moving from state k to state $k+1$ (state $k-1$) is the probability of nature drawing an agent who has played L (H) in the previous period times the probability that this agent will switch to H (L) in the current period. With the sum of the counter-probabilities, the process remains in state k.

In the following, transition probabilities will be represented by a transition matrix, $P = (p_{kj})_{k,j=0}^{n}$, where entry p_{kj} gives the probability that the chain will be in state j in the next period, given that it is currently in state k. Hence rows of the transition matrix represent states in the current period and columns represent states in the next period. Note that as only one agent can revise her strategy in each period, the Markov chain can only move to adjacent states or remain in the same state. Thus, the transition matrix has a tridiagonal form.

Example 1. Transition matrix. Consider a population of $n = 4$ and let $\gamma = 2$. Then the transition matrix is given by

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 1-\epsilon/2 & \epsilon/2 & 0 & 0 & 0 \\ (2-\epsilon)/8 & (3-\epsilon)/4 & 3\epsilon/8 & 0 & 0 \\ 0 & (2-\epsilon)/4 & 1/2 & \epsilon/4 & 0 \\ 0 & 0 & 3\epsilon/8 & (3-\epsilon)/4 & (2-\epsilon)/8 \\ 0 & 0 & 0 & \epsilon/2 & 1-\epsilon/2 \end{pmatrix} \end{array}.$$

Thus, for example, the probability of moving to state 1 when currently in state 0 is $p_{01} = \epsilon/2$. Note that for states $k \leq 2$, most of the probability mass is on lower states, while for $k > 2$, more probability mass is on higher states. ▲

The expected movement of a Markov chain is given by the powers of the transition matrix. That is, if the chain starts in $t = 0$ in some state k, then the expected distribution of states in $t = 1$ is the k'th row of P, in $t = 2$ the k'th row of P^2 and in period τ the k'th row of P^τ.

2.2 The best-reply dynamic with monitoring

Suppose now that there is a principal whose payoff increases in agents total effort level and who can change agents' incentives by changing γ. Let the principal's action space be

$A = \{0, 1, ..., n\}$ with typical action $a \in A$ and let $g : A \to K$ be a function that determines the threshold in the agents' stage game depending on the principal's action. It is assumed that $g(0) = \gamma$, that $g(n) = -1$ – so the principal can always influence agents' behavior – and that $g(a)$ is weakly decreasing in the principal's action $a \in A$. Agents' best-reply rule now depends on k and $g(a)$ and takes the form

$$Br(k, g(a)) = \begin{cases} H & \text{if} \quad k > g(a) \\ L & \text{if} \quad k \leq g(a) \end{cases}, \tag{4}$$

For concreteness, the principal's action can be understood as the number of agents the principal monitors. This scenario will be further pursued in section 2.3.

The principal's stage payoffs $\pi(k, a)$ depend on the (current) state of the best-reply dynamic and the principal's (current) action. Stage payoffs are assumed to be weakly increasing in $k \in K$ and weakly decreasing in $a \in A$. Furthermore, they are bounded to $|\pi(k, a)| \leq \bar{\pi} < \infty$ for all $k \in K$ and $a \in A$ and do not change over time.

The principal has an infinite time horizon $T = \{0, 1, ..., \infty\}$ and discounts future payoffs with a discount factor of $\lambda \in (0, 1)$. In the initial period, a state $k_0 \in K$ is exogenously given. The principal observes k_0 and sets an action $a_0 \in A$. In every subsequent period $t + 1$, one agent, denoted agent $t + 1$, can revise her strategy. This agent observes the pair (k_t, a_t) and updates her strategy according to best-reply rule $Br(k_t, a_t)$ with probability $(1 - \epsilon)$ and randomizes uniformly between L and H with probability $\epsilon \geq 0$. Then all agents exert effort according to their strategies. The principal observes the new effort level $k_{t+1} \in K$ and sets $a_{t+1} \in A$. In the monitoring scenario, this means that the principal verifies the effort levels of a sample of a_{t+1} agents and punishes shirkers. Note that the principal does not know which agent revised her strategy and that every agent has the same chance of being sampled. Afterwards, stage payoffs are realized and the period ends. As agents respond only myopically to the principal's intervention, this framework is an Markov decision process (MDP).

A *decision rule* $d = (d(0), ...d(n))$ is a vector that assigns an action for every state of the Markov decision process. As discussed in Appendix B, it is without loss of generality to consider deterministic decision rules – i.e. the principal does not randomize between actions. Furthermore, we can restrict attention to Markov decision rules that are conditioned only on the current state of the process rather than the entire history of states. Let the set of deterministic Markov decision rules be given by $D = A^{|K|}$.

A *policy* $\delta = (d_1, d_2, ...)$ assigns a decision rule for every period $t \in T$. A stationary policy $d^\infty = (d, d, ...)$ assigns the same decision rule d in every period $t \in T$. The transition probabilities resulting from a deterministic decision rule can be summarized by a transition matrix P_d, with component $p(k, j) = p_d(j|k, d(k))$. Thus a stationary policy induces a stationary Markov chain $(K, P_{d,\epsilon})$. A non-stationary policy $\delta = (d_1, d_2, ...)$ induces a non-stationary Markov chain $(K, P_{\delta,\epsilon}) = (K, \{P_{d_1,\epsilon}, P_{d_2,\epsilon}, ...\})$.

The principal's optimization problem is to find a policy that maximizes her expected discounted total payoff (expected payoff for short). Together with this optimality criterion, the Markov decision process outlined above constitutes a '*Markov decision monitoring*

problem' (MDM problem) that can be summarized by a tuple $[K, A, p(j|k,a), \pi(k,a), \lambda]$, where K and A are the state and action space respectively, $p(j|k,a)$ are the transition probabilities, $\pi(k,a)$ the stage payoffs and λ is the principal's discount factor.

2.3 Derivation of the best-reply rules from agents' payoffs

In the following, I will show how the best-reply rules in (1) and (4) can be derived from summary-statistic games. Note that by 'summary statistic' Crawford (1995) and Robles (1997) mean order statistics and convex combinations of order statistic. The present paper defines this term more broadly. In particular, I allow payoffs to depend on the number of agents who choose high effort (k) as in the total-effort games discussed by (Varian, 2004), provided that this total-effort game yields the same best-reply rule as an order-statistic game.

An order-statistic payoff technology as in the baseline case of Robles (1997) implies that an agent's payoff depends *only* on her own effort level and an order statistic of aggregated effort. If $s \in \{L, H\}^n$ is the strategy profile played by the entire population, then $f_\ell(s) \in \{L, H\}$ is the ℓth lowest order statistic. For example, $f_1(s)$ is the minimum effort level, $f_{(n+1)/2}(s)$ is the median effort level and $f_n(s)$ is the maximum effort level in s. Note that if f_ℓ is the order-statistic of the agents' stage game, the threshold in agent's best-reply rule is given by

$$\gamma = n - \ell. \tag{5}$$

An agent's payoff from effort i given order statistic is f_ℓ is $U_\ell(i, f_\ell)$. It is assumed that $U_\ell(i,i) > U_\ell(i,j)$ for $j \neq i$, and that $U_\ell(L,L) < U_\ell(H,H)$. The first condition requires that it is always best for agents to match the relevant order statistic. The second condition implies that agents get a higher utility from coordinating on H than from coordinating on L. Thus, All-H Pareto dominates All-L. For order statistics $\{f_2, ...f_{n-1}\}$, unilateral deviation from All-L or All-H does not change the relevant order statistic. Thus All-L and All-H are the strict Nash equilibria of the games with these order statistics. For order statistic f_1, unilateral deviation from All-H changes the order statistic from H to L. Yet, as $U(L,L) < U(H,H)$, this does not benefit the agent. Hence, again, All-L and All-H are the two strict strategy Nash equilibria of this game. To the contrary, in the case of f_n, unilateral deviation from All-L changes the order statistic from L to H. Thus, $U(L,L) < U(H,H)$ implies that unilateral deviation from All-L is profitable. Therefore, All-H is the only Nash equilibrium.

Following Robles (1997), I assume that agents do not consider themselves pivotal. That is, they want to match the order-statistic established in the previous period, even if they would have the chance to change the relevant order statistic and would benefit from this change.[5] Specifically, I assume that if $k_t = \gamma$ and the agent who can revise her strategy in $t+1$ has played L in the previous period, this assumption implies that this agent chooses

[5]This assumption implies that agents' best-replies depend only on aggregate behavior, and not on their own personal history, which greatly simplifies the analysis. Naturally, this assumption is more plausible if there are many agents, so that the likelihood to be pivotal is small.

to play L again – even though by playing H the agent could change the order statistic to H. Hence for all order statistics $f_1, ...f_n$, the agents' best-reply to All-L is to play L and the best-reply to All-H is to play H. Following Robles (1997), I add two order statistics where the agents's best-replies do not depend on the behavior of other agents. Let f_0 denote the order statistic that is equal to L even if agents are in All-H, and let $\gamma = n - 0 = n$ be the corresponding threshold. Similarly, let f_{n+1} be the order statistic that is equal to H given All-L, leading to a threshold of $\gamma = n - (n+1) = -1$.

As mentioned, one natural explanation why the principal could be able to change threshold γ is that the principal can verify agents' effort levels ex-post. In order to explore this monitoring scenario, we have to extend our focus from order-statistic games to coordination games with a more general payoff structure that are behaviorally equivalent to an order-statistic game, in that they induce the same best-reply rule.

Let $\Gamma[n, S, U]$ denote a general symmetric coordination game where all agents have strategy space S and payoff matrix U. Let $\Gamma[n, S, U_\ell]$ denote a game where payoffs are determined by the ℓth order statistic. Then games $\Gamma[n, S, U]$ and $\Gamma[n, S, U_\ell]$ induce the same best-reply rule, if for all $s_i, s_i' \in S$ and $s_{-i} \in S^{n-1}$, $U(s_i, s_{-i}) > U(s_i', s_{-i}) \Leftrightarrow U_\ell(s_i, s_{-i}) > U_\ell(s_i', s_{-i})$. That is, holding other agents' strategies fixed, if an agent gets a higher utility from playing L (H) than from playing H (L) in game $\Gamma[n, S, U]$, then this must also be true in game $\Gamma[n, S, U_\ell]$.

Example 2 illustrates the equivalence between a total-effort game and an order-statistic game. Example 3 shows how this total-effort game can be modified by monitoring agents such that the modified game is again equivalent to an order-statistic game.

Example 2. Agents' payoffs in the underlying coordination game. Consider a population of $n = 4$ agents playing symmetric coordination game $\Gamma[4, \{L, H\}, U]$ and let their utilities be given by the payoff matrix

$$U = \begin{array}{c} \\ L \\ H \end{array} \begin{array}{c} \begin{array}{cccccccc} LLL & LLH & LHL & HLL & LHH & HLH & HHL & HHH \end{array} \\ \begin{pmatrix} 5 & 10 & 10 & 10 & 15 & 15 & 15 & 20 \\ 2 & 7 & 7 & 7 & 12 & 12 & 12 & 25 \end{pmatrix} \end{array}.$$

Note that agents do not care about the identities of other players, so the payoff matrix can be shortened to

$$U = \begin{array}{c} \\ L \\ H \end{array} \begin{array}{c} \begin{array}{cccc} 0H & 1H & 2H & 3H \end{array} \\ \begin{pmatrix} 5 & 10 & 15 & 20 \\ 2 & 7 & 12 & 25 \end{pmatrix} \end{array}, \tag{6}$$

where 0H,...,3H indicate how much other agents exert high effort. Apparently, agents' utility depends on their own effort level and on the number of agents who choose high effort, so this is not a game with order-statistic payoff technology. However, this game has an equivalent payoff structure to a minimum effort game - that is, to a game with $\ell = 1$ and $\gamma = 4 - 1 = 3$, and a payoff matrix such as

$$U_2 = \begin{array}{c} \\ L \\ H \end{array} \begin{array}{c} \begin{array}{cccc} 0H & 1H & 2H & 3H \end{array} \\ \begin{pmatrix} 5 & 5 & 5 & 20 \\ 2 & 2 & 2 & 25 \end{pmatrix} \end{array}.$$

To see that $\ell = 1$ is the relevant order statistic, note that $f_1(H|HHH) = H$ while $f_1(H|LHH) = f_1(L|HHH) = L$. Thus, given that all others play H, the agent's action always matches the order statistic. As $U_\ell(H|H) > U_\ell(L|L)$ by assumption, the agent prefers to play H in that case. In any other case, the best-reply is to play L. ▲

Suppose now that the principal can verify agents' effort after observing the aggregate effort level.

