Learning to Promote Cooperation in the Collective Risk Dilemma

Learning to Promote Cooperation in the Collective Risk Dilemma

A THESIS PRESENTED
BY
ESTHER PLOTNICK
TO
THE DEPARTMENT OF MATHEMATICS
AND
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS (HONORS)
IN THE SUBJECT OF
MATH AND COMPUTER SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
NOVEMBER 2021

Thesis advisor: David Parkes, Clifford Taubes Esther Plotnick

Learning to Promote Cooperation in the Collective Risk Dilemma

Abstract

The first part of this thesis provides a survey of the game theory relevant to the analysis of multi-player player games, including the existence of mixed-strategy Nash equilibria, and the connections between Nash equilibria and evolutionary game theory through replicator dynamics. This exposition also reviews the Q-learning algorithm introduced in Watkins (1989), including a proof of its convergence.

The second part of this thesis is a study of social dilemmas, which are games where the payoff to an individual player for cooperative behavior is lower than for defecting behavior, but players are worse off if all defect. Sustaining cooperative action in social dilemmas is challenging due to this tension between unaligned individual short-term incentives and group long-term incentives. One solution is to introduce a social planner whose interventions can resolve the dilemma. In this work we see the application of a social planner to the *collective risk dilemma* (CRD). The CRD captures the setting in which individuals can contribute (*cooperate*) towards some collective target. If the target is not reached, i.e. too many agents did not contribute (*defect*), all agents suffer with some probability (the *risk*). Prior analysis of the CRD showed that there exist cooperative strategies which are *stable steady states* in settings with high risk Santos and Pacheco (2011). This work strengthens this result and shows that these cooperative strategies are also *evolutionary stable strategies*. Furthermore, this work specifically addresses the setting of low *perceived* risk in the population, which a social planner learns to mitigate through economic intervention in a new intertemporal framing of the CRD. Through Q-learning, a budgeted social planner can learn to push players from non-cooperative to cooperative equilibria and improve the level of cooperation.

Contents

Acknowledgments

Finishing this thesis took a village. I am indebted and grateful to all of my mentors and friends who supported me every step of the way.

To my thesis readers: thank you for your generosity and for giving your time to offer feedback.

To my thesis advisors Professor David Parkes and Professor Cliff Taubes: thank you for your unwavering patience, brilliant advice, and calm reassurance throughout this long process. David, thank you for your kind and generous mentorship over the past two years and for inspiring me to pursue the CS concentration and this research.

Thanks also to my Radcliffe Research Partnership mentor Dr. Ana Paiva for getting me started on this project and for her research mentorship. Thanks to Dr. Gianluca Brero for his RL guidance this semester, Peter S. Park for his helpful math consultations, and to Dr. Francisco Santos for helpful conversations about his work last semester. Going back to 2020, thanks to Dr. Sarah Keren for her mentorship during my PRISE fellowship summer and to Saffron Huang for an introduction to this research space through her reading course. I would also like to thank the MD4SG environment working group for many inspiring talks and discussions.

Finally, this thesis would not have been possible without the support and encouragement of my friends and family. Some special shout-outs: Conlan Olson for many thesis-writing accountability/support sessions this semester; Pratap Singh and Mason Meyer for their solidarity in off-cycle CS thesis-writing; Anca Dragulescu, Andrea Lamas-Nino, Anne Lheem, Dhilan Ramaprasad, Mridu Nanda, and Natalia Hajlasz for your generosity (whether through coffee deliveries or co-working sessions) and for being here for me; Amal Mattoo for his patient support and helpful math conversations; my brother Wen for his encouragement and for helping me troubleshoot LaTeX; and my parents for their unconditional support.

1
Introduction

In many situations, it pays for individuals to defect from cooperative behavior — yet if all defect, everyone is worse off. Such scenarios are known as social dilemmas. Because immediate gains for the individual are not aligned with long-term benefit for the group, it is a challenge to sustain cooperative action in these situations.

Social dilemmas such as action on global climate change, where incentives to free-riding behavior can lead to the tragedy of the commons, where individual use of a collective good – or alternatively the lack of maintenance of a collective good – reduces the quality for all Ostrom (1990). We may already be familiar with the Prisoner's dilemma as either a colloquialism or, as we will use it, a standard example in game theory. The story goes: two prisoners accused of some crime are interrogated separately and given the choice to betray their partner (*defect*, D)) or stay silent (*cooperate*, C). If both players betray each other (D, D), both serve 2 years in prison; if both stay silent (C, C), both serve 1 year on a lesser charge. However, if one prisoner betrays while the other stays silent ((C, D) or (D, C), one goes free while the other serves 3 years. There is higher reward for betrayal against a silent partner, so the purely reward-maximizing prisoner will betray their partner. But since both prisoners adopt this strategy, the result is that both prisoners serve more jail time than they would have if they both cooperated. This example demonstrates one type of social dilemma dynamics that can be used to model real-world dilemmas.

In this thesis, we use the *collective risk dilemma* (CRD) to model such real-world dilemmas. The CRD captures the setting in which individuals can contribute (*cooperate*) towards some collective target. If the target is not reached, i.e. too many agents did not contribute (*defect*), all agents suffer with some probability (the *risk*). Compared to previous work in MARL to address sequential social dilemmas Hughes et al. (2018) and Huang (2020), using the CRD as our model offers the additional complexity and richness of an N player game in which players contribute to a public good.

Furthermore, we explore strategic behavior in social dilemmas by introducing a social planner whose interventions can attempt to resolve the dilemma.

1.0.1 Contribution

This thesis makes both expository and research contributions to the analysis of the CRD. In Chapter 2, I give a clear exposition of fundamental game theory ideas and results, principally the existence of a mixed-strategy Nash equilibrium, and I connect the Nash solution concept with that in replicator dynamics and evolutionary stable strategies. This exposition sets the stage for the novel analysis of stability of CRD equilibria presented in Chapter 4. In Chapter 3, I provide a thorough explanation of the Q-learning algorithm and the proof of its convergence, which is both mathematically intriguing and important to understand before using the algorithm in Chapter 6. In Chapter 4, I begin with a survey of the current state of research on the CRD before I provide a new analysis of the MSNE, including the ESS solutions, of the CRD. In Chapter 5, I present a novel model with a social planner to approach improving levels of cooperation the CRD. In Chapter 6, I present our results from implementing this model with both a budgeted (action-constrained) and unbudgeted planner. In Chapter 7, I conclude and discuss future work, in particular extensions using multi-agent reinforcement learning.

2 Learning in games

This exposition aims to serve both as an extended background section and as a cohesive bridging of key mathematical concepts and results needed for studying the collective risk dilemma. This exposition draws from the textbooks Fudenberg (1991), Fudenberg (1998), and the Harvard course CS136 through Parkes and Seuken (2022). First, we review key game theory concepts and the classic result of mixed-strategy equilibrium existence. Then, we cover how Nash equilibria connect to equilibria concepts in replicator dynamics.

2.1 Game theory and Nash equilibria

Game theory offers a framework to analyze the world as strategic interaction among economic, or utility-maximizing agents. A game is defined by its players (or agents), the actions available to those players, and the players' utility functions over possible outcomes.

Definition 1 (Simultaneous move game)**.** *A finite, normal-form, simultaneous move game is defined by:*

- *A finite set of N players $i \in [N] = \{1, 2, \ldots, N\}$*
- *A finite action space $A = A_1 \times A_2 \times \cdots \times A_N$ where action $a_i \in A_i$ is the action played by agent i. Together, $a = (a_1, \ldots, a_n) \in A$ is an action profile.*

- *Utility functions $u_i : A \to \mathbb{R}$ for each agent i, which assigns a payoff for each action profile $a \in A$ (which determines the game's outcome)*

We adopt the perspective that each player aims to maximize their expected utility.

For two player games, it is convenient to represent these payoffs in a *normal-form/matrix representation* which gives the payoffs to each agent for each possible action profile; see 2.1.1.

	C	D
C	-1,-1	-3,0
D	0,-3	-2,-2

Figure 2.1.1: The canonical payoffs for a Prisoner's Dilemma game (discussed in 1).

Players employ *strategies* which specify a distribution over their choice of actions. More precisely,

Definition 2 (Mixed-strategy). *A mixed strategy s_i is a mapping from actions to a distribution $s_i : A_i \to [0, 1]$, $s_i \in \Delta(A_i)$, for agent i where, $\sum_{j \in A_i} s_i(j) = 1$ and $\Delta(A_i)$ is a probability simplex on A_i. Denote the space of mixed strategies for player i as Σ_i.*

A sample strategy in the Prisoner's Dilemma would be to play C 60% of the time and D the other40% of the time, which we would write as $s_1 = (.6, .4)$ for player 1. When there is some action $a \in A_i$ such that $s_i(a) = 1$ (and thus $s_i(b) = 0$ for all $b \in A \setminus \{a\}$), we call this a pure-strategy playing action a. Call the N-tuple of strategies for all agents i (assumed to be sampled independently) a *strategy profile*, $s = (s_1, \ldots, s_N) \in \Sigma$ where $\Sigma = \times_{i \in [N]} \Sigma_i$. Let s_{-i} denote the strategy profile without agent i, or $s_{-i} = (s_1, \ldots s_{i-1}, s_{i+1}, \ldots s_N)$.

When we introduce probabilities over actions, we need a notion of expected utility.

Definition 3 (Expected utility). *Let $p(a)$ give the probability of action profile a and let the strategy profile be s. Then the expected utility to player i is*

$$u_i(s) = \sum_{a \in A} u_i(a) \cdot p(a)$$

Game theory is interested in players' behavior when players' strategies rely on the chosen strategies of other players. We analyze such situations with the idea of *Nash equilibria* and players acting in best response to the actions of others.

Definition 4 (Nash equilibrium). *A mixed-strategy profile s^* is a mixed-strategy Nash equilibrium (MSNE) if for all players i,*

$$u_i(s^*, s^*_{-i}) \geq u_i(s_i, s^*_{-i})$$

for all mixed strategies s_i.

If s^* is a pure-strategy and satisfies the conditions, we call this a pure strategy Nash equilibrium.

The pure strategy Nash equilibria in the Prisoner's dilemma is (D, D), since for either player switching to cooperate from the (D, D) equilibria only decreases that players' payoff from -2 to -3 since we hold the opponent player's equilibrium strategy constant.

While pure-strategy Nash equilibria are not guaranteed to exist, MSNE are.

2.1.1 A fundamental result: the existence of mixed Nash equilibria

The guaranteed existence of a mixed strategy Nash equilibrium makes Nash equilibrium analysis applicable to any game in which randomized play is possible. We use Kakutani's fixed point theorem to show this.

