Games with vector payoffs : a dynamic programming approach

Vijay Kamble

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2015-201
http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-201.html

September 22, 2015

Games with Vector Payoffs : A Dynamic Programming Approach

by

Vijay Sukumar Kamble

A dissertation submitted in partial satisfaction

of the requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineeing and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jean Walrand, Chair
Professor Anant Sahai
Professor Christos Papadimitriou
Professor Shachar Kariv

Fall 2015

Games with Vector Payoffs : A Dynamic Programming Approach

by

Vijay Sukumar Kamble

Abstract

Games with Vector Payoffs : A Dynamic Programming Approach

by

Vijay Sukumar Kamble

Doctor of Philosophy in Engineering – Electrical Engineeing and Computer Sciences

University of California, Berkeley

Professor Jean Walrand, Chair

In several decision-making scenarios in adversarial environments, a decision-maker cares about multiple objectives at the same time. For example, in certain defense operations, an agent might be interested in simultaneously defending multiple targets from an enemy. In a repeated game against an unknown opponent, a player wants to minimize 'regret', i.e., to try to choose a strategy that performs well relative to each strategy in some given class of strategies in hindsight. In dynamic asymmetric information games where a player lacks some information that other players have, a typical goal is to choose a strategy that gives appropriate worst-case guarantees simultaneously on all possibilities. Many of these scenarios can be modeled as a vector-valued sequential game between the agent and an adversary. This thesis is concerned with characterizing and efficiently computing the optimal worst-case guarantees that an agent can achieve on the losses in such games.

The main contribution of this work is to show that for large classes of sequential games, these optimal guarantees can be characterized as the fixed point of a dynamic programming operator defined on the space of extremal (either maximal or minimal) elements of subsets of some partially ordered topological space. We first present this result in detail for the model of discounted repeated games with vector payoffs and then extend it to stochastic games with multiple states, and finally to reachability games (which model several types of pursuit-evasion games that arise in defense operations). For each of these models, we prove several structural properties of the set of these optimal guarantees and the corresponding optimal strategies. This approach opens up the possibility of using many well-known dynamic programming based methods and algorithms for approximating these guarantees and computing approximately optimal strategies. One such method based on approximate value-iteration is presented for the case of repeated games.

This approach results in the first characterization of the minmax optimal regret and the corresponding optimal strategy for expected regret minimization in repeated games with discounted losses. Further, it results in the first known procedure for efficiently computing an approximately optimal strategy for the uninformed player in Aumann and Maschler's celebrated model of zero-sum discounted repeated games with incomplete information on one side.

To my family

Contents

List of Figures

List of Tables

Acknowledgements

There has not been a single instance in the last five years when entering the Berkeley community to start my day was not a singular source of pure joy and inspiration, no matter how hard the times were, for it is always immediately evident all around that there is something clearly bigger than each of us flourishing majestically in this place. I cannot overstate the impact that this environment has had on me as a person and as a potential researcher, and I am immensely grateful to have been a part of it.

I feel incredibly lucky to have Jean Walrand as my advisor. A refined sense of understanding of what it means to be a researcher, of what is important and what is not (in life in general), is so clearly evident through his actions and conversations, and with results so compelling, that one cannot help but want to follow suit. The clarity of his thinking never fails to astound me. There have been numerous instances where I would just spread out a chaotic mass of ill-formed ideas in front of him and he would very cooly cut through it to carve out simple objects of interest, if there are any. This thesis is a result of several such interactions. I cannot express how deeply grateful I am for his guidance, care and patience over all these years.

Another micro-community that I have had a great pleasure of being a part of is the WiFo group (now called BLISS, but I am not quite used to calling it that yet). I have said this multiple times to people before, that the only thing I regret in my time as a PhD student is that I did not become a part of WiFo earlier (this happened only recently in early 2014). It really elevated the Berkeley experience to a whole another level. I am immensely grateful for the friendship and support that I have enjoyed from all its members, especially Kangwook Lee, Rashmi K.V., Nihar Shah, Gireeja Ranade, Po Ling Loh, Varun Jog, Giulia Fanti, Ramtin Pedarsani, Kate Harrison, Vasuki Narasimha Swamy, Sudeep Kamath, Venkatesan Ekambaram, Sameer Pawar, Naveen Goela, Fanny Yang, Reza Abbasi Asl, Vidya Muthukumar, Ashwin Pananjady, Payam Delgosha, Orhan Ocal and Dong Yin.

I would like to thank Anant Sahai, who has always supported me in my various endeavors in his capacity as my qualifying exam committee member and my dissertation committee member, and also for giving me an opportunity to collaborate on a very exciting research project that turned out to be a great learning experience for me. His help and guidance has been invaluable.

I sincerely thank the members of the Net-econ group: Venkat Anantharam, John Mussacchio, Abhay Parekh, Shyam Parekh and Galina Schwartz, for always providing me with timely feedback on my research. I would especially like to thank Abhay Parekh for the wonderful experience I had working as a GSI with him for EE126.

I would like to thank Patrick Loiseau, who hosted me at Eurecom in the French Riviera for a beautiful summer month in 2015, while collaborating on some key applications of my thesis work. I would also like to thank Nadia Fawaz and Fernando Silveira for hosting me for a fun summer internship at the Technicolor Lab in Palo Alto in 2012. Finally, I owe a huge debt of gratitude to Eitan Altman, who introduced me to research as an undergraduate student.