Example 3. Agents' payoffs in the MDM problem. Consider again the game $\Gamma[4, \{L, H\}, U]$, with the payoff matrix in equation 6. Let $A = K = \{0, 1, ..., 4\}$ and suppose that the principal verifies effort in a sample of $a \in A$, detects all agents who exert low effort in this sample and can punish shirkers by taking away their stage payoff. Let U' be agents' payoff matrix when they are audited for sure

$$U' = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 2 & 7 & 12 & 25 \end{pmatrix},$$

so that agents' expected utility for choice $a \in A$ is $U'' = (1 - a/n) U + (a/n) U'$, or

$$U'' = \begin{pmatrix} 5(1-a/n) & 10(1-a/n) & 15(1-a/n) & 20(1-a/n) \\ 2 & 7 & 12 & 25 \end{pmatrix}.$$

Consider an agent whose opponents all play L. Then the agent would be indifferent between playing L and H if $5(1 - a/4) = 2$ or $a = 12/5 \notin A$, and thus strictly prefers to play H if $a \in \{3, 4\}$ and to play L if $a \in \{0, 1, 2\}$. If one other agent plays H, the agent would be indifferent if $10(1 - a/4) = 7$ or $a = 6/5 \notin A$, so the agent strictly prefers to play H if $a \in \{2, 3, 4\}$ and to play L if $a \in \{0, 1\}$. If two other agents play H, the agent would be indifferent at $15(1 - a/4) = 12$ or $a = 4/5 \notin A$, so she strictly prefers to play H if $a \in \{1, 2, 3, 4\}$ and L if $a = 0$. Finally, if all opponents play H, the agent prefers to play H for all $a \in A$.

Recall however that agents always match the order statistic in the previous period, even if they can change the order statistic in their favor. To allow for the case where agents prefer to choose H even if the entire population - including themselves - played L in the previous period, suppose that agents strictly prefer to play H if $a = 4$ even in state $k = 0$. Then thresholds are given by the function $\gamma(a) = 3 - a$. Agents' best-reply rule can now directly be expressed as a function of states and actions

$$Br(k, a) = \begin{cases} H & \text{if} \quad k > 3 - a \\ L & \text{if} \quad k \leq 3 - a \end{cases},$$

rather than as function of k and $g(a)$. ▲

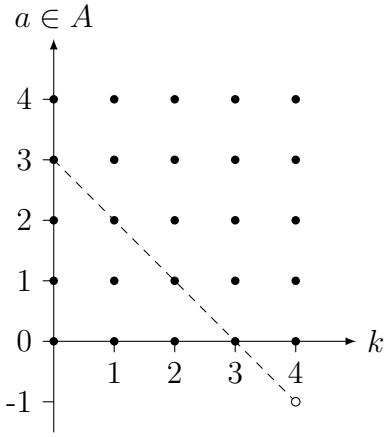

Figure 1. Best-replies in state-action space (k, a) for Example 3. The filled circles indicate feasible state-action pairs (k, a). The dashed line is the inverse of the threshold function $g(a) = 3 - a$. For pairs on or below the dashed line, that is, for (k, a) with $k \leq 3 - a$, the agents' best-reply is to play L, for pairs above the dashed line, the best-reply is H.

3 Optimal policies – computation and properties

In this section I introduce the main solution concept for infinite horizon Markov decision problems and characterize the set of optimal MDM policies. As mentioned, we can restrict attention to deterministic Markov decision rules $d \in D$. The principal's stage payoff when using such a decision rule is given by the payoff vector $\pi_d = (\pi_d(0), ..., \pi_d(n))^\top$. The principal's expected payoff from policy d^∞ and discount factor $\lambda \in (0, 1)$ is the column vector $v_\lambda^{d^\infty} = \{v_\lambda^{d^\infty}(k)\}_{k \in K}$ given by

$$\begin{aligned} v_\lambda^{d^\infty} &= \pi_d + \lambda P_d \pi_d + \lambda^2 P_d^2 \pi_d + ... \qquad (7) \\ &= \pi_d + \lambda P_d(\pi_d + \lambda P_d \pi_d + +\lambda^2 P_d^2 \pi_d + ...) \\ &= \pi_d + \lambda P_d v_\lambda^{d^\infty}. \qquad (8) \end{aligned}$$

Let V denote the set of expected payoffs and let

$$L_d v = \pi_d + \lambda P_d v \qquad (9)$$

be a linear transformation on V. Note that $L_d v \in V$ and $v_\lambda^{d^\infty} = (I - \lambda P_d)^{-1} \pi_d$ is the unique solution to $L_d v = v$. Consider the operator

$$Lv = \max_{d \in D} \{\pi_d + \lambda P_d v\}. \qquad (10)$$

and note that the equation $Lv = v$ is known as *optimality equation* or Bellman equation. As discussed in Appendix B.2, a policy is optimal if and only if it satisfies the optimality equation. This results suggests a fast algorithm to find the maximal expected payoff. The *value-iteration algorithm* selects an initial $v^0 \in V$ and repeatedly applies (10) until the optimality equation is satisfied.

Lemma 1. *There exists an optimal MDM policy that is stationary. That is,* $\delta^* = (d^*)^\infty$ *with* $d^* \in D$ *and* $v^* = v^{(d^*)^\infty} \in V$.

The proof is given in Appendix B.2. Intuitively, the lemma holds because given that K and A are finite, there must exist a decision rule $d \in D$ that satisfies the optimality equation. For future reference, note that the value-iteration algorithm can be applied component-wise. Let

$$w(k,a) = \pi(k,a) + \lambda \sum_{j=0}^{n} p(j|k,a)v(j). \tag{11}$$

Then $L_{d(k)}v(k) = w(k,d(k))$ is the k'th component of $L_d v$ and $Lv(k) = \max_{a\in A}\{w(k,a)\}$ is the k'th element of Lv. Furthermore, the optimality equation is satisfied if $Lv(k) = v(k)$ for all $k \in K$.

3.1 Optimal policies I: Minimally-invasive policies

In the proposed monitoring problem an increase in monitoring does not always change agents' best-reply rule. As monitoring is costly, we might suspect that for every state $k \in K$, the majority of monitoring levels is sub-optimal. In the language of Markov decision theory, this means that the optimal policy is structured – all candidates for optimality come from a strict subset of decision space D. Let $D^\sigma \subset D$ denote the set of *structured decision* rules and $V^\sigma \subseteq V$ the set of *structured values*. The optimality of a structure policy is established by the following lemma.

Lemma 2. (Puterman (2005), Theorem 6.11.1) *Suppose that*

a) $v \in V^\sigma$ *implies* $Lv \in V^\sigma$*;*
b) $v \in V^\sigma$ *implies there exists a* $d' \in D^\sigma \cap argmax_{d\in D} L_d v$*; and*
c) V^σ *is a closed subset of* V*, that is, for any convergent sequence* $\{v^\tau\} \subset V^\sigma$, $\lim_{\tau\to\infty} v^\tau \in V^\sigma$.

Then there exists an optimal policy $\delta^* = (d^*)^\infty$ *that is structured* ($d^* \in D^\sigma$) *and optimal values are structured as well* ($v^* \in V^\sigma$).

Lemma 2 can be applied to MDM problems as follows. Let a_k be the smallest action $a \in A$ for a given state $k \in K$ that makes playing H a best-reply. Note that a_k can be equal to zero. Formally, let $\hat{A}_k = \{a \in A : k > g(a) \text{for} k \in K\}$ and define $a_k = \{a_k, a'_k \in \hat{A}_k : a_k \leq a'_k\}$ and $A_k = \{0, a_k\}$. Note that a_k can be equal to zero, which makes A_k a singleton. In our case, let the set of structured policies be

$$D^S = A_0 \times ... \times A_n \tag{12}$$

and suppose that the set of structured values is given by the set of all weakly increasing functions on k,

$$V^S = \{v \in V : \forall k \in K\backslash\{n\},\ v(k) \leq v(k+1)\}. \tag{13}$$

The following proposition states that there exists an optimal policy that is structured. In particular, this policy is *'minimally-invasive'* in that for every state $k \in K$ it is either optimal for the principal to monitor just enough to make high effort a best-reply ($a(k) = a_k$) or not to monitor at all ($a(k) = 0$).

Proposition 1. (Minimally-invasive policies) *There exists an optimal MDM policy that is minimally-invasive. That is, $\delta^* = (d^*)^\infty$ with $d^* \in D^S$ and $v^* \in V^S$.*

Proof. We need to show that assumptions a)–c) in Lemma 2 are satisfied for structured policies D^S and structured values V^S.

ad a) (i) By Lemma 10, $\sum_{j=0}^{n} p(j|k,a)v(j)$ is non-decreasing in k. By assumption, $\pi(k,a)$ is increasing in k and decreasing in a. (ii) The sum of two increasing functions is increasing, so for every $a \in A$,

$$w(k,a) = \pi(k,a) + \lambda \sum_{j=0}^{n} p(j|k,a)v(j) \tag{14}$$

is increasing in $k \in K$. (iii) The maximum over a set of increasing functions is increasing, so $\max_{a \in A'}\{w(k,a)\}$ is increasing in $k \in K$ and therefore $Lv \in V^S$.

ad b) By definition, $\pi(k,a)$ is weakly decreasing in $a \in A$. In particular, $\pi(k,a') \leq \pi(k,0)$ for $0 < a' < a_k$. However, the agents' best-reply is to play L in both cases, which implies that $\sum_{j=0}^{n} p(j|k,a')v(j) = \sum_{j=0}^{n} p(j|k,0)v(j)$. Thus $w(k,a') \leq w(k,0)$. Similar reasoning reveals that for $a_k < a'' \leq n$, $\pi(k,a'') \leq \pi(k,a_k)$ while the agents' best-reply is to play H in both cases, so that $w(k,a'') \leq w(k,a_k)$. Thus, for all $v \in V^S$ there exists a decision rule with $d(k) \in \{0, a_k\}$ that maximizes $L_d v$ as required.

ad c) Lemma 11 states that V^S is a closed subset of V. ■

3.2 The reduced MDM problem

Proposition 1 also tells us that the principal's optimization problem can be reduced to the binary choice problem of when to monitor and when to abstain from monitoring. To see this more clearly, consider the following modification.

Suppose the principal's action space is $A' = \{a^-, a^+\}$ with $a^- < a^+$ and suppose that agents always consider it a best-reply to play L if the principal chooses a^- and to play H if the principal chooses a^+. That is, let the agent's best-reply rule be defined independently of k by

$$Br\,(a) = \begin{cases} H & \text{if} \quad a = a^+ \\ L & \text{if} \quad a = a^- \end{cases}. \tag{15}$$

Obviously, for $k \leq g(0)$, $a^- = 0$ and $a^+ = a_k$ and payoffs are given by $\pi(k,a^-) = \pi(k,0)$ and $\pi(k,a^+) = \pi(k,a_k)$. For $k > g(0) = \gamma$, we have $a_k = 0$. Suppose there is a hypothetical action a^- that nevertheless makes it a best-reply for the agents to choose L and suppose that the payoff of this hypothetical action is given by $\pi(k,a^-) = \pi(k,a^+) = \pi(k,a_k)$.

Let $D' = \{A'\}^{n+1}$ and let D^R be the set of decision rules $d' \in D'$ that are equivalent to a decision rule in $d \in D^S$ in that for $k \leq \gamma$ it holds that $d'(k) = a^+$ whenever $d(k) = a_k$ and $d'(k) = a^-$ whenever $d(k) = 0$ and in that for $k > \gamma$, $d'(k) = a^+$. Note that for every decision rule $d \in D^S$, there is exactly one decision rule $d' \in D^R$ that induces the same best-reply behavior and the same payoffs.

Let the monitoring problem where the principal has action space A' and restricts attention to decision rules D^R be called the *reduced MDM* problem. Formally, this problem is defined by the tuple $[K, A', p\left(j|k,a\right), \pi\left(k,a\right), \lambda]$.

Lemma 3. *It is without loss of generality to consider the reduced MDM problem.*

Proof. We need to show that (i) the optimal stationary policy in the reduced MDM problem uses a decision rule $d' \in D^R$ and (ii) that this decision rule is equivalent to the optimal decision rule of the general MDM problem $d^* \in D^S$.

(i) For states $k > \gamma$ we have introduced the hypothetical action a^- with the properties that playing L is a best-reply. As $v(j)$ is weakly increasing in $j \in K$ by Proposition 1, and a^+ shifts more probability mass on high states than a^-, we can conclude that $\lambda \sum_{j=0}^{n} p(j|k,a^-)v(j) \leq \lambda \sum_{j=0}^{n} p(j|k,a^+)v(j)$. Therefore, $w(k,a^-) \leq w(k,a^+)$ for all $k > \gamma$ and we can restrict attention to decision rules $d' \in D^R \subseteq D'$.