Theorem 5 (Kakutani's fixed point theorem). *The following conditions are sufficient for $r : \Sigma \to \mathcal{P}(\Sigma)$ to have a fixed point (i.e., a $\sigma \in \Sigma$ such that $\sigma \in r(\sigma)$): 1. Σ is a compact, convex, nonempty subset of a finite-dimension Euclidean space, 2. $r(\sigma)$ is nonempty for all σ, 3. $r(\sigma)$ is convex for all σ, and 4. $r(\cdot)$ has a closed graph, i.e. if the sequence $(\sigma^n, \hat{\sigma}^n) \to (\sigma, \hat{\sigma})$ with $\hat{\sigma}^n \in r(\sigma^n)$, then $\hat{\sigma} \in r(\sigma)$.*

Theorem 6 (Nash (1950)). *For every finite, normal-form game there exists a mixed-strategy equilibrium.*

Proof. Define player *i*'s *reaction correspondence*

$$r_i : \Sigma \to P(\Sigma_i), \sigma \mapsto \{s \in \Sigma_i \text{ maximizing } u_i(s, \sigma_{-i})\}$$

That is, r_i maps each strategy profile to the set of mixed strategies which maximize expected payoff to player i given σ_{-i}. Consider the Cartesian product of all r_i:

$$r : \Sigma \to P(\Sigma), \sigma \mapsto \times_{i \in [N]} r_i(\sigma)$$

We see that a fixed point of r would correspond to a Nash equilibrium; $\sigma \in r(\sigma)$ such that $\sigma_i \in r_i(\sigma)$.

We show that indeed r has a fixed point by satisfying the four conditions from Kakutani's fixed point theorem:

1. Σ is a compact, convex, nonempty subset of a finite dimensional Euclidean space. This follows from each Σ_i simply being a simplex of dimension $|A_i| - 1$.

2. $r(\sigma)$ is nonempty for all σ since each u_i is assumed to be linear, and thus continuous; continuous functions over compact sets achieve extrema, so $r(\sigma)$ is non empty.

3. $r(\sigma)$ is convex for all σ.

   Suppose for contradiction that $r(\sigma)$ were not convex. Then there would exist $\sigma' \in r(\sigma)$ and $\sigma'' \in r(\sigma)$ and $\lambda \in (0, 1)$ such that, by the definition of convexity,

$$\lambda\sigma' + (1 - \lambda)\sigma'' \notin r(\sigma).$$

   However, by linearity of the utility function,

$$u_i\left(\lambda\sigma'_i + (1 - \lambda)\sigma''_i, \sigma_{-i}\right) = \lambda u_i(\sigma'_i, \sigma_{-i}) + (1 - \lambda)u_i(\sigma''_i, \sigma_{-i})$$

   which implies that if both σ' and σ'' were best responses, their weighted average would also be a best response, which is a contradiction to non-convexity.

4. $r(\cdot)$ has a closed graph.

   Suppose this were not the case and there were some $\hat{\sigma} \notin r(\sigma)$ such that for $\hat{\sigma}^n \in r(\sigma^n)$, $(\sigma^n, \hat{\sigma}^n) \to (\sigma, \hat{\sigma})$. In particular this implies there is some $\hat{\sigma}_i \notin r_i(\sigma)$ for some player i and thus there exists an $\varepsilon > 0$ and a σ'_i such that $u_i(\sigma'_i, \sigma_{-i}) > u_i(\hat{\sigma}_i, \sigma_{-i}) + 3\varepsilon$. However, u_i is continuous and $(\sigma^n, \hat{\sigma}^n) \to (\sigma, \hat{\sigma})$ for sufficiently large n so

$$u_i(\sigma'_i, \sigma^n_{-i}) > u_i(\sigma'_i, \sigma_{-i}) - \varepsilon > u_i(\hat{\sigma}_i, \sigma_{-i}) + 2\varepsilon > u_i(\hat{\sigma}^n_i, \sigma^n_{-i}) + \varepsilon$$

   which implies σ'_i is strictly better than $\hat{\sigma}^n_i$ against σ^n_{-i}, which contradicts $\sigma^n_i \in r_i(\sigma^n)$.

□

2.2 Replicator dynamics and evolutionary stable strategies

Now that we have modeled strategic behavior with Nash equilibria, we introduce a model based on evolution. We review two key concepts in evolutionary game theory: replicator dynamics and evolutionary stable strategies (ESS). Replicator dynamics are a time-varying model that capture the emulation/imitation of strategies by other agents. ESS is a static property (like Nash equilibria) that captures the notion of resistance to an invading population of strategies.

2.2.1 Replicator Dynamics

For the discussion here, we make a few assumptions about the stage game and the players in the population: 1) that we have a symmetric stage game, 2) that there is a single homogenous population, and 3) that only pure strategies are allowed. However, we can interpret the fractions of players in a population playing a pure strategies as equivalent to the corresponding symmetric mixed strategy over those pure strategies.

Evolutionary dynamics uses similar but different notation than standard game theory, which we outline here. Note that utility $u(s, s')$ is now the utility *of the player playing strategy s*, the first argument, against another strategy s', the second argument.

- $u(s, s')$: payoff to the player using pure strategy s against pure strategy s'
- $u(s, \sigma) := \sum_{s'} \sigma(s')u(s, s')$: payoff to the player using pure strategy s against strategy profile σ; continuous in second variable on space of strategy profiles
- $u(\sigma, \sigma') := \sum_{s} \sigma(s)u(s, \sigma')$: average payoff of players in strategy profile σ against strategy profile σ'; continuous in both variables on space of strategy profiles
- $\varphi_t(s)$: measure of set of players using pure strategy s at time t
- $\theta_t(s) := \frac{\varphi_t(s)}{\sum_{s'} \varphi_t(s')}$: fraction of players using pure strategy s at time t
- $u_t(s) := \sum_{s'} \theta_t(s')u(s, s')$: expected payoff to player using pure strategy s at time t (equivalent to $u(s, \theta_t)$)
- $\overline{u}_t := \sum_s \theta_t(s)u_t(s)$ (equivalent to $u(\theta_t, \theta_t)$)

The intuition for replicator dynamics is that a player with strategy s is matched with a player of strategy s', with s' chosen proportionally to the measure of players in the population with that strategy.

We assume that each player's strategy is fixed, that this strategy is inherited, and that the rate of growth of the population using a strategy is proportional to its expected payoff at time t.

We can now define the dynamical system characterized by

$$\dot{\varphi}_t(s) = \varphi_t(s)u_t(s)$$

We want to work with just θ_t, so we derive what is called the *replicator equation*.

Lemma 7 (The replicator equation). *For the dynamical system characterized by φ_t as above*

$$\dot{\theta}_t(s) = \theta_t(s)(u_t(s) - \overline{u}_t(s))$$

Proof.

$$\begin{aligned}
\dot{\theta}_t(s) &= \frac{d}{dt}\frac{\varphi_t(s)}{\sum_{s'}\varphi_t(s')} \\
&= \frac{\dot{\varphi}_t(s)\sum_{s'}\varphi_t(s')}{\left(\sum_{s'}\varphi_t(s')\right)^2} - \frac{\varphi_t(s)\sum_{s'}\dot{\varphi}_t(s')}{\left(\sum_{s'}\varphi_t(s')\right)^2} \\
&= \frac{\dot{\varphi}_t(s)}{\sum_{s'}\varphi_t(s')} - \left(\frac{\varphi_t(s)}{\sum_{s'}\varphi_t(s')}\right)\left(\frac{\sum_{s'}\dot{\varphi}_t(s')}{\sum_{s'}\varphi_t(s')}\right) \\
&= \frac{\varphi_t(s)u_t(s)}{\sum_{s'}\varphi_t(s')} - \left(\frac{\varphi_t(s)}{\sum_{s'}\varphi_t(s')}\right)\left(\frac{\sum_{s'}\varphi_t(s)u_t(s')}{\sum_{s'}\varphi_t(s')}\right) \\
&= \left(\frac{\varphi_t(s)}{\sum_{s'}\varphi_t(s')}\right)u_t(s) - \left(\frac{\varphi_t(s)}{\sum_{s'}\varphi_t(s')}\right)\sum_{s'}\left(\frac{\varphi_t(s)}{\sum_{s'}\varphi_t(s')}\right)u_t(s') \\
&= \theta_t(s)u_t(s) - \theta_t(s)\overline{u}_t(s) \\
&= \theta_t(s)(u_t(s) - \overline{u}_t(s))
\end{aligned}$$

□

We first observe that the measure of set of players using strategy *s* will increase if expected payoff is positive and decrease if expected payoff is negative. We also observe that the population share of strategy *s* will increase if its payoff is above average and will decrease if its payoff is below average.

The next definition is a key concept in the study of dynamical systems.

Definition 8 (Steady state)**.** *A steady state is $\hat{\theta}$ such that if $\theta_t = \hat{\theta}$ then $\dot{\theta}_t(s) = 0$ for all s (i.e., population shares of strategies are constant).*

Every Nash equilibrium is a steady state: in Nash equilibrium all players use the same strategy (by symmetry between players), so $u_t(s) = \overline{u}_t(s)$, meaning $\dot{\theta}_t(s) = 0$. However, not every steady state is a Nash equilibrium: steadiness holds whenever all players have the same strategy, regardless of optimality (no entry of new strategies in this equilibrium).

When studying a steady state of a dynamical system, we are also interested in how the system behaves around that state.

Definition 9. *A stable steady state is a steady state $\hat{\theta}$ such that for every neighborhood $U \ni \hat{\theta}$ there is a neighborhood $U_1 \subset U$ such that if $\theta_0 \in U_1$ then $\theta_t \in U$ for $t > 0$ (i.e., if θ starts close enough to $\hat{\theta}$, it remains close by).*

The following result relates the local notion of stable steadiness to the global notion of Nash equilibrium.

Theorem 10. *If a steady state is stable, then it is a Nash equilibrium.*

Proof. Suppose $\hat{\theta}$ is a steady state, but the corresponding strategy profile σ^* is not a Nash equilibrium. Then there exists some pure strategies $s \in \text{supp}(\sigma^*)$ and s' with $u(s', \sigma^*) > u(s, \sigma^*)$. So there is some $\varepsilon > 0$ such that $u(s', \sigma^*) > u(s, \sigma^*) + 3\varepsilon$.

By continuity of u in the second variable, there is some neighborhood U of σ^* such that $|u(s, s^*) - u(s, \sigma^*)| < \varepsilon$ and $|u(s', s^*) - u(s', \sigma^*)| < \varepsilon$ for all $s^* \in U$.

Thus, we have $u(s', s^*) > u(s, s^*) + \varepsilon$ for all $s^* \in U$. Thus,

$$\begin{aligned} \theta_t \in U &\implies u_t(s') - \overline{u}_t > u_t(s) - \overline{u}_t + \varepsilon \\ &\implies \frac{\dot{\theta}_t(s')}{\theta_t(s')} > \frac{\dot{\theta}_t(s)}{\theta_t(s)} + \varepsilon \\ &\implies \frac{d}{dt}\ln(\theta_t(s')) > \frac{d}{dt}\ln(\theta_t(s)) + \varepsilon \\ &\implies \frac{d}{dt}\ln\frac{\theta_t(s')}{\theta_t(s)} > \varepsilon \end{aligned}$$

If $\hat{\theta}$ were stable, then $\theta_t \in U$ for $t > 0$. Then $\lim_{t\to\infty} \frac{\theta_t(s')}{\theta_t(s)} = \infty$ so $\lim_{t\to\infty} \theta_t(s) = 0$. Note that U could be chosen such that $s^*(s) > \varepsilon' > 0$ for all $s^* \in U$, so $\theta_t \notin U$ for some $t > 0$. This is a contradiction, so $\hat{\theta}$ cannot be stable. □

This argument relied only on the fact that growth rates of strategies are increasing functions of payoff; no additional structure of replicator dynamics was necessary.