(ii) According to Puterman (2005, Theorem 6.1.1), there is a unique decision rule $d^* \in D^S$ that satisfies $L_{d^*}v^* = v^*$ and by Lemma 8, the stationary policy $\delta^* = (d^*)^\infty$ is optimal. By design, for every decision rule $d^* \in D^S$ there is an equivalent decision rule $d' \in D^R$ such that $L_{d'}v = L_{d^*}v$. It follows that d' satisfies $L_{d'}v = v$ and thus, by Lemma 8, the stationary policy $\delta' = (d')^\infty$ has to be optimal. ■

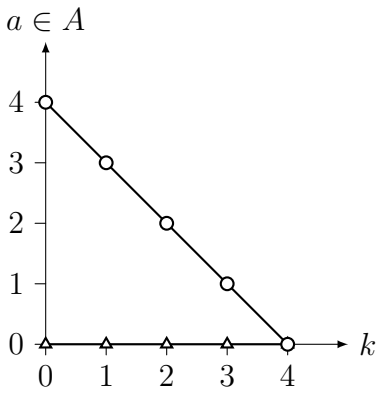

Figure 2.1

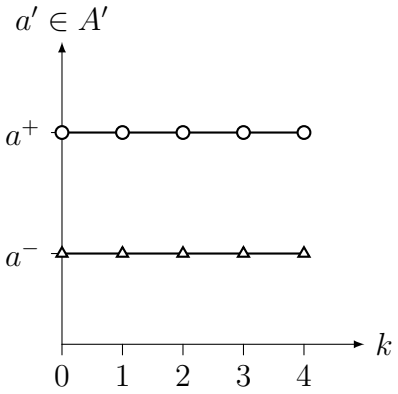

Figure 2.2

Figure 2. Equivalent decision rules $d \in D^S$ and $d' \in D'$. Figures 2.1 and 2.2 depict decisions on (k,a) space and (k,a') space respectively. Thereby, decision rules $d = (0,...,0)$ (triangles and the rightmost circle) and $d = (a_0,...,a_4)$ (circles) in Figure 2.1 are equivalent to $d' = (a^-,...,a^-)$ (triangles) and $d' = (a^+,...,a^+)$ (circles) in Figure 2.2.

Example 4. MDM and reduced MDM problem. Recall example 3 with $n = 4$ and $g(a) = 3 - a$ for $a \in A$. Then the lowest action sufficient to make playing H a best-reply satisfies $k > 3 - a$ which implies that $a_k = 4 - k$. Suppose that $k = 2$ and note that $a_2 = 4 - 2 = 2$. Let stage payoffs be given by

$$\pi(k, a) = k^2/4 - a$$

Note that this function is increasing in states $k \in K$ and decreasing in actions $a \in A$. Note that for all $0 < a' < a_k < a''$,

$$k^2/4 - a'' < k^2/4 - a_k < k^2/4 - a' < k^2/4,$$

while expected future payoffs are equivalent for actions a' and 0 and for actions a'' and a_k:

$$\sum_{j=0}^{n} p(j|k, a')v(j) = \sum_{j=0}^{n} p(j|k, 0)v(j), \text{ and}$$
$$\sum_{j=0}^{n} p(j|k, a'')v(j) = \sum_{j=0}^{n} p(j|k, a_k)v(j).$$

It follows that $w(k, a') < w(k, 0)$ and $w(k, a'') < w(k, a_k)$. Thus, optimal decisions $d^*(k)$ are in $A_k = \{0, a_k\}$ for each $k \in K$ as claimed by Proposition 1.

The MDM problem can be translated into a reduced MDM problem by letting $a^- = 0$ and $a^+ = 1$ and finding an equivalent payoff function. Note that $\pi(k, a_k) = k^2/4 - (4 - k)$. Thus, if we set

$$\pi(k, a') = k^2/4 - (4 - k)a'$$

we obtain $\pi(k, a^-) = k^2/4 = \pi(k, 0)$ and $\pi(k, a^+) = k^2/4 - (4 - k) = \pi(k, a_k)$, as required.

Figure 2 presents decision rules in structured decision space D^S and their equivalents in reduced decision space D^R. ▲

3.3 Optimal policies II: Monotone policies

In section 3.2 we have seen that the principal's optimization problem boils down to the question of when to monitor agents at all and when to 'give up' and abstain from monitoring. This formulation raises the additional question whether the optimal decision rule is monotone in the sense that agents are always monitored in state $k+1$ if they are monitored in state k.

Assumption 1. Stage payoff $\pi(k, a)$ is supermodular on $K \times A'$, meaning that for all $k^+ > k^-$ and for $a^+ > a^-$, it holds that

$$\pi(k^+, a^+) - \pi(k^+, a^-) \geq \pi(k^-, a^+) - \pi(k^+, a^-). \tag{16}$$

Intuitively, a function is supermodular if it has monotone increasing differences. Note however that $\pi(k,a)$ is assumed to be decreasing in the principal's actions. Therefore, $\pi(k,a^-) - \pi(k,a^+)$ is positive and supermodularity entails that the distance $\pi(k,a^-) - \pi(k,a^+)$ is decreasing in k.

Assumption 2. Stage payoff $\pi(k,a)$ is increasing at an increasing rate (IIR) in $k \in K$. That is, for all $k \in K\backslash\{0,n\}$ and $a \in A'$, it holds that

$$\pi(k+1,a) + \pi(k-1,a) \geq 2\pi(k,a). \tag{17}$$

Note that IIR can be thought of as discrete equivalent to convexity. Let the set of monotone decision rules be given by

$$D^M = \left\{ d \in D^R : \forall k \in K\backslash\{n\},\, d(k) \leq d(k+1) \right\}, \tag{18}$$

and let the set of 'monotone' values be given by the set of values that are (strictly) increasing in k at an (strictly) increasing rate,

$$V^M = \left\{ v \in V : \forall k \in K\backslash\{0,n\},\, v(k+1) + v(k-1) \geq 2v(k) \right\}. \tag{19}$$

The following proposition provides sufficient conditions for the optimality of monotone policies.

Proposition 2. (Monotone policies) *If assumptions 1 and 2 are satisfied, there exists an optimal MDM policy that is monotone. That is, $\delta^* = (d^*)^\infty$ with $d^* \in D^M$ and $v^* \in V^M$.*

Proof. Again we need to show that assumptions a)–c) in Lemma 2 are satisfied , this time for monotone policies $d \in D^M$ and monotone values $v \in V^M$.

ad a) (i) According to Lemma 12, $\sum_{j=0}^n p(j|k,a)v(j)$ is IIR in $k \in K$ if $v(j)$ is IIR in $j \in K$. Due to assumption (2) holds, $\pi(k,a)$ is IIR as well. (ii) The sum of two functions that are IIR is IIR and therefore,

$$w(k,a) = \pi(k,a) + \lambda \sum_{j=0}^{n} p(j,k,a)v(j) \tag{20}$$

is IIR in $k \in K$. (iii) Furthermore, the maximum of a set of functions that are IIR is also IIR. In particular, $max_{a\in A'}\{w(k,a)\}$ is IIR for all $k \in K$ and therefore, $Lv \in V^M$.

ad b) (i) By Lemma 13, $\sum_{j=0}^n p(j,k,a)v(j)$ is supermodular on $K \times A'$ if $v \in V^M$. By assumption 1, $\pi(k,a)$ is supermodular on $K \times A'$ as well. (ii) The sum of two supermodular functions is supermodular, so $w(k,a)$ is supermodular. (iii) A supermodular function has the property that it's maximizing arguments are monotone increasing. Thus, $d^*(k) \in \text{argmax}_{a\in A'}\{w(k,a)\}$ is monotone increasing in $d^*(k)$.

ad c) Lemma 14 states that V^M is closed in V. ■

Section 3.5 will discuss several examples for monotone and non-monotone optimal policies.

3.4 Comparative statics

For the reduced MDM problem, it is possible to establish monotone comparative statics for the exogenous parameters of the model.

Proposition 3. (Comparative Statics) *Suppose the optimal policy is monotone. Then the optimal monitoring rule $d^*(.)$ (i) is weakly increasing in discount rate λ, (ii) weakly decreasing in error rate ϵ, (iii) weakly decreasing in threshold γ and (iv) weakly increasing in population size n.*

Proof of Proposition 3. Note that for every $k \in K$, the principal weakly prefers a^+ if $w(k, a^+) \geq w(k, a^-)$. This weak inequality can be rearranged to

$$\begin{aligned}
\pi(k, a^+) + \lambda \sum_{j=0}^{n} p(j|k, a^+)v(j) &\geq \pi(k, a^-) + \lambda \sum_{j=0}^{n} p(j|k, a^-)v(j) \\
\lambda \sum_{j=0}^{n} \Big(p(j|k, a^+) - p(j|k, a^-) \Big) v(j) &\geq \pi(k, a^-) - \pi(k, a^+) \\
\lambda(1-\epsilon) \Big(\tfrac{k}{n}(v_k - v_{k-1}) + (1 - \tfrac{k}{n})(v_{k+1} - v_k) \Big) &\geq \pi(k, a^-) - \pi(k, a^+), \qquad (21)
\end{aligned}$$

where $v_k = v(k)$. Note that (21) represents the principal's trade-off between current payoffs (on the right hand side (RHS)) and future payoffs (on the left hand side (LHS)). As $\pi(k, a)$ is decreasing in $a \in A'$, we know that the RHS is non-negative. From Lemma 10, we know that $\sum_{j=0}^{n} p(j|k, a)v(j)$ is monotone increasing in $k \in K$ if $v(j)$ is monotone increasing in $j \in K$. Proposition 1 states that this is indeed the case. Hence, the LHS is non-negative as well.

ad i) Note that increasing λ increases the LHS, while the RHS remains unaffected. Therefore, the optimal decision rule $d^*(.)$ has to be weakly increasing in λ.

ad ii) Increasing ϵ decreases the LHS, while the RHS remains unaffected. Therefore, the optimal decision rule $d^*(.)$ has to be weakly decreasing in ϵ.

ad iii) Increasing γ does not affect the LHS but weakly increases the RHS, because the number of states for with $\pi(k, a^+) = \pi(k, a^+)$ is weakly decreasing in γ. Hence the optimal decision rule $d^*(.)$ has to be weakly decreasing in γ.

ad iv) If the optimal policy is monotone, optimal values are increasing in k at an increasing rate. As $(1 - k/n)$ increases in n, it follows that the LHS of (21) is increasing in n. As stage payoffs do not (directly) depend on n the RHS is unaffected for a given $k \in K$. Thus, $d^*(.)$ should be weakly increasing in n. ■

The intuition for the monotonicity of population size n is as follows: For a given k, an increase of the population size from n to $n' > n$ implies that a lower fraction of agents exerts high effort. Therefore, if high effort is the agents' best-reply, a larger fraction of agents can switch. For a given k, the transition becomes faster. Note however, that the proof of Proposition 3 implicitly assumes that the principal stage payoffs are increasing in total effort. Thus, an increase in population size n increases the principal's maximal payoffs $\pi(n, a)$ for $a \in A'$. However, it could be equally plausible that the principal's

payoff is increasing in the fraction of agents who exert high effort, $\pi(k/n, a)$. But then, $\pi(k/n, a^-) - \pi(k/n, a^+)$ and optimal $v(k/n)$ would also depend on n, thus making general statements on the effect of an increase of n impossible.

3.5 Examples

This section presents examples for monotone and non-monotone decision rules.

Example 5. A reduced MDM problem with monotone optimal policy. Consider the MDM problem outlined in 4 with $n = 4$, action space $A' = \{0, 1\}$ and stage payoffs be $\pi(k, a) = k^2/4 - (4 - k)a$. Plugging into (16) yields

$$\begin{aligned} \pi'(k+1,1) - \pi'(k+1,0) &\geq \pi'(k,1) - \pi'(k,0) \\ (k+1)^2/4 - (4-(k+1)) - ((k+1)^2/4) &\geq k^2/4 - (4-k) - (k^2/4) \\ -(4-k-1) &\geq -(4-k) \\ k+1 &\geq k, \end{aligned}$$

for all $k \in K\backslash\{n\}$, so assumption(1) is satisfied. Plugging into equation (17) yields

$$\begin{aligned} \pi'(k+1,a) + \pi'(k-1,a) \geq 2\pi'(k,a) &\text{ can be reduced to} \\ (k+1)^2 + (k-1)^2 &\geq 2k^2 \\ 2k^2 + 2 &\geq 2k^2 \end{aligned}$$

for all $k \in K\backslash\{0, n\}$ and $a \in A'$, so assumption (2) is satisfied as well. Thus, the optimal policy has to be monotone. Applying the value-iterations algorithm as outlined in Appendix B with $\epsilon = 0.01$, $\lambda = 0.7$ and convergence criterion $\eta = 0.01$ yields an optimal stationary policy $(d^*)^\infty$ with monotone conserving decision rule $d^*_{\lambda=0.7,\epsilon=0.01} = (a^-, a^-, a^+, a^+, a^+)$, as depicted in Figure 3.2. The optimal values (rounded to two digits) for these parameters are $v^*_{0.7,0.01} = (0.01, 0.54, 2.47, 7.48, 13.27)$.

Changing discount factor λ to 0.6 or 0.8 yields $d^*_{0.6,0.01} = (a^-, a^-, a^-, a^+, a^+)$ and $d^*_{0.8,0.01} = (a^-, a^+, a^+, a^+, a^+)$ respectively, so $d^*(k, \lambda, \epsilon)$ is indeed (weakly) increasing in λ as predicted by Proposition 3. Furthermore, changing mistake probability ϵ to 0.001 or 0.1 yields $d^*_{0.7,0.001} = d^*_{0.7,0.01} = (a^-, a^-, a^+, a^+, a^+)$ and $d^*_{0.7,0.1} = (a^-, a^-, a^-, a^+, a^+)$, so $d^*(k, \lambda, \epsilon)$ is indeed (weakly) decreasing in ϵ in this example. ▲

Example 6. A reduced MDM problem with non-monotone optimal policy. Consider a reduced MDM problem with $n = 4$, action space $A' = \{0, 1\}$ and stage payoffs $\pi(k, a) = k^{1/2} - a$ for $k \leq 3$ and $\pi(k, a) = k^{1/2}$ for $k = 4$. Note that for all $k \in K\backslash\{0, n-1, n\}$ and $a \in A'$,

$$((k+1)^{1/2} - a) + ((k-1)^{1/2} - a) < 2(k^{1/2} - a)$$

so Assumption (2) is violated. Applying the value-iterations algorithm with $\epsilon = 0.01$, $\lambda = 0.7$ and $\eta = 0.01$ yields an optimal stationary policy $(d^*)^\infty$ with non-monotone decision rule $d^*_{\lambda=0.7,\epsilon=0.01} = (a^+, a^-, a^-, a^-, a^+)$, as depicted in 3.4. ▲

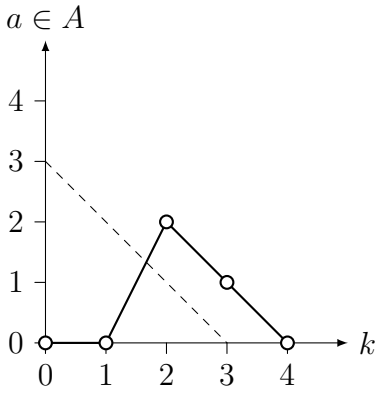

Figure 3.1

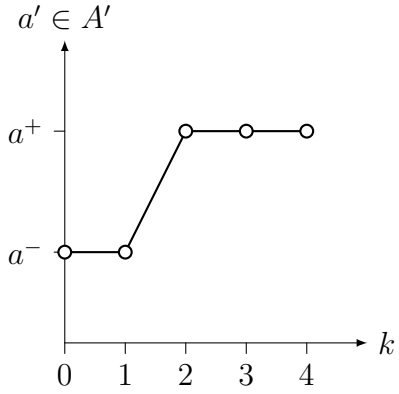

Figure 3.2

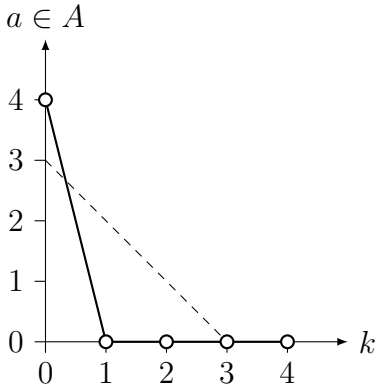

Figure 3.3

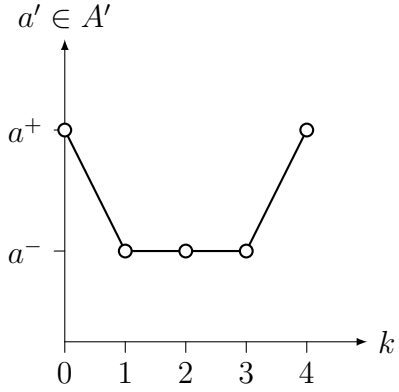

Figure 3.4

Figure 3. Monotone and non-monotone decision rules. Figure 3.2 depicts the *monotone* decision rule $d' = (a^-, a^-, a^+, a^+, a^+) \in D^M \subset D^R$. Figure 3.1 presents the equivalent decision rule on $d = (0, 0, a_2, a_3, a_4) = (0, 0, 2, 1, 0) \in D^S$.

To the contrary, Figure 3.4 depicts the *non-monotone* decision rule $d' = (a^+, a^-, a^-, a^-, a^+) \in D^R$. Decision rule $d = (a_0, 0, 0, 0, a_4) = (4, 0, 0, 0, 0) \in D^S$ in Figure 3.3 is equivalent. The dashed lines indicate the inverse of the threshold function $g(a) = 3 - a$ derived in Example 3.

Figure 3 illustrates monotone and non-monotone decision rules. Note that monotonicity is required for decision rules defined on *reduced* action space A', but not for their equivalent decision rules defined on A. Thus, a monotone decision rule defined on A' can be equivalent to a decision rule defined on A that is neither increasing nor decreasing (Figures 3.1 and Figures 3.2). To the contrary, a non-monotone decision rule on A' can be equivalent to a monotone decision rule on A.

4 Long-run implications of optimal policies

This section studies the long-run behavior induced by optimal policies. An optimal MDM policy induces a (regular) Markov chain $(K, P_{\gamma,\epsilon>0})$ that has a unique stationary distribution α that is approached from every initial distribution $\mu_0 \in \mathcal{P}(K)$, where $\mathcal{P}(K)$ be the set of probability vectors on K. The following lemma states that the stationary distribution can be computed by a simple matrix operation. The proofs for this section are given in Appendix A.1.

Lemma 4. (Grimmett and Stirzaker (2001), Ex. 6.6.5) *The unique stationary distribution of a regular stationary Markov chain $(K, P_{\gamma,\epsilon>0})$ is given by*

$$\alpha = e^{\top} (I - P + E)^{-1}, \tag{22}$$

where I is the identity matrix of appropriate dimension, E is a matrix with all entries one and e is a column vector with all entries one. Young (1993) shows that a regular Markov chain also has a stationary distribution when ϵ approaches zero. This so-called limit invariant distribution is given by $\alpha^* = \lim_{\epsilon \to 0} \alpha^{\epsilon}$. Stochastic stability is a qualitative feature of the limit invariant distribution. The set of stochastically stable states $K^* \in K$ is the set of all states that occur with positive probability in the limit invariant distribution,

$$K^* = \{k \in K : \alpha^*(k) > 0\}. \tag{23}$$

The set of stochastically stable states can be derived with the graph-theoretic methods described in Appendix A. Importantly, the set of stochastically stable states of a perturbed Markov chain is a subset of the set of recurrent classes (the limit set) of the unperturbed Markov chain $(K, P_{d,\epsilon=0})$.

Suppose the principal uses a monotone decision rule $d \in D^M$ and let $\gamma^d \in \{-1, 0, 1, ..., n\}$ be the highest state where agents' best-reply is to play L. The following proposition is a special case of Proposition 3.3 in Robles (1997).

Proposition 4. (Stochastic stability) *Consider an unperturbed Markov chain $(K, P_{d,\epsilon=0})$ induced by a monotone decision rule $d \in D^M$ with $\gamma^d \in \{-1, 0, 1, ..., n\}$. Then the set of stochastically stable states is*

$$K^* = \begin{cases} \{0\} & \text{if} \quad \gamma_d > (n-1)/2 \\ \{0, n\} & \text{if} \quad \gamma_d = (n-1)/2 \\ \{n\} & \text{if} \quad \gamma_d < (n-1)/2 \end{cases}. \tag{24}$$

The following example derives the stationary distribution of a Markov chain induced by a monotone policy.

Example 7. Let $n = 4$ and consider the Markov chain $(K, P_{d,\epsilon>0})$ induced by a monotone decision rule $d \in D^M$ with $\gamma_d = 1$. Note that example 5 has the optimal decision

rule $d^*_{\lambda=0.8,\epsilon=0.01} = (a^-, a^+, a^+, a^+, a^+)$ or a $\gamma_d = 1$. Following Lemma 4, the stationary distribution of induced Markov chain is given by

$$\alpha = \left(\frac{2\epsilon - \epsilon^2}{4 + 6\epsilon + 6\epsilon^2}, \frac{4\epsilon^2}{4 + 6\epsilon + 6\epsilon^2}, \frac{6\epsilon^2}{4 + 6\epsilon + 6\epsilon^2}, \frac{8\epsilon - 4\epsilon^2}{4 + 6\epsilon + 6\epsilon^2}, \frac{4 - 4\epsilon + \epsilon^2}{4 + 6\epsilon + 6\epsilon^2} \right).$$

Clearly, $\alpha^* = \lim_{\epsilon \to 0} \alpha = (0, 0, 0, 0, 1)$, meaning that state 0 (or All-L) is stochastically stable. ▲

If the principal chooses a non-monotone decision rule, there can be stochastically stable states that are not in $\{0, n\}$. This is illustrated in the following example.

Example 8. Let $n = 4$ and consider a Markov chain $(K, P_{d,\epsilon>0})$ induced by the non-monotone decision rule $d = (a^+, a^-, a^-, a^-, a^+)$ derived in example 6. The stationary distribution is

$$\alpha = \left(\frac{4 - 4\epsilon + \epsilon^2}{20 - 8\epsilon + 4\epsilon^2}, \frac{16 - 16\epsilon + 4\epsilon^2}{20 - 8\epsilon + 4\epsilon^2}, \frac{12\epsilon - 6\epsilon^2}{20 - 8\epsilon + 4\epsilon^2}, \frac{4\epsilon^2}{20 - 8\epsilon + 4\epsilon^2}, \frac{\epsilon^2}{20 - 8\epsilon + 4\epsilon^2} \right),$$

so that the limit invariant distribution is given by $\alpha^* = (0.2, 0.8, 0, 0, 0)$ which makes the recurrent class $\{0, 1\}$ stochastically stable.

Suppose instead that the optimal non-monotone decision rule is $d = (a^+, a^+, a^-, a^-, a^+)$. Then

$$\alpha = \left(\frac{\epsilon}{20 - 4\epsilon}, \frac{8 - 4\epsilon}{20 - 4\epsilon}, \frac{12 - 6\epsilon}{20 - 4\epsilon}, \frac{4\epsilon}{20 - 4\epsilon}, \frac{\epsilon}{20 - 4\epsilon} \right),$$

so that $\alpha^* = (0, 0.4, 0.6, 0, 0)$. Thus, the recurrent class $\{1, 2\}$ is stochastically stable. ▲

In summary-statistic games, agents' stage payoffs are highest if everyone coordinates on the same effort level. Following Proposition 4, the best-reply dynamic induced by a monotone policy will converge to All-L or All-H. To the contrary, Example 8 illustrates that a non-monotone policy might prevent agents to coordinate on the same effort level in the long run. Thus, we can conclude that in the long run, agents payoffs will be highest if the principal's monitoring policy is monotone.

In the MDM setting, monitoring is costly and punishment decreases agents' payoffs. Again, Proposition 4 and Example 8 suggest that we should expect more monitoring under a non-monotone policy. Thus, we can conclude that average stage payoffs will be higher under a monotone policy.

5 Concluding remarks

Workers, industries and entire economies can be trapped in inferior equilibria and many interventions can be interpreted as attempts to guide agents other (possibly more efficient) equilibria – be it bonus schemes that should induce workers to exert more effort in failing corporations as in Brandts and Cooper (2006) or governments that synchronize production levels as in Bryant (1983).

Results. My results suggest that interventions are most successful if they are conditioned on the agents' current strategy profile. In terms of our monitoring example, this is the total number of agents who exert high effort. First, this guarantees that the principal's monitoring has an effect on agents behavior. Second, this implies that transitions to efficient equilibria are rather cheap. Third, it implies that relapses are quickly dealt with. If the number of agents who exert high effort is decreasing, the principal automatically increases monitoring and thus prevents a slippery slope. However, my results also imply that there can be situations where it is better for the principal to give up. If the cost of monitoring does not cover the expected gains from higher effort levels, it is better for the principal to leave the system to itself until one day, the time is right to give it another try. Thus, we can conclude that the optimal monitoring policy is *more responsive* but *less invasive* than for example a policy where everyone is monitored all the time or a 'bang-bang' policy, where periods where all agents are monitored alternate with periods where no-one is monitored. Furthermore, if the principal's stage payoffs are supermodular and increasing at an increasing rate (IIR), the principal's expected payoff is necessarily supermodular and thus the optimal policy is monotone increasing in the number of agents who exert high effort. Monotone policies have attractive long-term properties. They allow agents to coordinate on the same effort level almost all of the time and minimize the number of periods where the principal monitors. Furthermore, they boil the principal's optimization problem down to a simple stopping problem.

Robustness. How robust are the obtained results? Most importantly, the proposed model rests on the assumption that agents base their beliefs on observed behavior from the past, even in a non-stationary environment. However, as argued in the introduction, experimental evidence suggests that this is a plausible assumption. Within the framework of best-reply learning, the result that the optimal policy is minimally-invasive appears to be fairly robust. As shown in the proof of Proposition 1, this result depends on the observation that increasing monitoring levels weakly increases costs, but might not change agents behavior – either because monitoring is not sufficient to make high effort a best-reply, or because agents would exert high effort for even lower monitoring levels. Thus, the majority of monitoring levels is weakly dominated.