To explore the converse, we will employ the slightly different notion of asymptotic stability.

Definition 11 (Asymptotically stable steady states)**.** *A steady state $\hat{\theta}$ is asymptotically stable if for every neighborhood $U \ni \hat{\theta}$ there is a neighborhood $U_1 \subset U$ such that if $\theta_0 \in U_1$ then $\lim_{t\to\infty} \theta_t = \hat{\theta}$ (i.e., if θ starts close enough to $\hat{\theta}$, it will converge to it).*

It turns out that asymptotic stability refines the notion of Nash equilibrium. We introduce some more concepts to explain how.

- A steady state $\hat{\theta}$ is *(locally) isolated* if there exists a neighborhood $U \ni \hat{\theta}$ containing no other steady states.

- A *perturbed game* is one in which only totally mixed strategies are allowed (each strategy is played with positive probability).

- A strategy s in a game G is *trembling hand perfect* if there is a sequence of perturbed games converging to G for which there is a sequence of Nash equilibria converging to s (intuitively, equilibrium still roughly holds even if players diverge slightly with small enough probability).

Theorem 12. *If a stable steady state is asymptotically stable, then it corresponds to a Nash equilibrium that is trembling-hand perfect and isolated.*

See Bomze (1986) for the proof.

This theorem has a number of consequences for us. In games that do not yield Nash equilibria that are trembling-hand perfect and isolated, asymptotic stability will not hold under our model, so such equilibria will be hard to find. The next section will address a stronger notion of stability than asymptotic stability,

2.2.2 Evolutionary stable strategies

Informally, we are interested in an equilibrium that can repel invading evolutionary strategies. Although it is formally a static property, we may intuit it dynamically as an equilibrium where invaders die off immediately.

Definition 13 (Evolutionary stable strategy (ESS))**.** *If a population is at some profile σ and a small ε play σ', σ is an ESS if the resulting mixture has lower payoff than the existing population under σ, or*

$$u(\sigma, (1-\varepsilon)\sigma + \varepsilon\sigma') > u(\sigma'(1-\varepsilon)\sigma + \varepsilon\sigma')$$

for all sufficiently small ε.

Lemma 14. *A profile σ corresponds to an ESS if and only if one of the following holds*

- $u(\sigma, \sigma) > u(\sigma, \sigma)$
- $u(\sigma, \sigma) = u(\sigma, \sigma)$ *and* $u(\sigma, \sigma) > u(\sigma, \sigma)$

Proof. This follows directly from expanding the expressions in Definition 14 by linearity of expectations. □

We are interested in relating this new notion to the previous equilibrium refinements, in particular asymptotically stable strategies. First, we introduce local Lyapunov functions for equilibria. Then we show if there exists a local Lyapunov function for equilibrium σ, then σ is an asymptotically stable equilibrium. Then we can prove that if σ is an ESS, then it is an ASSS by exhibiting such a function.

Definition 15 (local Lyapunov)**.** *A local Lyapunov function V for an equilibrium σ is a real-valued C^1 function on a neighborhood $U \ni \sigma$ satisfying*

- $V \geq 0$ *and* $V(\theta) = 0 \iff \theta = \sigma$
- $\frac{d}{dt}V(\theta_t) < 0$ *for* $\theta_t \neq \sigma$ *and* $\frac{d}{dt}V(\theta_t) = 0$ *for* $\theta_t = \sigma$

Lemma 16. *If there exists a local Lyapunov function for equilibrium σ, then σ is a stable equilibrium.*

Proof. Choose $\varepsilon > 0$ such that $\overline{B_\varepsilon(\sigma)} \subset U$. Then V attains a minimum m on the (compact) boundary $\partial\overline{B_\varepsilon(\sigma)}$, and $m > 0$ since $V > 0$ on $U \setminus \{\sigma\}$. Since V is continuous, we can pick $\delta > 0$ such that $V(\theta) < m$ for $\theta \in B_\delta(\sigma)$.

We will show that $\theta_0 \in B_\delta(\sigma) \implies \theta_t \in B_\varepsilon(\sigma)$ for all t. Assume for the sake of contradiction that this fails. Then by continuity there is some $T > 0$ such that $\theta_T \in \partial\overline{B_\varepsilon(\sigma)}$ (i.e., the trajectory must cross the boundary as it exits the ball), so $V(\theta_T) \geq m$. However, $\frac{d}{dt}V(\theta_t) < 0$ for all $t < T$, so $V(\theta_T) < V(\theta_0) < m$, which is a contradiction. □

Theorem 17. *If there exists a local Lyapunov function for equilibrium σ, then σ is an asymptotically stable equilibrium.*

Proof. By Lemma 16, σ is stable. Thus, there is some r such that $\theta_0 \in B_r(\sigma) \implies \theta_t \in \theta_0 \in B_R(\sigma) \subset U$ for all $t \geq 0$. And to show $\theta_t \in B_\varepsilon(\sigma)$ for all $t \geq T$, it suffices to show $\theta_T \in B_\delta(\sigma)$ for some δ.

Assume for the sake of contradiction that $\theta_t \in \overline{B_R(\sigma)} \setminus B_\delta$ for all $t \geq 0$. By continuity and compactness, $\frac{d}{dt}V(\theta_t)$ attains a maximum $-\mu$, and $-\mu < 0$ since $\frac{d}{dy}V(\theta_t) < 0$ on U. Then we have

$$\begin{aligned} V(\theta_T) &= V(\theta_0) + \int_0^T \frac{d}{dt}V(\theta_t)dt \\ &\leq V(\theta_0) + \int_0^T -\mu dt \\ &= V(\theta_0) - T\mu \end{aligned}$$

For $T > \frac{V(\theta_0)}{\mu}$ this implies $V(\theta_T) < 0$, which is a contradiction. Thus, for all t large enough, θ_t is contained inside all $B_\varepsilon(\sigma)$, so $\lim_{t\to\infty} \theta_t = 0$. □

Theorem 18. *If σ is an ESS, then it is an ASS.*

Proof. We construct a local Lyapunov function as follows.

$$E_\sigma(\theta) := 1 - \prod_s \left(\frac{\theta(s)}{\sigma(s)}\right)^{\sigma(s)}$$

To see that E_σ has a unique global minimum at σ, note that, with the inequality resulting from Jensen's inequality and the concavity of log,

$$\begin{aligned}\log(1 - E_\sigma(\theta)) &= \log\left(\prod_s \left(\frac{\theta(s)}{\sigma(s)}\right)^{\sigma(s)}\right) \\ &= \sum_s \sigma(s)\log\left(\frac{\theta(s)}{\sigma(s)}\right) \\ &\leq \log\left(\sum_s \sigma(s)\frac{\theta(s)}{\sigma(s)}\right) \\ &= \log\left(\sum_s \theta(s)\right) \\ &= 0\end{aligned}$$

Thus, $E_\sigma(\theta) \geq 0$. And the equality case of Jensen's inequality holds if and only if $\frac{\theta(s)}{\sigma(s)}$ is the same for all s, and by normalization that implies $\theta = \sigma$.

To see that E_σ decreases with time near σ, we compute

$$\begin{aligned}\frac{\frac{d}{dt}(E_\sigma(\theta_t))}{1 - E_\sigma(\theta_t)} &= \frac{-\frac{d}{dt}(1 - E_\sigma(\theta_t))}{1 - E_\sigma(\theta_t)} \\ &= -\frac{d}{dt}\log\left(1 - E_\sigma(\theta_t)\right) \\ &= -\frac{d}{dt}\sum_s \sigma(s)\log\theta_t(s) \\ &= -\sum_s \sigma(s)\frac{\dot{\theta}_t(s)}{\theta_t(s)} \\ &= -\sum_s \sigma(s)[u_t(s) - \overline{u}_t] \\ &= -\sum_s \sigma(s)[u(s,\theta) - u(\theta,\theta)] \\ &= u(\theta,\theta) - u(\sigma,\theta)\end{aligned}$$

By Lemma 14 and continuity of u, we have $u(\theta,\theta) < u(\sigma,\theta)$ for θ in some neighborhood $U \ni \sigma$. Combined with the fact $1 - E_\sigma(\theta_t) > 0$, we have $\frac{d}{dt}E_\sigma(\theta_t) < 0$ on U.

□

In summary, we have established the following broad relationships between notions of equilibria and stability which will be important for my analysis of the CRD in Chapter 4.

$$\begin{tikzcd}
\text{ESS} & \subseteq & \text{Asymp. SSS} & \subseteq & \text{SSS} & \subseteq & \text{NE} & \subseteq & \text{SS} \\
& \text{Thm. 18} \arrow[u] & & & & \text{Thm. 10} \arrow[u] & & &
\end{tikzcd}$$

Figure 2.2.1: An overview of the relations between the types of equilibria that we have established: in addition to the normal abbreviations SS stands for steady state and SSS stands for stable steady state and.

3

Q-learning

3.1 Q-learning

In this section, we 1) explain the Q-learning algorithm and 2) prove its convergence to an optimal solution.

We first define a *Markov decision process* (MDP) which is central to the Q-learning model and algorithm. The MDP gives us a framework for modeling an agent's decision making in an environment in which both randomness in the environment and decisions (actions) change outcomes. The notation used in the definition below is used throughout this chapter.

Definition 19 (Markov decision process (MDP)). *A Markov decision proces (MDP) is defined by X, a finite set of states; A, a finite set of actions, $P_a(x, y)$, the transition probability for action a from state x to y for $x, y \in S$, and the reward function $r(x, a, y)$ which gives the specific reward from action a causing the transition from state x to y. Assume r is bounded and deterministic.*

A classic figure, taken from Sutton (2018) so forgive the slight capitalization difference in notation, helps us visualize the process.

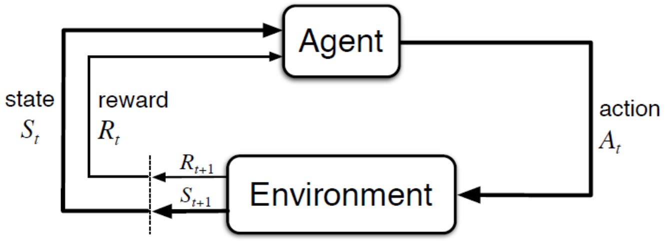

Figure 3.1.1: The agent-environment interaction in an MDP

3.1.1 Q-learning algorithm

This exposition draws from the treatment of Q-learning in the textbook Sutton (2018).