To the contrary, the optimality of monotone policies requires stronger assumptions that ensure that supermodularity is preserved in every step of the value-iteration algorithm.

First, the model assumes that in every period, only one agent updates her strategy. Inertia is a central feature of best-reply learning models (Kandori et al., 1993), so this assumption is a natural starting place. If more than one agent updates at a time, the transition matrix developed in Section 2.1 loses it's tri-diagonal form and it gets more difficult to establish supermodularity analytically. On the other hand, increasing the number of agents who update in each period implies that transitions occur faster. Therefore, states with very high and very low payoffs are reached faster, which makes monitoring more attractive. As illustrated in Example 9, this might make it even easier to achieve monotonicity.

Second, the model assumes that errors are uniformly distributed. Alternatively, it could

be plausible that agents errors depend on states, time or the agents' payoffs. However, as shown by Bergin and Lipman (1996), almost any equilibrium can be supported in best-reply dynamics if agents' errors are state dependent. Thus, depending on the specification of errors, it might get easier or more difficult to show the optimality of monotone policies.

Further research. One avenue of further research would be to extend the proposed model to multiple effort levels. Another avenue would be to explore alternative update and error modalities. However, the calibration of these parameters is essentially an empirical question. Thus, an experimental investigation of the proposed model would be another avenue of further research.

References

Alos-Ferrer, C. and Ania, A. B. (2005). The evolutionary stability of perfectly competitive behavior. *Economic Theory*, 26(3):497–516.

Bergin, J. and Lipman, B. L. (1996). Evolution with state-dependent mutations. *Econometrica*, 64(4):943–956.

Brandts, J. and Cooper, D. J. (2006). A change would do you good.... an experimental study on how to overcome coordination failure in organizations. *The American Economic Review*, pages 669–693.

Bryant, J. (1983). A simple rational expectations keynes-type model. *The Quarterly Journal of Economics*, 98(3):525–28.

Crawford, V. P. (1995). Adaptive dynamics in coordination games. *Econometrica: Journal of the Econometric Society*, 63(1):103–143.

Devetag, G. (2005). Precedent transfer in coordination games: An experiment. *Economics Letters*, 89(2):227–232.

Devetag, G. and Ortmann, A. (2007). When and why? a critical survey on coordination failure in the laboratory. *Experimental Economics*, 10(3):331–344.

Ellison, G. (2000). Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution. *The Review of Economic Studies*, 67(1):17–45.

Freidlin, M. I. and Wentzell, A. D. (1984, 2012). *Random perturbations of dynamical systems*. Springer, New York.

Grimmett, G. R. and Stirzaker, D. R. (2001). *Probability and random processes*. Oxford University Press, New York.

Jensen, M. K. (2006). Aggregative games. *Discussion Paper, University of Birmingham*, 10.

Jensen, M. K. (2010). Aggregative games and best-reply potentials. *Economic Theory*, 43(1):45–66.

Kandori, M., Mailath, G. J., and Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica*, 61(1):29–56.

Kemeny, J. G. and Snell, J. L. (1976). *Finite Markov chains*, volume 210. Springer, New York.

Kim, Y. (1996). Equilibrium selection in n-person coordination games. *Games and Economic Behavior*, 15(2):203–227.

Puterman, M. L. (2005). *Markov decision processes: discrete stochastic dynamic programming*, volume 414. Wiley & Sons, Hoboken (NJ).

Robles, J. (1997). Evolution and long run equilibria in coordination games with summary statistic payoff technologies. *Journal of Economic Theory*, 75(1):180–193.

Schelling, T. C. (1978). *Micromotives and macrobehavior*. W.W. Norton & Company Inc., London (UK), New York (NY).

Van Huyck, J. B., Battalio, R. C., and Beil, R. O. (1990). Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review*, 80(1):234–248.

Van Huyck, J. B., Battalio, R. C., and Beil, R. O. (1991). Strategic uncertainty, equilibrium selection, and coordination failure in average opinion games. *The Quarterly Journal of Economics*, 106(3):885–910.

Varian, H. (2004). System reliability and free riding. *Economics of information security*, pages 1–15.

Weber, R. A. (2006). Managing growth to achieve efficient coordination in large groups. *The American Economic Review*, 96(1):114–126.

Young, H. P. (1993). The evolution of conventions. *Econometrica*, 61(1):57–84.

Young, H. P. (2001). *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press, Princeton (NJ).

Appendix A Markov chains

A finite *Markov chain* (K, P) is a discrete-time stochastic process on a finite state space K that follows a transition rule $P : K \times K \to [0, 1]$ and has the property that the probability to be in a certain state in a given period depends *only* on the state the chain occupied in the previous period Kemeny and Snell (1976, definitions 2.1.1 and 2.1.2). The transition rule can be conveniently be represented as a transition matrix $P = (p_{kj})_{k,j=0}^{n}$, where entry p_{kj} gives the probability that the chain will be in state j in the next period, given that it is currently in state k.

A stationary (time-homogeneous) Markov chain has the same transition matrix P in every period. For this class of chains, it is particularly easy to compute the expected movement of a Markov chain. Let $\mathcal{P}(K)$ be the set of probability vectors on K, with typical member being a row vector $\mu = (\mu(0), ..., \mu(n))$. Note that $\mu(k) \in [0, 1]$ and $\sum_k \mu_k = 1$. The initial distribution of states is given by some $\mu_0 \in \mathcal{P}(K)$. The expected distribution of states in period $t + 1$ when the chain is currently in period t is $\mu_{t+1} = \mu_t P$. The expected distribution in $t + 2$ is $\mu_{t+2} = \mu_{t+1} P = (\mu_t P) P = \mu_t P^2$ and in general, the expected distribution in period $t + \tau$ is

$$\mu_{t+\tau} = \mu_{t+\tau-1} P = \mu_{t+\tau-2} P^2 = ... = \mu_t P^\tau.$$

A non-stationary (time-inhomogeneous) Markov chain has a different transition matrix P_t in every period. Here the expected distribution in period $t+\tau$ is $\mu_{t+\tau} = \mu_t P_t \times ... \times P_{\tau-1}$. In the remainder of this appendix, I will consider only stationary Markov chains.

A recurrent class is a set of states $KR \subseteq K$ that once entered cannot be left again. States that belong to a recurrent class are called recurrent, other states are called transient and are summarized in transient sets $KT \subset K$.

A Markov chain is *absorbing* if it contains both recurrent and transient states. Such a chain will end up in one of its' recurrent classes after a finite number of periods. To which recurrent class the chain converges is path dependent. The recurrent classes of an absorbing chain are sometimes called limit states $\Omega \subseteq K$. The unperturbed Markov chain $(K, P_{\gamma,\epsilon=0})$ is absorbing, as $q_k = 1$ for $k > \gamma$ and $q_k = 0$ for $k \leq \gamma$ in this case. If $k > \gamma$, the Markov chain can move to higher states or remain in the same state as in the previous period, but cannot move to lower states. Once state n has been reached (and this will occur in finite time with probability one), the chain remains in this state forever. To the contrary, if $k \leq \gamma$, the chain can only move towards lower states and once state 0 is reached, it remains there. Thus, states 0 and n are absorbing. The recurrent classes of the unperturbed Markov chain are called limit sets.

The *basin of attraction* of a recurrent class of an absorbing Markov chain is the set of states from where the chain converges to this recurrent class with probability one.

A Markov chain is *regular* if it is irreducible and acyclic (Kemeny and Snell, 1976). A chain irreducible if it contains only one recurrent class and if this recurrent class coincides with the state space, meaning that every state can be reached from any other state (possibly in more than one step). A chain is acyclic, if it does not move through states in a definite order. Following Kemeny and Snell (1976, Theorem 4.1.2), a Markov chain is regular if

and only if P^t has no zero entries for some $t > 0$. Perturbed Markov chains $(K, P_{\gamma,\epsilon>0}))$ are regular. To see this, suppose that the chain starts in state 0. Then it can be in states 0 or 1 in the next period and in states 0, 1, 2 in the subsequent period. After n periods, the chain has a positive probability to be in every state of the chain. By the same argument, every state can be reached from every other state with positive probability after at most n periods. To the contrary, as $p_{kk} > 0$, the Markov chain also has a positive probability to stay in any state it has reached for any finite number of periods. Thus, if $\epsilon > 0$, the Markov chain representing the learning dynamic is regular.

Stationary distribution. A regular Markov chain has a unique *stationary distribution* α that is approached from every initial distribution $\mu_0 \in \mathcal{P}(K)$ as time goes to infinity Kemeny and Snell (1976, Theorem 4.1.3 and 4.1.4). Intuitively, this is the case because the distance between the largest and the smallest element in each column of P^t approaches zero as t becomes large. The stationary distribution has the property that $\alpha = \alpha P$. That is, α is a left eigenvector of P, more precisely the unique left eigenvector of P with an eigenvalue of 1, as probabilities have to be between zero and one and have to sum to one. According to Lemma 4, the stationary distribution is given by

$$\alpha = e^{\top} (I - P + E)^{-1} \tag{25}$$

That is, once the transition matrix of a finite Markov chain is known, the stationary distribution can be computed by simple matrix operations.
Young (1993) shows that the regular Markov chain $(K, P_{\gamma,\epsilon})$ also has a stationary distribution when ϵ approaches zero. This so-called *limit invariant distribution* is given by $\alpha^* = \lim_{\epsilon \to 0} \alpha^{\epsilon}$. *Stochastic stability* is a qualitative feature of the limit invariant distribution. The set of stochastically stable states $K^* \in K$ is the set of all states that occur with positive probability in the limit invariant distribution,

$$K^* = \{k \in K : \alpha^*(k) > 0\}. \tag{26}$$

Building on Freidlin and Wentzell (2012), Young (1993) uses graph-theoretic techniques to derive the set of stochastically stable states. This approach is simplified by the observation that the set of stochastically stable states of the perturbed Markov chain $(K, P_{\gamma,\epsilon\to 0})$ has to be contained in the recurrent classes (the limit sets Ω) of the unperturbed Markov chain $(K, P_{\gamma,\epsilon\to 0})$. Consider the set of directed graphs that represent the movement of the learning process. The vertices in this set are all states $k, j \in K$, and a transition from state k to state j is denoted by a directed edge $[k, j]$.

A *j-tree* T_j is a directed graph with a unique sequence of directed edges (directed path) connecting any vertex $k \neq j$ to j. The set of j-trees is given by $\mathscr{T}_j$.

The *resistance* $r(k, j)$ is the minimum number of mistakes needed for a direct transition from state k to state j if k and j are neighbouring states. If state j cannot be directly accessed from k, $r(k, j) = \infty$. In our case, this means that if a transition from k to $j \in \{k-1, k, k+1\}$ occurs with positive probability in the unperturbed process, then $r(k, j) = 0$, and if such a transition does not occur with positive probability in the unperturbed process,

then $r(k,j) = 1$. A least-resistance j-tree is a tree that minimizes $\sum_{[k,j]\in T_j} r(k,j)$ over all $T_j \in \mathscr{T}_j$. The resistance of the least-resistance j-tree is known as *stochastic potential*, denoted $\rho(j)$. As shown in the following lemma, the set of stochastically stable states is the set with minimal stochastic potential $\rho(k)$.

Lemma 5. (Young (1993), Theorem 4(ii)) *Consider the regular Markov chain given by $(K, P_{\gamma,\epsilon>0})$. State $k \in K$ is stochastically stable if and only if $k \in \Omega$ and $\rho(j) \leq \rho(k)$ for all $j \in K$.*

Passage times. The *first passage time* is the time needed to move from one state to another. Kemeny and Snell (1976, chap. 4.4-4.5) show how the mean and the variance of the first passage time can be computed. The *maximum expected waiting time* (MEWT) is an upper bound for first passage times in the perturbed Markov chains $(K, P_{\gamma,\epsilon\to 0})$ proposed by Ellison (2000). The *co-radius* $CR(k)$ of a limit set of the unperturbed chain is the minimum number of mistakes needed to move into its' basin of attraction from another recurrent class. For sufficiently small $0 < \epsilon < \epsilon'$, the MEWT $\eta_\epsilon(k)$ is bounded from above by $\eta_\epsilon(k) < c\epsilon^{-CR(k)}$ for some positive constant c.

A.1 Proofs for Section 4

Proof of Lemma 4. Note that $\alpha = \alpha P$ can be rearranged to

$$\begin{aligned} \alpha(I-P) &= 0, \\ \alpha(I-P) + e^\top - e^\top &= 0 \text{ and as } \alpha E = e^\top, \\ \alpha(I-P+E) &= e^\top \text{and thus} \\ \alpha &= e^\top (I-P+E)^{-1}. \end{aligned}$$

Note that for regular Markov chains, the matrix $(I-P+E)$ has full rank. Therefore, the matrix is invertible. ■

Proof of Proposition 4. Recall that only the limit states of the unperturbed Markov chain, $k = 0, n$, are candidates for stochastically stable states and that the agent chooses H if $k > \gamma$ and L if $k \leq \gamma$, where $\gamma \in \{0, 1, ..., n-1\}$. T_0 is the least-resistance 0-tree, a directed graph from 0 to n passing exactly once through every intermediate state, $0 \to 1 \to ... \to n-1 \to n$, where the resistance $r(k, k+1)$ is one for $k \leq \gamma$ and zero for $k > \gamma$. The stochastic potential of T_0 is thus $\rho(0) = \sum_{k=0}^{n-1} r(k, k+1) = \gamma + 1$. To the contrary, T_n is the least resistance n-tree, with a resistance $r(k, k-1)$ of one for $k > \gamma$ and a resistance of zero for $k \leq \gamma$. Thus, the stochastic potential of T_n is $\rho(n) = \sum_{k=1}^{n} r(k, k-1) = n - \gamma$. Hence, states 0 and n have the same stochastic potential if $\gamma + 1 = n - \gamma$ or $\gamma = (n-1)/2$. Furthermore, $\rho(0) < \rho(n)$ if $\gamma > (n-1)/2$ and $\rho(n) < \rho(0)$ if $\gamma < (n-1)/2$. ■

Appendix B Markov decision processes

Consider a decision maker who can affect the state of a system by choosing an appropriate action and whose payoff from affecting the system depends only on current states and actions. A discrete-time *Markov decision process* is a model of sequential decision making under uncertainty with the property that the state in the following period and the decision maker's stage payoff depend only on the current state and the current action.

Formally, an (infitine-horizon discrete time) Markov decision process is given by a tuple $[K, A, p_t(j|k,a), \pi_t(k,a)]$, where $T = \{0,1,2,...\}$ is the discrete time horizon, K is the state space, A is the action space for state $k \in K$, $p_t(j|k,a)$ gives the probability (in period $t \in T$) that the system will occupy state $j \in K$ in the following period when the current state is $k \in K$ and the current action is $a \in A$. Finally, $\pi_t(k,a)$ denotes the decision maker's payoff in period $t \in T$. Together with an optimality criterion, a Markov decision process constitutes a Markov decision problem. The Markov decision problem outlined in section 3.3 belongs to the class of *discounted MD problems*, where the decision maker maximizes her expected total discounted payoff for some discount factor $\lambda \in [0,1)$.

A *decision rule* $d = (d(0), ...d(n))$ is a vector that assigns an action for every state of the Markov decision problem. A decision rule is called Markov if it depends only on the current state of the system and is called history-dependent if it depends on the (entire) history of the Markov decision problem. A decision rule is called deterministic if for every state it assigns an action with probability one. Otherwise, it is called randomized. Let D^{MD} and D^{MR} be the sets of deterministic and randomized Markov decision rules respectively and let D^{HD} and D^{HR} be the sets of deterministic and randomized history-dependent decision rules. Note that $D^{MD} \subset D^{MR} \subset D^{HR}$ and $D^{MD} \subset D^{HD}$.

A *policy* $\delta = (d_0, d_1, d_2, ...)$ assigns a decision rule for every period $t \in T = \{0,1,...,\infty\}$. A Markov policy assigns only Markov decision rules, a history-dependent policy assigns history-dependent decision rules. A stationary policy assigns the same Markov decision rule in every period. Let Δ^{SD} and Δ^{SR} be the sets of policies that use only deterministic and randomized stationary decision rules respectively. Similarly, let Δ^{MD} and Δ^{MR} be the sets of deterministic and randomized Markov policies and Δ^{HD} and Δ^{HR} the sets of deterministic and randomized history-dependent policies. Note that $\Delta^{SD} \subset \Delta^{SR} \subset \Delta^{MR} \subset \Delta^{HR}$ and $\Delta^{SD} \subset \Delta^{MD} \subset \Delta^{HD} \subset \Delta^{HR}$.

A large part of the Markov decision process literature concerns the question whether it is appropriate to restrict attention to a more specific policy space. Puterman (2005, Theorem 5.5.1) states that for every randomized history dependent policy, it is possible to construct a randomized Markov policy that induces the same transition probabilities $p_t(k|k,a)$ in every period $t \in T$. Puterman (2005, Proposition 6.2.1) shows that for discrete state and action spaces it is without loss of generality to consider deterministic policies. This result follows from the observation that the supremum of a function on a discrete domain is larger than or equal to all probability distributions on this function. Finally, in many Markov decision problems it is possible to restrict attention to stationary policies. After a brief introduction to normed vector spaces in Appendix B.1, these results are reviewed in Appendix B.2.

B.1 Normed vector spaces

Let V be the set of bounded real-valued functions on K. Then v is an element of V if $v : K \to \mathbb{R}$ and if there exists a constant C such that $|v(k)| \leq C$ for all $k \in K$. The maximum norm of $v \in V$ is given by

$$||v|| = \max_{k \in K} |v(k)| \quad \forall v \in V. \tag{27}$$

V is closed under addition and scalar multiplication and endowed with a norm, it is a normed vector space.

A *Cauchy sequence* $\langle v^m \rangle \in V$ has the property that for every $\eta > 0$ there exists an M such that whenever $m, o > M$, it holds that $||v_m - v_o|| < \eta$. It can be shown that every Cauchy sequence in V contains a limit in V. This makes V a complete normed linear space (Banach space). It is furthermore assumed that V is partially ordered, meaning that if $u, v \in V$ and $u(k) \leq v(k)$ for all $k \in K$, then $u \leq v$. Consider the matrix norm

$$||Q|| = \max_{k \in K} \sum_{k \in K} |q(k, j)|. \tag{28}$$

A *linear transformation* Q on V is bounded if there exists a constant $K > 0$ such that $||Qv|| \leq K||v||$ for all $v \in V$. Probability matrices Q are bounded linear transformations as $||Q|| = 1$ in that case. Puterman (2005, Lemma C.1) notes that if V is a Banach space, the set of bounded linear transformations on V, denoted $S(V)$, is bounded as well. $Q \in S(V)$ is a positive linear transformation if $Qv \geq 0$ whenever $v \geq 0$.

An *operator* $T : V \to V$ is a *contraction mapping* on a Banach space V if for all $u, v \in V$ and some $\lambda \in (0, 1)$, it holds that $||Tv - Tu|| \leq \lambda||v - u||$. The following version of the Banach Fixed-Point Theorem is due to Puterman (2005, Theorem 6.2.3).

Lemma 6. (Banach Fixed-Point Theorem) *Suppose V is a Banach space and $T : V \to V$ is a contraction mapping. Then*

a. there exists a unique v^* in V such that $Tv^* = v^*$; and
b. for arbitrary v^0 in V, the sequence $\langle v^m \rangle$ defined by

$$v^{m+1} = Tv^m = T^{m+1}v^0 \tag{29}$$

converges to v^*.

B.2 Optimality of stationary policies

Recall that D is the set of deterministic Markov decision rules. The principal's stage payoff from state $k \in K$ when using decision rule $d \in D$ is $\pi_d(k) = \pi(k, d(k))$. A payoff vector is a column vector $\pi_d = (\pi_d(0), ..., \pi_d(n))^\top$ that summarizes stage payoffs for decision rule $d \in D$. The probability that the process is in state $j \in K$ in the next period when the decision rule is $d \in D$ and the current state is $k \in K$ is $p_d(j|k) = p(j|k, d(k))$. Let P_d

be the transition matrix with component $p(k,j) = p_d(j|k, d(k))$. Recall that there is a distinct transition matrix for every decision rule $d \in D$. Thus, a policy $\delta \in \Delta^{MD}$ induces a non-stationary (or time-inhomogeneous) Markov chain $(K, \{P_{d_0}, P_{d_1}, ...\})$. The principal's expected payoff from policy $\delta = (d_0, d_1, ...)$ can be summarized by the column vector

$$\begin{aligned} v_\lambda^\delta &= \pi_{d_0} + \lambda P_{d_0}\pi_{d_1} + \lambda^2 P_{d_0}P_{d_1}\pi_{d_2} + ... \\ &= \pi_{d_0} + \lambda P_{d_0}\left(\pi_{d_1} + \lambda P_{d_1}\pi_{d_2} + \lambda^2 P_{d_1}P_{d_2}\pi_{d_3} + ...\right) \\ &= \pi_{d_0} + \lambda P_{d_0} v_\lambda^{\delta'} \end{aligned} \tag{30}$$

for $\delta' = (d_1, d_2, ...)$. That is, the principal's optimization problem is equivalent to a one-period problem with $v_\lambda^{\delta'}$ as terminal reward. A deterministic stationary policy $\delta = (d, d, ...)$ induces a stationary Markov chain (K, P_d) and yields expected payoff $v_\lambda^{d^\infty} = r_d + \lambda P_d v_\lambda^{d^\infty}$ so that $v_\lambda^{d^\infty} = (I - \lambda P)^{-1} r_d$.

Recall that V is the set of real-valued bounded functions on K and note that $v_\lambda^\delta \in V$. The *(optimal) value* of a Markov decision problem with discount factor $\lambda \in [0, 1)$ is given by

$$v_\lambda^* = \max_{\delta \in \Delta^{MD}} v_\lambda^\delta. \tag{31}$$

Note that V is a partially ordered set that is bounded, so there is a greatest element in V. A policy $\delta^* \in \Delta^{MD}$ is *discount optimal* for a given $\lambda \in [0, 1)$ if

$$v_\lambda^{\delta^*}(k) = v_\lambda^*(k) \quad \forall k \in K. \tag{32}$$

Recall the L-operator $Lv = \max_{d \in D}\{\pi_d + \lambda P_d v\}$.

Lemma 7. (Puterman (2005), Theorem 6.2.6) *A policy $\delta^* \in \Delta^{HR}$ is optimal if and only if $v_\lambda^{\delta^*}$ is a solution to the optimality equation $Lv = v$.*

Intuitively, Lemma 7 holds because once the maximal expected payoff (or optimal value) v_λ^* is reached, it is not possible to improve on that. Puterman (2005, Proposition 6.2.4) establishes that for $\lambda \in [0, 1)$, the L-operator is a contraction mapping on the set of bounded real-valued functions V. Thus, the Banach Fixed-Point Theorem tells us that there is a unique value u^* that satisfies $Lv = v$ and that this value is the limit of a Cauchy sequence obtained by repeatedly applying the L-operator. Hence we know that the optimal value $v_\lambda^* = u^*$ can be found by repeatedly applying the L-operator. This is the idea behind the value-iteration algorithm discussed in Appendix B.3.

According to Puterman (2005), Lemma 5.6.1, $L_d v = r_d + \lambda P_d v$ is a positive bounded linear transformation on V and therefore $L_d v \in V$. A decision rule $d^* \in D$ is called *conserving*, if it satisfies $L_d^* v_\lambda^* = v_\lambda^*$.

Lemma 8. (Puterman, 2005, Theorem 6.2.6 b.) *Suppose there exists a conserving decision rule $d^* \in D$. Then the deterministic stationary policy $(d^*)^\infty$ is optimal.*

Proof. According to Puterman (2005), Theorem 6.2.2(c), the optimality equation $Lv = v$ has a unique solution that is equal to the optimal value of the Markov decision process, v_λ^*. It follows that

$$L_{d^*} v_\lambda^* = v_\lambda^* = L v_\lambda^*. \tag{33}$$

Following Puterman (2005), Theorem 6.1.1, $v_\lambda^{d^\infty}$ is the unique solution to the equation $L_d v = v$. In particular, this implies that $v_\lambda^{(d^*)^\infty}$ is the unique solution to $L_d^* v_\lambda^* = v_\lambda^*$ and therefore we can conclude that $v_\lambda^* = v_\lambda^{(d^*)^\infty}$. The optimality of policy $(d^*)^\infty$ follows. ■

Lemma 9. *Suppose that state space K and action space A are finite. Then there exists a conserving decision rule $d^* \in D$.*

Proof. Following Puterman (2005, Theorem 6.2.7 a.), there exists a conserving decsion rule if K is discrete (so in particular if it is finite) and if $Lv = max_{d \in D}\{r_d + \lambda P_d v\}$ can be attained for all $v \in V$. Following Puterman (2005, 6.2.10), Lv can be attained if A is finite. ■

Corollary 1. (Proof of Lemma 1) *For the MDM problem there exists a conserving decision rule $d^* \in D$ and the stationary policy $\delta^* = (d^*)^\infty$ is optimal.*

B.3 The value-iteration algorithm

The value-iteration algorithm is a procedure for finding optimal stationary policies. The following characterization is due to Puterman (2005, p.161), adapted to the notation in the current paper.

1. Select $v^0 \in V$, specify tolerance $\eta > 0$ and set $m = 0$.

2. For each $k \in K$, compute $v^{m+1}(k)$ by

$$v^{m+1}(k) = \max_{a \in A} \left\{ \pi(k,a) + \lambda \sum_{j \in K} p(j|k,a) v^m(j) \right\}. \tag{34}$$

3. If

$$||v^{m+1} - v^m|| < \eta(1-\lambda)/2\lambda, \tag{35}$$

   go to step 4. Otherwise, increment m by 1 and return to step 2.