Q-learning is an off-policy, temporal difference control algorithm. The algorithm learns a function Q which approximates the optimal action-value function independent of the policy that is followed (beyond the fact that the policy dictates which state-action pairs are visited). The only assumptions needed for convergence are that all state-action pairs are visited infinitely many times and some typical stochastic approximation conditions on the learning rate; we discuss these conditions in depth in the next section 3.1.2.

Q-learning applies an *update rule* to approach a determination of an optimal value function. At its heart the algorithm is a alternate expression of the Bellman equation, which includes both immediate reward and discounted future values.

$$Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \alpha_t(x_t, a_t)[r_t + \gamma \max_{b \in A} Q_t(x_{t+1}, b) - Q_t(x_t, a_t)]$$

To illustrate how this update rule is applied to the MDP, we provide this procedural outline:

Algorithm 1 Q-learning algorithm

```
Algorithm parameters: step size $\alpha \in [0, 1)$, small $\varepsilon > 0$ **repeat**
    **for** *each episode* **do**
        Initialize $S$ **for** *each step of episode* **do**
            Choose $A$ from $S$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy) Take action $A$, observe $R, S'$
            $Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \max_a Q(S', a) - Q(S, A)]$ $S \leftarrow S'$
        **end**
    **end**
**until** *$S$ is terminal/have reached maximum $t$*;
```

Typical implementations of Q-learning employ an *ε-greedy* approach: with probability ε, the algorithm *explores* a random action, otherwise it *exploits* the action with the maximum Q value for a given state.

3.1.2 Convergence of Q-learning

The convergence of Q-learning was introduced by Watkins in 1989 and proved by Watkins and Dayan in 1992 Wat. The result was generalized independently in 1993/1994 by Jaakkola et al. (1993) and Tsitsiklis (1994) by using ideas from stochastic approximation theory rather than using a construction specific to Q-learning as in Wat. Though we only discuss Q-learning here, the proof of Jaakkola et al. (1993) applies more generally. [1] This exposition draws from Melo (2021) and Jaakkola et al. (1993) and expands upon their exposition.

The optimal Q-function

Q-learning makes continuous updates to learn the *optimal Q-function*. We first formalize some key definitions to understand the optimal Q-function.

The value of a state x is the expected discounted reward from a sequence of actions starting from x:

Definition 20 (Value of a state)**.** *Let the value of a state $x \in X$ for a sequence of actions $\{A_t\}$ be represented by $J : X \times A^{\mathbb{N}} \to \mathbb{R}$,*

$$J(x, \{A_t\}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(X_t, A_t) | X_0 = x\right]$$

where X_t, A_t are random variables reflecting the state and action in time step t respectively, R is shorthand for $\mathbb{E}[R(X_t, A_t)] = \sum_{y \in \mathcal{X}} P_{A_t}(x, y) \cdot r(X_t = x, A_t = a, X_{t+1} = y)$ and the discount factor $\gamma \in (0, 1)$.

We use this value of a state to define the optimal value function, which simply takes the maximum expected value of a state over all possible sequences of actions leaving x.

Definition 21 (Optimal value function)**.** *$V^*(x)$ is the optimal value function*

$$V^*(x) = \max_{\{A_t\}} \sum_{t=0}^{\infty} \mathbb{E}\left[\gamma^t R(X_t, A_t) | X_0 = x\right]$$

Crucially, this formulation gives us the following identity, which we quickly derive:

Lemma 22.

$$V^*(x) = \max_{a \in A} \sum_{y \in X} P_a(x, y)[r(x, a, y) + \gamma V^*(y)]$$

Proof. Consider $J(x, \{A_t\})$ for some starting state x and action sequence $\{A_t\}$. We expand on the first

[1] Jaakkola et. al extend their proof for convergence of Q-learning to $TD(\lambda)$ learning, which unifies Monte Carlo methods and temporal difference learning.

iteration, apply linearity of expectation, and re-index to express the value of in a convenient form:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R(X_t,A_t)|X_0=x\right] &= \sum_{y\in\mathcal{X}} P_{A_0}(x,y)\left[r(x,A_0,y)+\mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^t R(X_t,A_t)|X_1=y\right]\right]\\
&= \sum_{y\in\mathcal{X}} P_{A_0}(x,y)\left[r(x,A_0,y)+\gamma\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R(X_t,A_{t+1})|X_0=y\right]\right]\\
&= \sum_{y\in\mathcal{X}} P_{A_0}(x,y)\left[r(x,A_0,y)+\gamma J(y,\{A_t\}_{t\geq 1})\right]
\end{aligned}$$

Now, we plug this back into $V^*(x)$, and by independence of $A_0, A_{t\geq 1}$, and after some re-labeling:

$$\begin{aligned}
V^*(x) &= \max_{\mathcal{A}_t}\sum_{y\in\mathcal{X}} P_{A_0}(x,y)\left[r(x,A_0,y)+\gamma J(y,\{A_t\}_{t\geq 1})\right]\\
&= \max_{A_0\in\mathcal{A},\mathcal{A}_{t\geq 1}}\sum_{y\in\mathcal{X}} P_{A_0}(x,y)\left[r(x,A_0,y)+\gamma J(y,\{A_t\}_{t\geq 1})\right]\\
&= \max_{A_0\in\mathcal{A}}\sum_{y\in\mathcal{X}} P_{A_0}(x,y)\left[r(x,A_0,y)+\max_{\mathcal{A}_{t\geq 1}}\gamma J(y,\{A_t\}_{t\geq 1})\right]\\
&= \max_{a\in\mathcal{A}}\sum_{y\in\mathcal{X}} P_a(x,y)\left[r(x,a,y)+\gamma V^*(y)\right]
\end{aligned}$$

□

We see the similarity of this value function with the Q-learning update function. We now have the pieces to define the optimal Q^* function; formally,

Definition 23 (Optimal Q-function). *The optimal Q-function is*

$$Q^*(x,a)=\sum_{y\in X}P_a(x,y)[r(x,a,y)+\gamma V^*(y)]$$

We now see that the optimal Q-function Q^* is a rephrasing of this identity so that $V^*(x)=\max_{a\in\mathcal{A}}Q^*(x,a)$, which makes the problem of determining the optimal Q-function equivalent to the value optimization RL control problem. The mathematical convenience of using the Q function is that the maximum appears inside the expectation in the Q update function.

Return to the proof of Q-learning convergence

To prove converge of Q-learning to Q^*, it will be helpful to show that Q^* is a fixed point of a function which can be applied to the update rule. The recursive identity of V^* leads us to consider the following operator, call it H, over the space of possible Q functions $\{q|q:X\times A\to\mathbb{R}\}$:

$$H(q(x,a)) = \sum_{y \in X} P_a(x,y)[r(x,a,y) + \gamma \max_{b \in A} q(y,b)]$$

Lemma 24. *H is a contraction in the sup-norm*[2].

Proof. We want to show $||H(q_1) - H(q_2)||_\infty \leq \gamma ||q_1 - q_2||_\infty$. Quickly expanding,

$$\begin{aligned}
||H(q_1) - H(q_2)||_\infty &= \max_{x,a \in X,A} \left[\sum_{y \in X} P_a(x,y)[r(x,a,y) + \gamma \max_{b \in A} q_1(y,b) - r(x,a,y) - \gamma \max_{b \in A} q_2(y,b)] \right] \\
&= \max_{x,a \in X,A} \gamma \left[\sum_{y \in X} P_a(x,y)[\max_{b \in A} q_1(y,b) - \max_{b \in A} q_2(y,b)] \right] \\
&\leq \max_{x,a \in X,A} \gamma \sum_{y \in X} P_a(x,y) \max_{z,b} [q_1(z,b) - q_2(z,b)] \\
&= \max_{x,a} \gamma \sum_{y \in X} P_a(x,y) ||q_1 - q_2||_\infty \\
&= \gamma ||q_1 - q_2||_\infty
\end{aligned}$$

□

Theorem 25. *Q^* is a fixed point of H*

Proof. Since H is a contraction by 24, by the Banach fixed point theorem there will be a unique fixed point. We demonstrate that Q^* is this fixed point:

$$\begin{aligned}
H(Q^*)(x,a) &= \sum_{y \in X} P_a(x,y)[r(x,a,y) + \gamma \max_{b \in A} \sum_{z \in X} P_b(x,z)[r(x,b,z) + \gamma V^*(z)]] \\
&= \sum_{y \in X} P_a(x,y)[r(x,a,y) + \gamma V^*(y)] \\
&= Q^*(x,a)
\end{aligned}$$

□

Finally before proving convergence, we present a result from stochastic approximation. The proof given in Jaakkola et al. (1993) is quite technical and is not presented here.

Theorem 26. *Jaakkola et al. (1993) The random process $\{\Delta_t\}$ taking values in $\mathbb{R}^n$ and defined as*

$$\Delta_{t+1}(x) = (1 - a_t(x))\Delta_t(x) + a_t(x)F_t(x)$$

[2] Note that the choice of norm is for convenience but, by the equivalence of norms for finite dimensional vector spaces over complete valued fields (e.g. the function space we consider), is not limiting.

converges to zero with probability 1 under the following assumptions: let W be some weighted maximum norm,

- $0 \leq \alpha_t \leq 1$, $\sum_t \alpha_t(x) = \infty$ *and* $\sum_t \alpha_t^2(x) < \infty$
- $\|\mathbb{E}[F_t(x)|\mathcal{F}_t]\|_W \leq \gamma \|\Delta_t\|_W$ *with* $\gamma < 1$
- $Var[F_t(x)|\mathcal{F}_t] \leq C(1 + \|\Delta_t\|_W^2)$

We now have the pieces to prove the convergence of Q-learning.

Theorem 27 (Convergence of Q-learning)**.** *Given a finite MDP* $(\mathcal{X}, \mathcal{A}, P, r)$*, the Q-learning algorithm with the update rule*

$$Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \alpha_t(x_t, a_t)[r_t + \gamma \max_{b \in \mathcal{A}} Q_t(x_{t+1}, b) - Q_t(x_t, a_t)]$$

and with learning rate properties for all $(x, a) \in \mathcal{X} \times \mathcal{A}$

$$\sum_t \alpha_t(x, a) = \infty, \quad \sum_t \alpha_t^2(x, a) < \infty$$

converges with probability 1 to the optimal Q-function.