4. For each $k \in K$, choose

$$d_\eta(k) \in \arg\max_{a \in A} \left\{ \pi(k,a) + \lambda \sum_{j \in K} p(j|k,a) v^m(j) \right\} \tag{36}$$

   and stop.

Following Puterman (2005) convergence of the value iteration algorithm is linear in λ. The speed of convergence can be increased further if we restrict the set of decision rules. The set of deterministic stationary decision rules D has a cardinality of $|A|^{|K|} = (n+1)^{(n+1)}$. Decision space D' has a cardinality of $|K|^2 = (n+1)^2$ and the set of minimally-invasive decision rules $D^S = A_0 \times ... \times A_n$ has a cardinality of $(\gamma+1)^{26}$.

B.4 Lemmas for Propositions 1 and 2

Recall the definition of transition probabilities from (2). For future reference, note that transition probabilities for states $k' \in \{k-1, k, k+1\}$ can be rearranged to

$$
\begin{aligned}
p_{k-1,k-2} &= \tfrac{k-1}{n}(1-q_{k-1}),\\
p_{k-1,k-1} &= \tfrac{k-1}{n}q_{k-1} + \left(1-\tfrac{k-1}{n}\right)(1-q_{k-1}) = \tfrac{k-1}{n}q_{k-1} + \left(1-\tfrac{k+1}{n}\right)(1-q_{k-1}) + \tfrac{2}{n}q_{k-1},\\
p_{k-1,k} &= \left(1-\tfrac{k-1}{n}\right)q_{k-1} = \left(1-\tfrac{k+1}{n}\right)q_{k-1} + \tfrac{2}{n}q_{k-1},
\end{aligned}
$$

$$
\begin{aligned}
p_{k,k-1} &= \tfrac{k}{n}(1-q_k) = \tfrac{k-1}{n}(1-q_k) + \tfrac{1}{n}(1-q_k),\\
p_{k,k-1} &= \tfrac{k}{n}q_k + \left(1-\tfrac{k}{n}\right)(1-q_k) = \tfrac{k-1}{n}q_k + \left(1-\tfrac{k+1}{n}\right)(1-q_k) + \tfrac{1}{n},\\
p_{k,k-1} &= \left(1-\tfrac{k}{n}\right)q_k = \left(1-\tfrac{k+1}{n}\right)q_k + \tfrac{1}{n}q_k,
\end{aligned}
\tag{37}
$$

$$
\begin{aligned}
p_{k+1,k} &= \tfrac{k+1}{n}(1-q_{k+1}) = \tfrac{k-1}{n}(1-q_{k+1}) + \tfrac{2}{n}(1-q_{k+1}),\\
p_{k+1,k+1} &= \tfrac{k+1}{n}q_{k+1} + \left(1-\tfrac{k+1}{n}\right)(1-q_{k+1}) = \tfrac{k-1}{n}q_{k+1} + \left(1-\tfrac{k+1}{n}\right)(1-q_{k+1}) + \tfrac{2}{n}q_{k+1},\\
p_{k+1,k+2} &= \left(1-\tfrac{k+1}{n}\right)q_{k+1},
\end{aligned}
$$

Let $P^a_{k'} = (p_{k',k-2}, ..., p_{k',k+2})$ with $k' \in \{k-1, k, k+1\}$ and $k \in N$ be a row vector and $u = (u_{k-1}, u_k, u_{k+1}, u_{k+2})^\top$ be a column vector with $u_k = v(k)$ for all $k \in K$. If $0 < k < n$, it is clear that u is a subset of v. Thus, it can be easily seen that $P^a_k u = \sum_{j=0}^n p(j|k,a)v(j)$. For $k \in \{0, n\}$, there are elements in u that have no counterpart in v. However, there is no positive probability on these elements either. Thus $P^a_k u = \sum_{j=0}^n p(j|k,a)v(j)$ for all $k \in K$.

Structured Policies. Recall that the set of structured decision rules is $D^S = \times_{k=0}^n A_n$. The set of structured values $V^S \subset V$ is the space of monotone increasing bounded real-valued functions $v : K \to \mathbb{R}$,

$$V^S = \{v \in V : \forall k \in K \backslash \{n\},\ v(k) \leq v(k+1)\}.$$

Lemma 10. *If $v(j)$ is weakly increasing in $j \in K$, $\sum_{j=0}^n p(j|k,a)v(j)$ is weakly increasing in k for all $a \in A$.*

[6] To see this, recall that there are two actions that are candidates for optimality for all states $k = 0, 1, ..., \gamma$, while there is only one such candidate for states $k = \gamma+1, ..., n$

Proof. We need to show that $(P^a_{k'} - P^a_k)u \geq 0$ for all $k, k' \in K$ with $k' \geq k$ and $a \in A$. Note that if this condition holds locally for all k and $k' = k+1$, it must hold for any $k' \geq k$. Recall that q_k is the probability that an agent chooses H given (k, a). For a given $a \in A$, q_k is non-decreasing in k. Thus, the worst case is to assume that $q_k = q_{k+1} = q$. It follows that $(P^a_{k'} - P_k)u \geq 0$ can be rearranged to

$$\begin{aligned}(P^a_{k+1} - P^a_k)u = {} & \tfrac{k-1}{n}(1-q)(u_k - u_{k-1}) + \left(\tfrac{k-1}{n}q + \left(1 - \tfrac{k+1}{n}\right)(1-q)\right)(u_{k+1} - v_k) + \\ & + \left(1 - \tfrac{k+1}{n}\right) q(u_{k+2} - u_{k+1}) + \tfrac{1}{n}(1-q)(u_k - u_{k-1}) + \\ & + \tfrac{1}{n}q(u_{k+1} - u_k) + \tfrac{1}{n}q(u_k - u_k) = \\ = {} & \tfrac{k}{n}(1-q)(u_k - u_{k-1}) + \left(\tfrac{k}{n}q + \left(1 - \tfrac{k+1}{n}\right)(1-q)\right)(u_{k+1} - v_k) + \\ & + \left(1 - \tfrac{k+1}{n}\right) q(u_{k+2} - u_{k+1}). \end{aligned} \tag{38}$$

For $k < n$ and u monotone increasing, all terms of this expression are non-negative, so we can conclude that $(P^a_{k+1} - P^a_k)u \geq 0$, as required. ■

Lemma 11. *V^S is a closed subset of V.*

Proof. To show that V^S is a closed subsets of V, we need to show that every Cauchy sequence starting in V^S has a limit in V^S. Suppose $\langle v_n \rangle$ is an arbitrary Cauchy sequence in V^S with $v_n \to v$ and suppose to the contrary that $v \notin V^S$. That is, suppose there are elements $x, y \in K$ with $x < y$ for which $v(x) > v(y)$. Let $\psi = v(x) - v(y) > 0$. As $\langle v_n \rangle$ is a Cauchy sequence, we know that there exists an integer $N \in X$ such that $\|v_n - v\| < \psi/2$. In component notation, this implies that for all $z \in K$, $|v_n(z) - v(z)| < \psi/2$. In particular

$$\begin{aligned} &|v_n(x) - v(x)| < \psi/2 \text{ and thus } -v_n(x) < -v(x) + \psi/2 \\ &|v_n(y) - v(y)| < \psi/2 \text{ and thus } v_n(y) < v(y) + \psi/2. \end{aligned}$$

Adding up both inequalities yields

$$\begin{aligned} &v_n(y) - v_n(x) < v(y) - v(x) + 2\psi/2 = -\psi + 2\psi/2 \text{ and thus} \\ &v_n(y) - v_n(x) < 0, \end{aligned}$$

which is a contradiction to the assumption that $\langle v_n \rangle$ is in V^S. Therefore the limit v has to be an element of V^S and we are done. ■

Monotone Policies. Recall that the set of of monotone decision rules is given by

$$D^M = \left\{ d \in D^S : \forall k \in K \backslash \{n\},\, d(k) \leq d(k+1) \right\},$$

while the set of monotone values is the space of bounded real-valued functions $v : K \to \mathbb{R}$ that are increasing in $k \in K$ at an increasing rate,

$$V^M = \left\{ v \in V : \forall k \in K \backslash \{0, n\},\, v(k+1) + v(k-1) \geq 2v(k) \right\}.$$

Lemma 12. *If $v(j)$ is IIR in $j \in K$, $\sum_{j=0}^{n} p(j|k,a)v(j)$ is IIR in k for all $a \in A'$.*

Proof. We need to show that $(P^a_{k+1} + P^a_{k-1} - 2P^a_k)u \geq 0$ for all $k \in K\backslash\{0,n\}$ and $a \in A'$. By design, $q_{k-1} = q_k = q_{k+1} = q$ for all $a \in A'$. Recall the presentation of transition probabilities in (37) and note that $(P^a_{k+1} + P^a_{k-1} - 2P^a_k)u$ can be arranged to

$$\begin{aligned}(P^a_{k+1} + P^a_{k-1} - 2P^a_k)u = {} & \tfrac{k-1}{n}(1-q)(u_k + u_{k-2} - 2u_{k-1}) + \\ & + \left(\tfrac{k-1}{n}q + \left(1 - \tfrac{k+1}{n}(1-q)\right)\right)(u_{k+1} + u_{k-1} - 2u_k) + \\ & + \left(1 - \tfrac{k+1}{n}\right)(u_{k+2} + u_k - 2u_{k+1}).\end{aligned} \tag{39}$$

For $0 < k < n$ and u IIR in k, this expression has only non-negative entries and we can conclude that $(P^a_{k+1} + P^a_{k-1} - 2P^a_k)u \geq 0$. ■

Lemma 13. *If $v(j)$ is IIR in $j \in K$, $\sum_{j=0}^{n} p(j|k,a)v(j)$ is supermodular on $K \times A'$.*

Proof. $\sum_{j=0}^{n} p(j|k,a)v(j)$ is supermodular on $K \times A'$ if $(P^{a^+}_{k+1} - P^{a^-}_{k+1})u \geq (P^{a^+}_{k-1} - P^{a^-}_{k-1})u$ for all $k \in K\{n\}$. Note that $q_k(a^-) = \epsilon/2$ and $q_k(a^+) = 1 - \epsilon/2$. Note that

$$\begin{aligned}(P^{a^+}_k - P^{a^-}_k)u &= (1-\epsilon) * (0, -\tfrac{k}{n}, \tfrac{k}{n} - \left(1 - \tfrac{k}{n}\right), 1 - \tfrac{k}{n}, 0)u = \\ &= (1-\epsilon)\left(\tfrac{k+1}{n}(u_{k+1} - u_k) + \left(1 - \tfrac{k+1}{n}\right)(u_{k+2} - u_{k+1})\right) \text{ and} \qquad (40) \\ (P^{a^+}_{k+1} - P^{a^-}_{k+1})u &= (1-\epsilon) * (0, 0, -\tfrac{k+1}{n}, \tfrac{k+1}{n} - \left(1 - \tfrac{k+1}{n}\right), 1 - \tfrac{k+1}{n})u \\ &= (1-\epsilon)\left(\tfrac{k}{n}(u_k - u_{k-1}) + \left(1 - \tfrac{k}{n}\right)(u_{k+1} - u_k)\right), \qquad (41)\end{aligned}$$

so that $(P^{a^+}_{k+1} - P^{a^-}_{k+1})u \geq (P^{a^+}_k - P^{a^-}_k)u$ can be simplified to the non-negativity condition

$$\tfrac{k}{n}(u_{k+1} + u_{k-1} - 2u_k) + \left(1 - \tfrac{k+1}{n}\right)(u_{k+2} + u_k - 2u_{k+1}) \geq 0. \tag{42}$$

As u is IIR on k, this condition is satisfied for all $k \in K\backslash\{n\}$ ■

Lemma 14. *V^M is a closed subset of V.*

Proof. The proof is similar to the proof for Lemma 11, that is, we have to show that a Cauchy sequence in V^M has a limit in V^M. Suppose $\langle v_n \rangle$ is an arbitrary Cauchy sequence in V^M with $v_n \to v$ and suppose to the contrary that $v \notin V^M$. That is, suppose there are elements $x-1, x, x+1 \in K$ for which $v(x+1) + v(x-1) < 2v(x)$. Let $\psi = 2v(x) - v(x+1) - v(x-1) > 0$. As $\langle v_n \rangle$ is a Cauchy sequence, we know that there exists an integer $N \in X$ such that $\|v_n - v\| < \psi/4$. In component notation, this implies that for all $z \in K$, $|v_n(z) - v(z)| < \psi/4$, so that

$$\begin{aligned}|v_n(x) - v(x)| < \psi/4 \text{ and thus } -v_n(x) &< -v(x) + \psi/4, \\ |v_n(x+1) - v(x+1)| < \psi/4 \text{ and } v_n(x+1) &< v(x+1) + \psi/4, \\ |v_n(x-1) - v(x-1)| < \psi/4 \text{ and } v_n(x-1) &< v(x-1) + \psi/4.\end{aligned}$$