Proof. To apply 3.1.2 we rewrite the Q-learning update rule as

$$Q_{t+1}(x_t, a_t) = (1 - \alpha_t(x_t, a_t))Q_t(x_t, a_t) + \alpha_t(x_t, a_t)[r_t + \gamma \max_{b \in A} Q_t(x_{t+1}, b)]$$

Let $\Delta_t(x, a) = Q_t(x, a) - Q^*(x, a)$. Then

$$\Delta_t(x_t, a_t) = (1 - \alpha_t(x_t, a_t))\Delta_t(x_t, a_t)) + \alpha_t(x, a)[r_t + \gamma \max b \in AQ_t(x_{t+1}, b) - Q^*(x_t, a_t)]$$

and letting $X(x, a)$ be a random sample state obtained from the relevant Markov chain $(\mathcal{X}, P_a)$, we can write

$$F_t(x, a) = r(x, a, X(x, a)) + \gamma \max_{b \in A} Q_t(y, b) - Q^*(x, a)$$

so we have

$$\mathbb{E}[F_t(x, a)|\mathcal{F}_t] = \sum_{y \in \mathcal{X}} P_a(x, y)[r(x, a, X(x, a)) + \gamma \max_{b \in A} Q_t(y, b) - Q^*(x, a)]$$

which is conveniently equivalent to

$$\mathbb{E}[F_t(x, a)|\mathcal{F}_t] = (H(Q_t))(x, a) - Q^*(x, a)$$

and since Q^* is a fixed point of H

$$\mathbb{E}[F_t(x,a)|\mathcal{F}_t] = (H(Q_t))(x,a) - (H(Q^*))(x,a)$$

now from the fact that H is a contraction in the sup norm, it is immediate that

$$\|\mathbb{E}[F_t(x,a)|\mathcal{F}_t]\|_\infty \leq \gamma\|Q_t - Q^*\|_\infty = \gamma\|\Delta_t\|_\infty$$

and finally we compute variance.

$$\begin{aligned}
\operatorname{Var}[F_t(x)|\mathcal{F}_t] &= \mathbb{E}[(r(x,a,X(x,a)) + \gamma \max_{b\in A} Q_t(y,b) - Q^*(x,a) - (H(Q_t))(x,a) + Q^*(x,a))^2] \\
&= \mathbb{E}[(r(x,a,X(x,a)) + \gamma \max_{b\in A} Q_t(y,b) - (H(Q_t))(x,a))^2] \\
&= \operatorname{Var}[r(x,a,X(a,a)) + \gamma \max_{b\in A} Q_t(y,b)|\mathcal{F}_t]
\end{aligned}$$

and since r is bounded we can conclude that for some constant C

$$\operatorname{Var}[F_t(x)|\mathcal{F}_t] \leq C(1 + \|\Delta_t\|_W^2)$$

so by Δ_t converges to zero with probability 1 so Q_t converges to Q^* with probability 1. □

Proving convergence has given us a closer look at the theoretical underpinnings of the Q-learning algorithm and an understanding of the conditions needed for convergence. In practice we can satisfy these conditions by: for $\sum_t \alpha_t = \infty$, visiting each state-action pair infinitely often through an ε-greedy policy, and for $\sum_t \alpha_t^2 < \infty$, reducing the learning rate over time. In my implementation of Q-learning in Chapter 6, choices for these parameters are discussed.

4

The Collective Risk Dilemma

The remainder of this thesis addresses the *collective risk dilemma*. This chapter gives a thorough exposition of the collective risk dilemma (CRD), contextualizing the CRD within the larger body of work on social dilemmas, unifying many distinct treatments of the CRD in the literature, and providing precise analysis of its equilibria.

In the next two chapters (5, 6), I introduce an intertemporal extension of the CRD allowing for the intervention of a social planner and present the results of using Q-learning for the planner.

4.1 Background: social dilemmas

Social dilemmas are well-studied in game theory and capture settings in which Nash equilibria characterized by defective behavior result in outcomes below the *Pareto optimal*, which is the property of an outcome where no player can be strictly better off without making at least one other individual strictly worse off Ostrom and Walker (2003).

The following definition is adapted from the broader definition found in Hughes et al. (2018).

Definition 28 (2 action, N-player social dilemma)**.** *An N-player social dilemma is defined by a tuple (G, σ), with G an N player game with 2 actions, cooperate, C, and defect, D, available to each player, and $\sigma = (\underbrace{C, \dots C}_{\ell}, \underbrace{D, \dots D}_{m})$ representing ℓ cooperators and m defectors, $\ell + m = N$. Given that there are ℓ*

cooperators, denote the expected payoff from C (D) as $R_C(\ell)$ $(R_D(\ell))$. Then (G, σ) is a social dilemma if

- *Mutual cooperation is preferred to mutual defection, $R_C(N) > R_D(0)$*
- *Mutual cooperation is preferred to being exploited with defectors, $R_C(N) > R_C(0)$*
- *Both or one of the fear or greed property is satisfied:*
  - *Fear: mutual defection is preferred to cooperating with defectors, $R_D(i) > R_C(i)$ for sufficiently small i*
  - *Greed: defecting against cooperators is preferred to mutual cooperation, $R_D(i) > R_C(i)$ for sufficiently large i*

Revisiting the 2-player games Stag Hunt and Prisoner's dilemma, we see from their payoff matrices that they are social dilemmas. We later show in Lemma 30 that some collective risk dilemmas are social dilemmas.

4.2 The collective risk dilemma

The N-player game which we study is called the *collective risk dilemma* (CRD), introduced in Milinski et al. (2008). The CRD is a modified *threshold game*, which simply refers to a multiplayer game defined by some minimum number of cooperators (*threshold*) for all players to receive some benefit.

Definition 29. *The collective risk dilemma (CRD) is an N-player simultaneous move game $CRD(N, M, c, r, b)$ specified by: N players, an integer threshold $M < N$, a fractional (percent) cost of cooperation $c \in (0, 1)$, a probability representing level of risk $r \in (0, 1)$, and a starting endowment b symmetric[1] for all players. Each player in the CRD has two actions available: either cooperate (C), paying a fraction of their endowment $c \cdot b$, or defect (D), paying nothing. If the total number of cooperators is below the threshold, then with probability r a risky event occurs and players lose their remaining endowment.*

Note that since there are only two actions, C, D, when we refer to a strategy we need only specify the probability of cooperating, referred to in this section as x. We often consider the incentives for a single agent in the CRD and as such are interested in the number of *other* cooperators, which we model as $k \sim \text{Bin}(N-1, x)$, representing the other $N-1$ players each adopting strategy x.

We see in Figure 4.2.1 that the expected payoffs depend non-linearly on the number of cooperators, which especially distinguishes analysis of the CRD from the analysis of two-player games.

[1] Other treatments of the CRD consider the effective of inequality in endowments, see

	k$\geq M$	k=M-1	k<M-1
C	b(1-c)	b(1-c)	b(1-c)(1-r)
D	b	b(1-r)	b(1-r)

Figure 4.2.1: Expected payoffs in the CRD.

Now I show that some CRDs are social dilemmas.

Lemma 30. *For $c < r$, $M > 1$, $CRD(N, M, c, r, b)$ satisfies the definition of a N-player social dilemma.*

Proof. We demonstrate each condition:

- With N cooperators, cooperators have expected payoff $b(1 - c)$. With 0 cooperators, defectors have expected payoff $b(1 - r)$. Since $c < r$, $R_C(N) = b(1 - r) < b(1 - c) = R_D(N)$, so mutual cooperation is preferred to mutual defection.

- With N cooperators, cooperators have expected payoff $b(1 - c)$. With 0 cooperators, cooperators have expected payoff $b(1 - c)(1 - r)$. Since $R_C(N)b(1 - c) > b(1 - c)(1 - r) = R_C(0)$, mutual cooperation is preferred to being exploited with defectors.

- Both Greed and Fear hold since:
  - Defecting against cooperators is preferred to mutual cooperation for $k > M$ other cooperators, $R_D(k) = b > b(1 - c) = R_C(k)$ satisfying Greed.
  - Mutual defection is preferred to cooperating with defectors for $k < M - 1$ other cooperators, $R_D(k) = b(1 - r) > b(1 - c)(1 - r) = R_C(k)$ satisfying Fear.

□

However, the framework of a social dilemma does not fully capture the CRD; as is evident from the Schelling diagram, the CRD has unique behavior for $k = M - 1$, where the number of other cooperators is on the threshold.

We can also visualize the CRD dynamics with the Schelling diagram[2] in Figure 4.2.2.

[2] Schelling diagrams are useful representations of games, see Schelling (1973) for more.

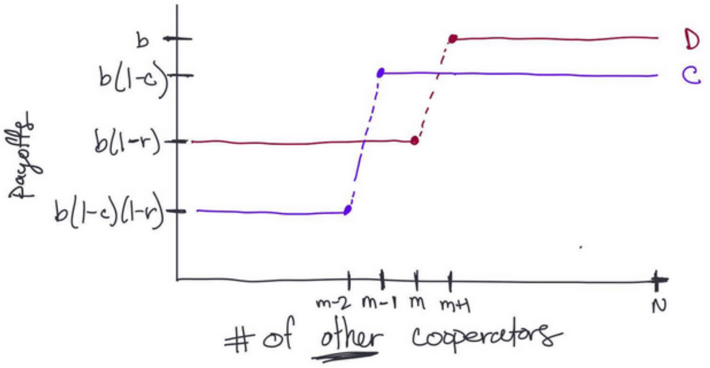

Figure 4.2.2: Schelling diagram for the CRD to illustrate the payoff dynamics given in the proof above.

Now that we have reviewed the basics of the CRD, we will next highlight some related work. Since its introduction in Milinski et al. (2008), the CRD has primarily been studied from a replicator dynamics perspective and with behavioral experiments. Using replicator dynamics, Santos and Pacheco (2011) finds that cooperation will be maximized when risk is high and N is small, and provides analysis of the stable and unstable steady states. Numerous behavioral experiments which model the CRD have verified some of these observations, strengthening the case for applicability of the CRD to real-world social dilemmas Fernández Domingos et al. (2021).

4.3 CRD Nash equilibria characterization

Prior work by Santos and Pacheco (2011) characterized the steady states of the CRD at all-defect $x = 0$, all-cooperate state at $x = 1$, "on the threshold" at $\overline{x} = M - 1/N - 1$, and two mixed strategy cooperative states x_L, x_R with $x_L < x_R$. I show that the all-defect steady state, the steady state at $\overline{x} = M - 1/N - 1$, and x_L, x_R are Nash equilibria, which is consistent with what we know from 2 – Nash equilibria are a subset of steady states.

There are two cases of pure strategy Nash which are intuitive.

1. The first pure NE is all defect; for $M > 1$, switching to cooperate when $k = 0$ only decreases a player's payoff from $b(1 - r)$ to $b(1 - r)(1 - c)$.

2. The second pure NE holds for $c < r$ and is "on the threshold" when the number of total cooperators is M. In this equilibrium, if a player is defecting then switching to cooperate only decreases their payoff from b to $b(1 - c)$. However, if a player is playing action C, then switching to D brings the total number of cooperators below the threshold, switching that player's payoff from $b(1 - c)$ to $b(1 - r)$– this is only a Nash equilibrium when $c < r$. Note that this constraint is the same as in Lemma 30.

We now directly derive the mixed Nash equilibria and show that those two mixed strategy cooperative states. We first assume symmetric agents and strategies, so we can model the total number of cooperators as sampled from a $\text{Bin}(N, x)$ distribution where x is the probability of cooperation for the symmetric strategy.