Adding these three inequalities yields

$v_n(x+1) + v_n(x-1) - 2v_n(x) < v(x+1) + v(x-1) - 2v(x) + 2\psi/4 + 2\psi/4 = -\psi + 4\psi/4$ and thus
$v_n(x+1) + v_n(x-1) - 2v_n(x) < 0$,

which is a contradiction to the assumption that $\langle v_n \rangle$ is in V^M. Therefore the limit $v \in V^M$ as required. ■

B.5 Simultaneous updating

If more than one agents updates in every period, the probability that a certain number of updating agents has played high effort in the previous period follows a hypergeometric distribution. That is, we can understand the process of selecting agents who update as urn experiment with multiple draws without replacement, where ν is the number of updating agents (no. of draws), κ is the number the number of updating agents who played H in the previous period (no. of successes), n is the population size and k_{t-1} the total number of updating agents in the previous period (total no. of successes). The probability mass function of this distribution is given by

$$f_{\kappa,k_{t-1}} = \frac{\binom{k_{t-1}}{\kappa}\binom{n-k_{t-1}}{\nu-\kappa}}{\binom{n}{\nu}}. \tag{43}$$

Recall from 3 that the probability that an agent plays H in the current period is given by q_{k_t}. The probability that k_t agents play H in period t is thus equal to the probability that k_{t-1} agents played high effort in the previous period times the probability that $x = k_t - k_{t-1}$ agents change their behavior in the current period. Formally, this is given by

$$\phi_{k_t,k_{t-1}} = f_{\kappa,k_{t-1}} \times q_{k_t}^x (1-q_{k_t})^{\nu-x}. \tag{44}$$

The probability that $k_t = k$ is then given by

$$\sum_{k_{t-1}} \phi_{k_{t-1},k}. \tag{45}$$

Thus, for example, if $\nu = 2$ agents update in every period, the probability that the total effort level does not change, that is, the probability that $k_t = k_{t-1} = k$ is given

$$\sum_{k_{t-1}} \phi_{k_{t-1},k} = (1-q_k)^2 f_{\kappa=0,k_{t-1}} + 2q_k(1-q_k) f_{\kappa=1,k_{t-1}} + q_k^2 f_{\kappa=2,k_{t-1}}.$$

That is, the probability that $k_{t-1} = k_t$ is given by the probability that both updating agents exerted L and L in the previous and the current period, plus two times the probability that they exerted L and H in the previous period and the current period plus the probability that they exerted H and H in the previous period and do the same in the current period.

The following examples suggest that monotonicity of policies might be preserved if the number of updating agents is $\nu > 1$.

Example 9. Consider a reduced MDM problem with $n = 4$, $\gamma = 3$, $\epsilon = 1/100$ and $\lambda = 1/2$ and let stage payoffs be given by $\pi(k, a) = k^2/4 - (4 - k)a$ for $a \in A'$ as in Example 5. Then the optimal decision rules for ν updating agents are given by $d^*_{\nu=1} = d^*_{\nu=2} = \{a^-, a^-, a^-, a^+, a^+\}$, $d^*_{\nu=3} = \{a^-, a^-, a^+, a^+, a^+\}$ and $d_{\nu=4} = \{a^-, a^+, a^+, a^+, a^+\}$ respectively. ▲

Example 10. Consider again the reduced MDM problem with $n = 4$, $\gamma = 3$, $\epsilon = 1/100$ and $\lambda = 1/2$, but let stage payoffs be given by $\pi(k, a) = k^{1/2} - a$ for $k \leq 3$ and $a \in A'$ and $\pi(k, a) = k^{1/2}$ for $k = 4$. As we have seen in Example 6, if only one agent is updating, the optimal decision rule is $d^*_{\nu=1} = \{a^+, a^-, a^-, a^-, a^+\}$, which is non-monotone. To the contrary, if more than one agent is updating, the optimal decision rule is $d^*_{\nu=2} = d^*_{\nu=3} = d^*_{\nu=4} = \{a^+, a^+, a^+, a^+, a^+\}$, which is monotone. ▲

Thus, Examples 9 and 10 suggest that it might even get easier to achieve monotonicity if more than one agent updates her strategy in every period.

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

http://eeecon.uibk.ac.at/wopec/

2013-28 **Dominik Erharter:** Promoting coordination in summary-statistic games

2013-27 **Dominik Erharter:** Screening experts' distributional preferences

2013-26 **Loukas Balafoutas, Rudolf Kerschbamer, Matthias Sutter:** Second-degree moral hazard in a real-world credence goods market

2013-25 **Rudolf Kerschbamer:** The geometry of distributional preferences and a non-parametric identification approach

2013-24 **Nadja Klein, Michel Denuit, Stefan Lang, Thomas Kneib:** Nonlife ratemaking and risk management with bayesian additive models for location, scale and shape

2013-23 **Nadja Klein, Thomas Kneib, Stefan Lang:** Bayesian structured additive distributional regression

2013-22 **David Plavcan, Georg J. Mayr, Achim Zeileis:** Automatic and probabilistic foehn diagnosis with a statistical mixture model

2013-21 **Jakob W. Messner, Georg J. Mayr, Achim Zeileis, Daniel S. Wilks:** Extending extended logistic regression to effectively utilize the ensemble spread

2013-20 **Michael Greinecker, Konrad Podczeck:** Liapounoff's vector measure theorem in Banach spaces *forthcoming in Economic Theory Bulletin*

2013-19 **Florian Lindner:** Decision time and steps of reasoning in a competitive market entry game

2013-18 **Michael Greinecker, Konrad Podczeck:** Purification and independence

2013-17 **Loukas Balafoutas, Rudolf Kerschbamer, Martin Kocher, Matthias Sutter:** Revealed distributional preferences: Individuals vs. teams

2013-16 **Simone Gobien, Björn Vollan:** Playing with the social network: Social cohesion in resettled and non-resettled communities in Cambodia

2013-15 **Björn Vollan, Sebastian Prediger, Markus Frölich:** Co-managing common pool resources: Do formal rules have to be adapted to traditional ecological norms?

2013-14 **Björn Vollan, Yexin Zhou, Andreas Landmann, Biliang Hu, Carsten Herrmann-Pillath:** Cooperation under democracy and authoritarian norms

2013-13 **Florian Lindner, Matthias Sutter:** Level-k reasoning and time pressure in the 11-20 money request game *forthcoming in Economics Letters*

2013-12 **Nadja Klein, Thomas Kneib, Stefan Lang:** Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data

2013-11 **Thomas Stöckl:** Price efficiency and trading behavior in limit order markets with competing insiders *forthcoming in Experimental Economics*

2013-10 **Sebastian Prediger, Björn Vollan, Benedikt Herrmann:** Resource scarcity, spite and cooperation

2013-09 **Andreas Exenberger, Simon Hartmann:** How does institutional change coincide with changes in the quality of life? An exemplary case study

2013-08 **E. Glenn Dutcher, Loukas Balafoutas, Florian Lindner, Dmitry Ryvkin, Matthias Sutter:** Strive to be first or avoid being last: An experiment on relative performance incentives.

2013-07 **Daniela Glätzle-Rützler, Matthias Sutter, Achim Zeileis:** No myopic loss aversion in adolescents? An experimental note

2013-06 **Conrad Kobel, Engelbert Theurl:** Hospital specialisation within a DRG-Framework: The Austrian case

2013-05 **Martin Halla, Mario Lackner, Johann Scharler:** Does the welfare state destroy the family? Evidence from OECD member countries

2013-04 **Thomas Stöckl, Jürgen Huber, Michael Kirchler, Florian Lindner:** Hot hand belief and gambler's fallacy in teams: Evidence from investment experiments

2013-03 **Wolfgang Luhan, Johann Scharler:** Monetary policy, inflation illusion and the Taylor principle: An experimental study

2013-02 **Esther Blanco, Maria Claudia Lopez, James M. Walker:** Tensions between the resource damage and the private benefits of appropriation in the commons

2013-01 **Jakob W. Messner, Achim Zeileis, Jochen Broecker, Georg J. Mayr:** Improved probabilistic wind power forecasts with an inverse power curve transformation and censored regression

2012-27 **Achim Zeileis, Nikolaus Umlauf, Friedrich Leisch:** Flexible generation of e-learning exams in R: Moodle quizzes, OLAT assessments, and beyond

2012-26 **Francisco Campos-Ortiz, Louis Putterman, T.K. Ahn, Loukas Balafoutas, Mongoljin Batsaikhan, Matthias Sutter:** Security of property as a public good: Institutions, socio-political environment and experimental behavior in five countries

2012-25 **Esther Blanco, Maria Claudia Lopez, James M. Walker:** Appropriation in the commons: variations in the opportunity costs of conservation

2012-24 **Edgar C. Merkle, Jinyan Fan, Achim Zeileis:** Testing for measurement invariance with respect to an ordinal variable *forthcoming in Psychometrika*

2012-23 **Lukas Schrott, Martin Gächter, Engelbert Theurl:** Regional development in advanced countries: A within-country application of the Human Development Index for Austria

2012-22 **Glenn Dutcher, Krista Jabs Saral:** Does team telecommuting affect productivity? An experiment

2012-21 **Thomas Windberger, Jesus Crespo Cuaresma, Janette Walde:** Dirty floating and monetary independence in Central and Eastern Europe - The role of structural breaks

2012-20 **Martin Wagner, Achim Zeileis:** Heterogeneity of regional growth in the European Union

2012-19 **Natalia Montinari, Antonio Nicolo, Regine Oexl:** Mediocrity and induced reciprocity

2012-18 **Esther Blanco, Javier Lozano:** Evolutionary success and failure of wildlife conservancy programs

2012-17 **Ronald Peeters, Marc Vorsatz, Markus Walzl:** Beliefs and truth-telling: A laboratory experiment

2012-16 **Alexander Sebald, Markus Walzl:** Optimal contracts based on subjective evaluations and reciprocity

2012-15 **Alexander Sebald, Markus Walzl:** Subjective performance evaluations and reciprocity in principal-agent relations

2012-14 **Elisabeth Christen:** Time zones matter: The impact of distance and time zones on services trade

2012-13 **Elisabeth Christen, Joseph Francois, Bernard Hoekman:** CGE modeling of market access in services

2012-12 **Loukas Balafoutas, Nikos Nikiforakis:** Norm enforcement in the city: A natural field experiment *forthcoming in European Economic Review*

2012-11 **Dominik Erharter:** Credence goods markets, distributional preferences and the role of institutions

2012-10 **Nikolaus Umlauf, Daniel Adler, Thomas Kneib, Stefan Lang, Achim Zeileis:** Structured additive regression models: An R interface to BayesX

2012-09 **Achim Zeileis, Christoph Leitner, Kurt Hornik:** History repeating: Spain beats Germany in the EURO 2012 Final

2012-08 **Loukas Balafoutas, Glenn Dutcher, Florian Lindner, Dmitry Ryvkin:** The optimal allocation of prizes in tournaments of heterogeneous agents

2012-07 **Stefan Lang, Nikolaus Umlauf, Peter Wechselberger, Kenneth Harttgen, Thomas Kneib:** Multilevel structured additive regression

2012-06 **Elisabeth Waldmann, Thomas Kneib, Yu Ryan Yu, Stefan Lang:** Bayesian semiparametric additive quantile regression

2012-05 **Eric Mayer, Sebastian Rueth, Johann Scharler:** Government debt, inflation dynamics and the transmission of fiscal policy shocks *forthcoming in Economic Modelling*

2012-04 **Markus Leibrecht, Johann Scharler:** Government size and business cycle volatility; How important are credit constraints? *forthcoming in Economica*

2012-03 **Uwe Dulleck, David Johnston, Rudolf Kerschbamer, Matthias Sutter:** The good, the bad and the naive: Do fair prices signal good types or do they induce good behaviour?

2012-02 **Martin G. Kocher, Wolfgang J. Luhan, Matthias Sutter:** Testing a forgotten aspect of Akerlof's gift exchange hypothesis: Relational contracts with individual and uniform wages

2012-01 **Loukas Balafoutas, Florian Lindner, Matthias Sutter:** Sabotage in tournaments: Evidence from a natural experiment *published in Kyklos*

University of Innsbruck

Working Papers in Economics and Statistics

2013-28

Dominik Erharter

Promoting coordination in summary-statistic games

Abstract
This paper studies how external incentives can help agents to coordinate in summary-statistic games. Agents follow a myopic best-reply rule and face a trade-off between efficiency and strategic uncertainty. A principal can help agents to coordinate on the Pareto optimal equilibrium by monitoring an appropriate number of agents. The optimal monitoring policy is 'minimally-invasive' - for every strategy profile of the agents, the principal either monitors just enough agents to make high effort a best-reply or does not monitor at all. Furthermore, given the principal's payoffs are supermodular and increasing at an increasing rate, the optimal monitoring policy is monotone in the number of agents who choose high effort.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)