The first quantity in which we are interested are the expected payoffs for playing C, D for some agent i. We present this in terms of the number of *other* cooperators k, $p_c(x) = P(k \geq M - 1)$ (probability that threshold is reached with agent i's cooperation), and $p_d(x) = P(k \geq M)$ (probability that threshold is reached without agent i's cooperation). Following the CRD specification,

$$\pi_i(C) = b(1 - c)\left[p_c(x) + (1 - r)(1 - p_c(x))\right]$$

$$\pi_i(D) = b\left[p_d(x) + (1 - r)(1 - p_d(x))\right]$$

A particular value of interest is the difference between these payoffs. Let $k \sim \text{Bin}(N - 1, x)$. Then

$$\pi_i(C) - \pi_i(D) = b \cdot (rP(k = M - 1) - c(1 - rP(k \leq M - 2)))$$

See 8 for the algebra.

We can then derive the mixed strategy Nash equilibria directly; an agent will mix when the expected payoffs are equivalent and $\pi_i(C) - \pi_i(D) = 0$. Note that the value of the starting endowment does not affect the MSNE since we assume symmetric endowments.

Theorem 31 (Characterization of CRD MSNE). *Let the equilibrium strategy of the MSNE be $(x, 1 - x)$, where x represents the probability of playing C. Let $k \sim Binom(N - 1, x)$. Then x must satisfy*

$$c = \frac{rP(k = M - 1)}{1 - rP(k \leq M - 2)}$$

Proof. We simply manipulate

$$\begin{aligned}
0 &= \pi_i(C) - \pi_i(D) \\
0 &= b \cdot (rP(k = M - 1) - c + crP(k \leq M - 2)) \\
0 &= rP(k = M - 1) - c + crP(k \leq M - 2) \\
c &= \frac{rP(k = M - 1)}{1 - rP(k \leq M - 2)}
\end{aligned}$$

□

We inspect the MSNE graphically. When there is a $CRD(N, M, c, r)$, x satisfying Theorem 31 there are two MSNE.

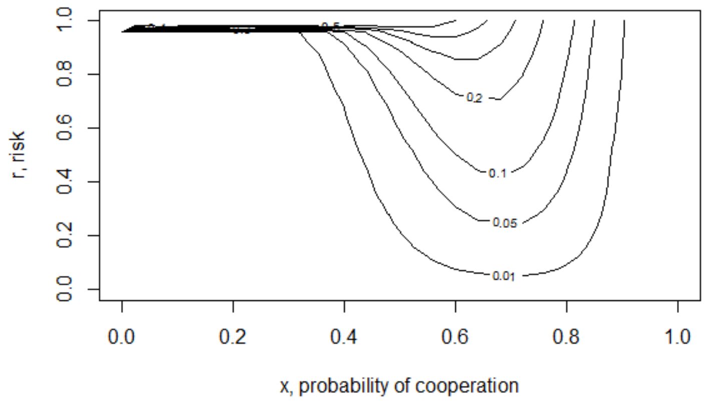

Figure 4.3.1: For $N = 20, M = 15$, mixed-strategy Nash equilibria given by the probability of cooperation x on the x axis for CRD games with risk r on y axis and cooperation cost c given as level sets. To read this graph, imagine a horizontal line for a fixed r intersecting the level set for a fixed c, and the value for x at that point gives the MSNE strategy.

Now that we have analyzed the CRD Nash equilibria, we investigate their *stability*. Stability is important in this work both

4.4 CRD ESS characterization

Prior work by Santos and Pacheco (2011) showed that the stable steady states of the CRD are the all-defect state and the mixed strategy x_R (the higher cooperative mixed Nash strategy). In this section I further show that these stable steady states are also evolutionary stable strategies (ESS), a stronger notion as discussed in 2. I characterize the ESS using the approach outlined in Bach et al. (2006) for N player threshold games.

We first want to determine the increase in expected payoff for a single agent playing C rather than D, or the *gain function*. Adopting similar notation to that in the EGT treatment in 2, we notate the payoff in the CRD of playing pure strategies C, D against k collaborators as $u(C, k), u(D, k)$, respectively. We first determine the additional expected payoff a single agent will obtain by switching from D to C if k other players cooperate; call this $\Delta(k)$.

$$\Delta(k) = u(C, k) - u(D, k) = b[(1 - c)p_c(x) - c(1 - r)(1 - p_d(x)) - p_d(x)]$$

Now we can define the gain function.

Definition 32 (Gain function for CRD). $g(x) = \sum_{k=0}^{N-1} \binom{N-1}{k} x^k (1 - x)^{N-1-k} \Delta(k)$

Note that $g(x)$ is a polynomial on interval $[0, 1]$.

We use the gain function to define the expected payoff from a single agent playing strategy y while the remaining $N-1$ players are playing strategy x, which as in Bach et al. (2006) we call W.

$$W(y,x)=\sum_{k=0}^{N-1}\binom{N-1}{k}x^k(1-x)^{N-1-k}u(D,k)+y\cdot g(x)$$

Since there are mixed Nash only when this payoff is independent of y Bach et al. (2006), $x\in(0,1)$ can only be a Nash equilibrium when $g(x)=0$ (which is if and only if $\Delta(k)=0$ for all k, consistent with our earlier NE analysis).

We now characterize the ESS of the game with respect to the gain function. I follow the theorem and proof in Bach et al. (2006), with only a difference in the referenced gain function [3].

Theorem 33 (Characterization of ESS of the CRD). *We characterize when the Nash equilibria of the CRD are also ESS.*

- *If $g(0)<0$, then $x=0$ is an ESS.*
- *If $g(1)<1$, then $x=1$ is an ESS.*
- *If $g(x)=0$ and $g'(x)<0$, then x is an ESS.*

Proof. Using the definition of W and the condition for a strategy being an ESS in Broom et al. (1997) for all $y\in[0,1]$, $y\neq x$, and ε smaller than some $\varepsilon(y)$

$$\sum_{k=0}^{N-1}\frac{\varepsilon^k}{k!}g^{(k)}(x)(y-x)^{k+1}<0$$

Since ε can be arbitrarily small, we need only concern ourselves with the first nonzero term of the Taylor expansion. This implies that $x=0$ is an ESS if $g'(0)<0$ and that $x=1$ is an ESS if $g'(1)>0$. Note also that if $g(0)>0$, then $y=1$ is the unique best response to $x=0$ (following a computation of W), so $x=0$ is not a NE. A similar argument holds for $x=1$ not being a NE if $g(1)<0$.

For the mixed strategies $x\in(0,1)$, from the inequality we see that if $g(x)\neq 0$ the the inequality cannot be satisfied for all y since the term $(y-x)^{k+1}$ can be both positive and negative. However, if for the odd integer k such that it is the lowest nonzero term $g^{(k)}(x)\neq 0$, then if $g^{(k)}(x)<0$ then the inequality always holds. Thus for $g(x)=0$ and $g'(x)<0$, x is an ESS. □

To apply this theorem, we inspect the graph of the gain function.

[3] Note also that Bach et al. (2006) identify further special cases of ESS for their general threshold game, but these special cases do not apply to the CRD due to the structure of the CRD gain function.

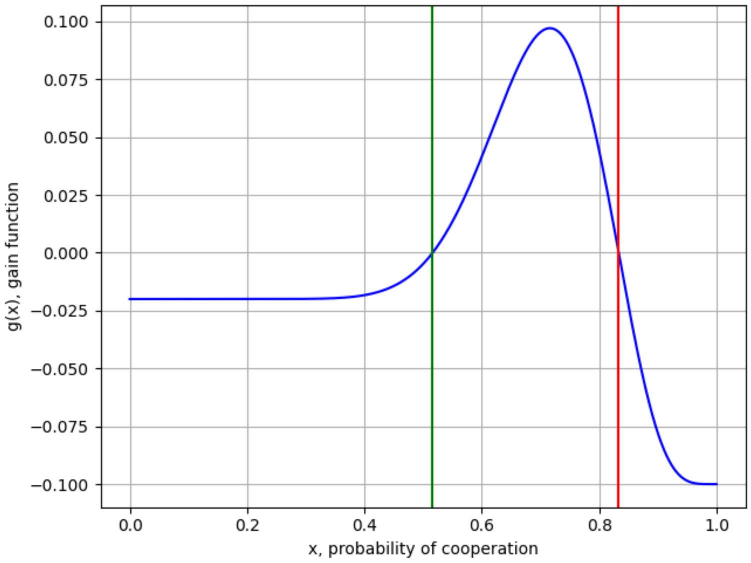

Figure 4.4.1: Plot of the gain function for $CRD(N = 20, M = 15, c = .1, r = .8)$. Green line gives x_L, red line gives x_R.

From the graph we verify that our MSNE calculated in the previous section, x_L, x_R, satisfy $g(x_L) = 0, g(x_R) = 0$ and see that x_R is an ESS since clearly $g'(x_R) < 0$. However, x_L is not an ESS because $g'(x_L) \not< 0$; this is is consistent with the finding in Santos and Pacheco (2011) that x_L is not a stable steady state since ESS are a subset of stable steady states 2.

Showing that the all defect equilibria and x_R are ESS strengthens the case for using these strategies as the base of my model in Chapter 5 and 6.

5 Model

This model introduces a new extension of the CRD covered in Chapter 4. The extensions aim to capture the role of a social planner intervening periodically in a population.

5.1 Motivation

Consider the following motivating setting: a government planner wants to prevent pollution from individual actors but is unable to effectively prevent it. However, the government is able to offer non-targeted interventions such as pollution prevention education or subsidizing technology with safety standards that reduce pollution. Furthermore, as is the case in many environmental settings, assume that the perceived risk among individual actors does not match the true risk posed by further pollution.

In this setting we introduce an extended CRD which allows such a social planner to intervene by changing the CRD's cost of cooperation parameter c and address a *low perceived risk* among individual agents in the CRD. The goal of the social planner is to reduce pollution actions, i.e defecting actions. In the CRD context, this means the objectives can be both to increase the level of cooperation played at CRD equilibrium by manipulating c, r, to avert evolutionary shifts towards the evolutionary stable strategy of all-defect, and to ensure that the CRD agents achieve the threshold.

Furthermore, the CRD as currently defined has one final conclusion – either the risky event occurred or it did not. If in the extreme the risky event represents something especially undesirable (perhaps "catastrophic climate change"), we may want periodic checkpoints before we flip the risk coin. To capture the potential lasting impact of a defect action (e.g. pollution staying in the environment), we introduce an intertemporal version of the CRD in which the CRD is played at each time step of an MDP with respect to the true (not perceived) risk.

The action designed so that it is the planner is able to reverse both the negative effects of a low perceived risk in the population and the worsening effects of increased cooperation cost as CRD games fail, reflecting the cascading effects of defection such as pollution.

The action we consider is an economic intervention to subsidize the cost of cooperation for the agents. The subsidy action is modeled as a percentage decrease in the cost of cooperation; we refer to the resulting cost as the *subsidized cost*.

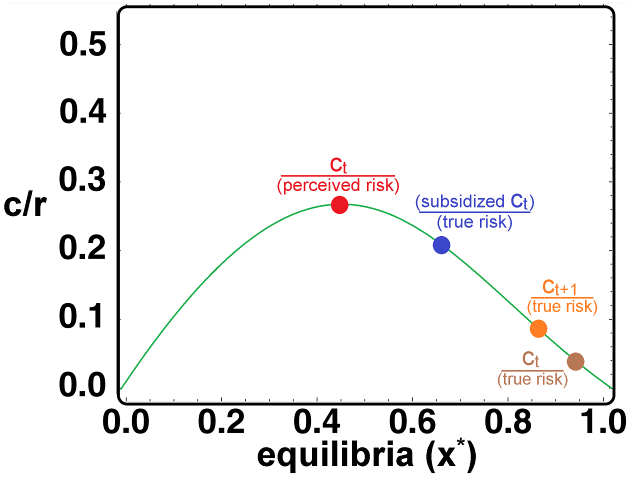

Figure 5.1.1: The curve pictured gives the MSNE x^* for some (c, r), plotted on the y-axis as the relation c/r. With no action from the planner, the CRD agents play an equilibria at the red dot with respect to the true cooperation cost and their perceived risk. This level of cooperation is much lower than that of the true equilibrium at the brown dot, which is with respect to the true cooperation cost and the true risk. If the planner subsidized the cost, the "correct equilibria" with respect to the true risk would be at the blue dot. The CRD agents with the subsidized cost would play at an equilibrium between the red and blue dot, which is higher than their default. The intertemporal aspect of the environment is represented by a cooperation cost which increases over time from time step t to $t + 1$, making it more difficult to sustain higher levels of cooperation – this is representative of higher cleanup costs in pollution, for example.

5.1.1 Formal model

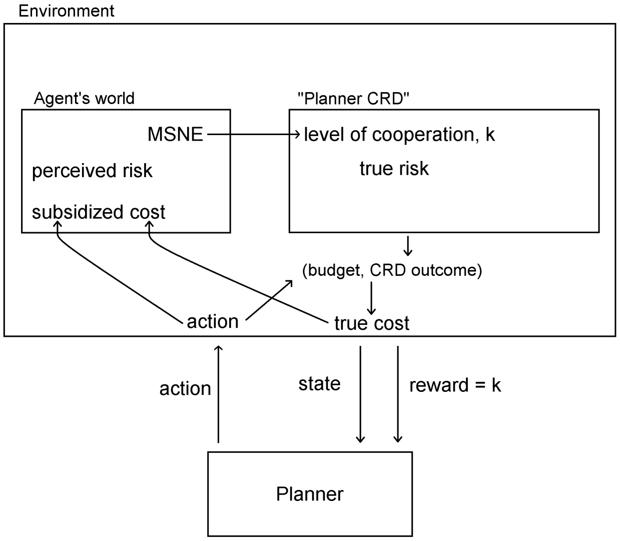

Figure 5.1.2: Diagram indicating the high-level relationship between moving parts in the environment, the factors that go into computing reward, and the effect of the planner's action.

The MDP induced by the above description is defined as follows. We specify only 3 actions for simplicity.

Hyperparameters: $\beta \in (0, 1)$, determines the rate at which cooperation increases / time step

- State:
  - Fixed parameters: $N \in \mathbb{Z}^+$, number of agents, $M > N$, threshold of cooperative actions, $r' \in (0, 1)$ perceived risk, $r \in (0, 1)$, true risk
  - $c \in (0, 1)$, cost of cooperation
  - $c' \in (0, 1)$, subsidized cost of cooperation
  - $p \in (0, 1)$, probability of cooperation in CRD equilibrium with respect to r', c'
  - $k = N \cdot p$
  - `safe=1` if $k > M$; else `safe=0` with probability r, `safe=1` otherwise.
- Actions: $\{a_0 = \text{No subsidy}, a_1 = \text{subsidy}, a_2 = \text{higher subsidy}\}$
- Transition: $(s^{(t)}, a) \mapsto s^{(t+1)}$
  - if `not safe`: $c \mapsto c \cdot \beta$
  - $c', a_i \mapsto c' \cdot (1 - i/10)$
- Reward: $R(s, a) = s.k$

6
Results

In this chapter, we provide 1) a brief overview of the implementation of the model given in Chapter 5, 2) a justification of the environment and learning parameters used to generate results, 3) training results for both unbudgeted and budgeted planners.

6.1 Model specification

In this section we review further details needed for the implementation of the model. Note that we use an OpenAI gym environment to facilitate implementation. First to clarify terminology, *step* refers to steps within one episode, and an *episode* refers to a run of a reset environment over steps, and an *epoch* refers to a run of a reset agent over episodes so that Q-values are updated (# of steps) · (# of episodes) times per epoch.

The RL planner is implemented using Q-learning, as reviewed in Chapter 3. The RL planner is given the entire state space along with the time step as its observation. The planners actions are as described in 5.1.1.

We review the additional specifications for the environment and the CRD agents. Given a cooperation cost, we compute the minimum and maximum risk such that a MSNE exists. We scale the minimum and maximum risk by the `challenge` parameter, increasing the minimum risk and decreasing the maximum risk, to yield the the perceived risk and the true risk respectively. This has the effect of making the planner's game achievable; if the perceived risk is exactly on the boundary, a slight increase in cooperation cost would push the CRD agents into an all-defect state, and the planner's actions would have to overcome completely ever increase in cooperation cost to succeed. Again reference Figure 5.1.1 for a visual representation.

We assume that agents in the inner population will play either of the two ESS: the stable MSNE with respect to their perceived risk and subsidized cooperation cost, or the all-defect NE. To compute this MSNE, we follow the analysis from 4 and implement a binary search for the stable MSNE.

6.1.1 Parameter selection

The results use the following parameters.

6.1.2 Environment parameters

There are two key environment parameters: the risk perception difference between the agent population and the true risk, which I call the "challenge" parameter, and the percentage β by which we increase the cooperation cost when the CRD outcome is failure. In the implementation, these values are set to $challenge = .1$, `beta`. The environment parameters were chosen carefully to allow a range of behavior and outcomes. I note three characteristics of the environment. First we choose a finite time step horizon, number of steps/episode $t = 15$ for ease of simulation and explainability.

- If no subsidy action is taken at any step, then before halfway through the end of the episode agents the cooperation cost has increased such that agents now play the all-defect Nash equilibrium.
- If the maximum subsidy action is taken at each step, the cost of cooperation does not increase since the stable MSNE brings the population above the cooperation threshold in expectation.
- A mixed policy can bring a population back from an all-defect state.

We represent these three properties graphically, plotting the agents' strategies over the course of an episode (15 time steps).

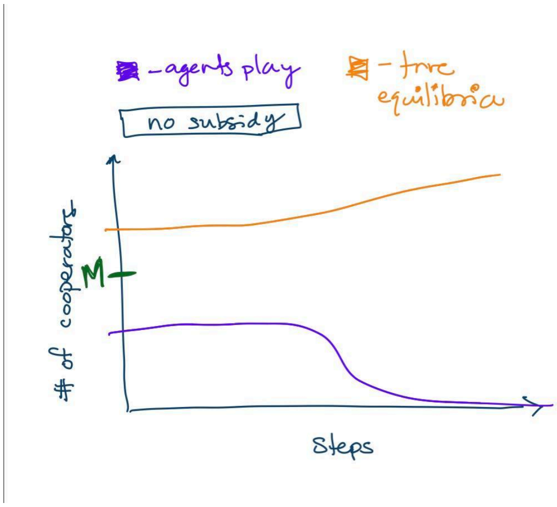

Figure 6.1.1: Without any subsidies, agents will not cooperate above the threshold, and as the CRD fails the cooperation cost will rise in the environment, forcing the agents' equilibrium to the all-defect.

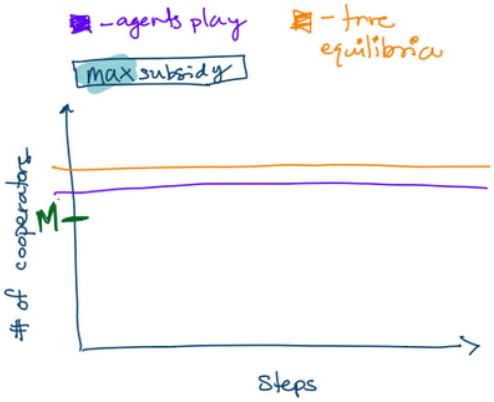

Figure 6.1.2: With max subsidies, agents cooperate above the threshold, and since the CRD never fails the cooperation cost does not increase (and thus does not change the true equilibria), allowing continued cooperation.

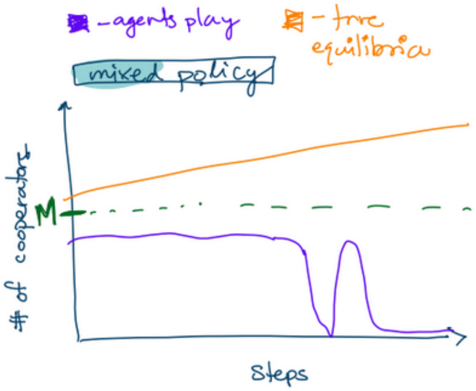

Figure 6.1.3: With a mixed RL planner policy, it is possible for the planner's action to push the agents' equilibrium strategy from all-defect back to the stable MSNE. Note that this policy is not ultimately learned (indeed, it is not optimal for the agents to be in the all-defect state) but is one that could be explored in the learning process.

6.1.3 Epsilon-Greedy Q-learning parameters

To inform these parameter choices, I referenced Calvano et al. (2020) and experimented with different parameters.

Exploration/exploitation tradeoff

Since the environment is especially sensitive to ill-timed low subsidy actions, exploration can often result in significantly lower reward. To address this, I apply a time-declining exploration rate where for episode i,

$$\varepsilon_t = e^{-\beta i}$$

for some $\beta > 0$, where larger β results in faster diminishing of exploration.

Learning rate

The learning rate α is directly tied to the Q-learning update function; the larger α is, the more that Q-values are sensitive to changes in reward and expected value. At the beginning of training, a high enough α is needed to learn from new actions/states; at the same time, a large α can disrupt learning when there is frequent exploration since the Q-learning algorithm would too easily "forget" what it learned before. Due to the complex state space, a relatively high α is chosen but is also subject to a learning rate decay to balance these considerations.

Discount factor

The discount factor γ weights future rewards; since we are especially interested in the reward in the last time step (whether the population was prepared for the "true" risky event), we set $\gamma = .9$ to still allow convergence but better capture this cumulative reward over steps.

6.2 Planner results

6.2.1 Unbudgeted planner

The unbudgeted RL planner learns a strategy which approaches and often reaches optimal reward (≈ 229). Note that this optimal reward is only achievable if the agents do not play the all-defect equilibria. However, due to the environment dynamics, there are multiple policies that may achieve optimal/near-optimal reward.

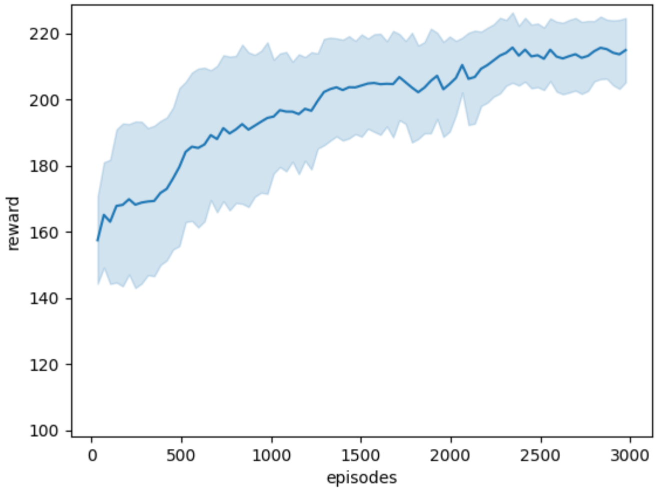

Figure 6.2.1: It learns! This plot shows cumulative reward per episode over episodes averaged over 10 epochs, with standard error[1] shaded. Learning parameters $\alpha = .7$ with annealing, $\varepsilon = .05$ with annealing, 3000 episodes. No budget. The range on the y-axis is the range of rewards possible given the environment parameters.

We see that a planner can learn to intervene effectively in the intertemporal CRD to both increase cooperation and avert the risky event occurring in the last time step's CRD.

A sample learned policy is $\pi = [2, 2, 2, 1, 1, 2, 1, 2, 2, 0, 2, 2, 2, 2, 0]$ over the 15 time steps. We see that intervening with a higher subsidy is preferable, and perhaps necessary to stay on track following an instance of no subsidy. Since the RL planner's actions are not constrained by a budget, it almost always acts with the maximum subsidy. We note that the learned CRD agent policies are like those in 6.1.2.

6.2.2 Budgeted planner

The budget is implemented coarsely, with much room available for further analysis of the effect of the budget. The planner is limited by some budget parameter B: the budget is decremented proportional to the action the planner took in the previous time step, and when no budget remains the agent's chosen action has no effect on the environment. We plot our results for planners constrained by different degrees of budget.

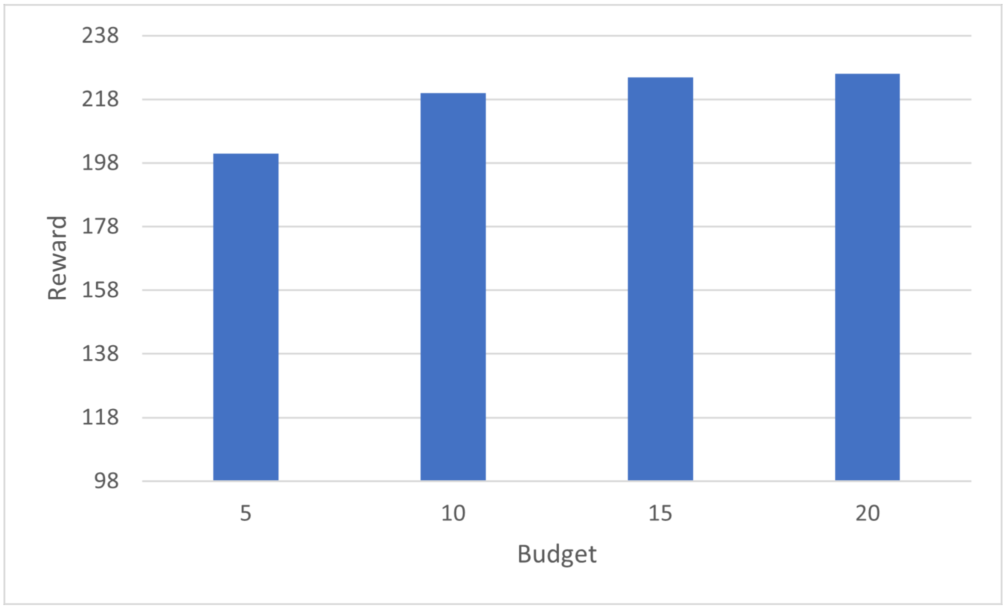

Figure 6.2.2: This bar chart plots the max. achievable reward learned by a planner with budget B. Note that the budget allowing for maximum subsidizing action each time step is $B = 30$, and the minimum budget such that the agent can still act is $B = 1$, and again the range in reward is from 98 to 230.

We find that the planner is still able to learn high-reward policies in this challenging environment, even for low budgets which only allow a few subsidizing actions.

7
Conclusion

This thesis provided 1) cohesive exposition on notions of equilibria and stability in game theory and evolutionary game theory, 2) equilibria and stability analysis of the collective risk dilemma, proving the strength of stability of the stable states identified in prior work, and 3) extended the CRD model to capture changes over time in the parameters of the CRD, and 4) successfully used Q-learning to implement a social planner which is able to prompt agents with a low perceived risk, who would otherwise play the MSNE with an insufficient level of cooperation, to overcome the CRD threshold and prevent CRD failure.

There are many avenues for future work. In this model, we allow the CRD agents to play their evolutionary stable strategies in response to planner interventions. It would be illuminating to use replicator dynamics or multi-agent reinforcement learning (MARL) to model the CRD agents to see 1) how long convergence (and if convergence occurs in the MARL case) to evolutionary stable strategies and 2) how effective planner interventions are when CRD agents don't immediately play their MSNE/evolutionary stable strategies – indeed, it could be easier for the planner since the effect of an increase in cooperation cost in the environment would not immediately reduce levels of cooperation in the population. This extension would better model the real world in which human's strategies may be resistant to change or are more discounted. Another extension could leverage the population model of the CRD, i.e. playing the CRD with N players sampled from some larger population. Using this version of the CRD, we could analyze communication among agents in a population and capture any spatial aspects of a real world environment (e.g. like in the ASM setting). Ultimately, further study using this extended CRD model will advance our understanding of how to address social dilemmas in the real world by capturing *N* player game dynamics under a social planner.

8 Appendix

8.0.1 Appendix to Chapter 4

Algebraic derivation for the change in expected payoff from C to D

To be completely certain of our algebraic simplification of the characterization of the MSNE, we provide a line by line derivation.

$$\begin{aligned}
\pi_i(C) - \pi_i(D) &= b(1-c)\left(p_c + (1-r)(1-p_c)\right) - b\left(p_d + (1-r)(1-p_d)\right) \\
&= b\left(p_c + (1-r)(1-p_c) - cp_c - c(1-r)(1-p_c) - p_d - (1-r)(1-p_d)\right) \\
&= b\left(p_c + 1 - r - p_c + rp_c - cp_c - c(1 - r - p_c + rp_c) - p_d - (1 - r - p_d + rp_d)\right) \\
&= b\left(rp_c - c + cr - crp_c - rp_d\right) \\
&= b\left(r(p_c - p_d) - c + cr(1-p_c)\right) \\
&= b\left(r \cdot P(k = M-1) - c(1 - r \cdot P(k \leq M-2)\right)
\end{aligned}$$

with the last line as a result of the equalities

$$1 - p_c = 1 - (1 - P(k \leq M-2)) = P(k \leq M-2)$$

$$p_c - p_d = (1 - P(k \leq M-2)) - (1 - P(k \leq M-1)) = P(k \leq M-1) - P(k \leq M-2) = P(k = M-1)$$

References

L. Bach, T. Helvik, and F. Christiansen. The evolution of N-player cooperation—threshold games and ESS bifurcations. *Journal of Theoretical Biology*, 238(2):426–434, 2006.

I. M. Bomze. Non-cooperative two-person games in biology: A classification. *International Journal of Game Theory*, 15(1):31–57, 1986.

M. Broom, C. Cannings, and G. T. Vickers. Multi-player matrix games. *Bulletin of Mathematical Biology*, 59(5):931–952, 1997.

E. Calvano, G. Calzolari, V. Denicolo, and S. Pastorello. Artificial Intelligence, Algorithmic Pricing, and Collusion. *The American Economic Review*, 110(10):3267–3297, 2020.

E. Fernández Domingos, J. Grujić, J. Burguillo, F. Santos, and T. Lenaerts. Modeling behavioral experiments on uncertainty and cooperation with population-based reinforcement learning. *Simulation Modelling Practice and Theory*, 109:102299, 03 2021. doi: 10.1016/j.simpat.2021.102299.

D. Fudenberg. *Game Theory*. MIT Press, Cambridge, Mass., 1991.

D. Fudenberg. *The Theory of Learning in Games*. MIT Press series on economic learning and social evolution ; 2. MIT Press, Cambridge, Mass., 1998.

S. Huang. Bi-level multi-agent reinforcement learning for intervening in intertemporal social dilemmas, 2020.

E. Hughes, J. Z. Leibo, M. G. Phillips, K. Tuyls, E. A. Duéñez-Guzmán, A. G. Castañeda, I. Dunning, T. Zhu, K. R. McKee, R. Koster, H. Roff, and T. Graepel. Inequity aversion improves cooperation in intertemporal social dilemmas. 2018.

T. Jaakkola, M. I. Jordan, and S. P. Singh. *On the Convergence of Stochastic Iterative Dynamic Programming Algorithms*. 1993.

F. S. Melo. Convergence of Q-learning: A simple proof, 2021.

M. Milinski, R. D. Sommerfeld, H.-J. Krambeck, F. A. Reed, and J. Marotzke. The Collective-Risk Social Dilemma and the Prevention of Simulated Dangerous Climate Change. *Proceedings of the National Academy of Sciences - PNAS*, 105(7):2291–2294, 2008.

J. F. Nash. Equilibrium Points in n-Person Games. *Proceedings of the National Academy of Sciences - PNAS*, 36(1):48–49, 1950.

E. Ostrom. Governing the commons: The evolution of instituitions for collective action, 1990.

E. Ostrom and J. Walker. *Trust and Reciprocity: Interdisciplinary Lessons for Experimental Research*. Russell Sage Foundation Series on Trust. Russell Sage Foundation, New York, 2003.

D. Parkes and S. Seuken. *Algorithmic Economics: A Design Approach*. Cambridge University Press, 2022.

F. C. Santos and J. M. Pacheco. Risk of collective failure provides an escape from the tragedy of the commons. *Proceedings of the National Academy of Sciences - PNAS*, 108(26):10421–10425, 2011.

T. C. Schelling. Hockey Helmets, Concealed Weapons, and Daylight Saving: A Study of Binary Choices with Externalities. *The Journal of Conflict Resolution*, 17(3):381–428, 1973.

R. S. Sutton. *Reinforcement Learning: An Introduction*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts ; London, England, second edition. edition, 2018.

J. Tsitsiklis. Asynchronous Stochastic-Approximation and Q-Learning. *Machine Learning*, 16(3): 185–202, 1994.

C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK, May 1989. URL http://www.cs.rhul.ac.uk/~chrisw/new_thesis.pdf.