I am extremely thankful to have been surrounded by a wonderful group of friends here in Berkeley (apart from the WiFo folks): my roommate Arka Bhattacharya, my (slightly

younger) academic sibling Stephan Adams, my local guardians Raj Shekhar Singh and Momo Zheng, Aamod Shanker, Debanjan Bhowmik, Varun and Vivek Mishra, just to name a few. Also a few very close friends from college: Anuj Gupta, Pavan Nithin and Mayank Kedia, curiously managed to make their way back into my life at a time when it felt like that was exactly what the doctor would have ordered.

Finally, I would like to thank my family: my wife Madhura, my sister Shweta, and my parents Shubhangi and Sukumar, whose unwavering love and support are the single greatest source of my strength. I love them beyond words can express. This dissertation is dedicated to them.

Chapter 1

Introduction

In several decision-making scenarios in adversarial environments, an agent cares about multiple objectives at the same time. Typically in these cases her actions can have different implications for the different objectives, and she wants to take some action that performs appropriately well on all the objectives simultaneously. Such scenarios can be effectively modeled as a game with vector-valued payoffs. Following are a few examples.

- **Pursuit-evasion games of target defense:** Consider the following pursuit-evasion game that arises in defense operations. A defender wants to protect 3 different targets, A, B and C from being destroyed by an enemy in a field. To achieve this, the defender tries to apprehend the enemy (say by coming within some striking distance) before the enemy is able to reach the targets. Since the targets are in different locations, the defender needs to keep track of the effects of his movements on the vulnerability of the different targets, which may be quite different: a move that keeps the enemy away from one target may make the other target more vulnerable. One can model this situation as a vector-valued dynamic game with three components, one for each target. If the enemy reaches a target i, the defender faces a loss of l_i on the corresponding component.

- **Games of incomplete information:** In many situations, an agent may not have sufficient information about the game that she is involved in. For instance, all that is known is that the underlying game is one of a finite set of possible games. Such games of incomplete information can be modeled as a vector-valued game, where each component of the vector represents the loss corresponding to a particular possibility. Consider the following example. Alice and Bob decide to play the following simple game that we will call 'Bluff'. The game consists of two stages. In the first stage, they simultaneously place a bet, which could be either 20 cents (low) or \$1 (high). If they both bet low, they keep their money and the game is over. If one bets high and the other bets low, the low player loses his/her bet to the high player and the game is over. If both bet high, then the game proceeds to stage two where they play a game

of chess. The winner of the game takes all the money and the game is over. Now both Bob and Alice know that Alice is an average chess player, but only Bob knows his expertise, which could be 'Expert' or 'Poor'. If Bob is an expert then he definitely beats Alice, where as if he is a poor player then Alice definitely beats him. Thus in the event that Bob is an expert, Alice always wants to bet low, whereas if he is a poor player then she wants to bet high. One can model this situation as the vector-valued game shown in Figure 1.1 with the two components corresponding to Alice's loss in the two different possibilities. Alice's goal is then to minimize her losses simultaneously

		Bob: H	Bob: L
Alice	H	(1, -1)	(-0.2, -0.2)
	L	(0.2, 0.2)	(0, 0)

Figure 1.1. A vector valued zero-sum game.

on the two possibilities.

- **Regret minimization in repeated decision-making:** One of the best studied problems in online decision-making is that of regret minimization in repeated games (see e.g., [9] for a survey). Imagine that Alice can choose one of two paths to go to work each day. Each path may be congested, leading to a loss of 1 for Alice, or it may be clear, leading to no loss. Each day, without knowing the state of the paths, Alice chooses some path, possibly randomly. After having taken this path, she learns the congestion levels of the two paths in hindsight. This model can be represented by the matrix game in Table 1.1. The rows correspond to the choice made by Alice and the columns correspond to the four different possibilities: either both paths are congested, or both are clear, or one is congested and the other one is clear. Alice would like to minimize her expected worst-case 'regret', defined as the difference between her actual expected loss and the loss incurred by the better of the two paths in the long run.

  Now consider the vector-valued game in Table 1.2. For any entry, the first component is the additional loss incurred in a single stage relative to the loss incurred by choosing the first path, and the second component is the additional loss incurred relative to the loss incurred by choosing the second path, called the single-stage regret with respect to

Table 1.1. Possible loss scenarios

Path 1	1	0	1	0
Path 2	0	1	1	0

Table 1.2. Single-stage regret w.r.t. Path 1 and 2 (in that order)

Path 1	(0,1)	(0,-1)	(0,0)	(0,0)
Path 2	(-1,0)	(1,0)	(0,0)	(0,0)

Path 1 and Path 2 respectively. The goal of minimizing regret can be formally thought of as simultaneously minimizing the cumulative loss on the different components in a repeated play of this vector-valued game.

The goal of the decision-maker in these games is to choose actions that perform well on the different components despite the adversarial environment. Playing a strategy that guarantees a low loss on one component could at the same time make the potential loss high on the other components. Thus the problem for the decision-maker is to characterize the optimal tradeoff between the guarantees on the different components and find strategies that achieve the different points on this tradeoff. This thesis is concerned with characterizing and computing this tradeoff and strategies that achieve it for different classes of dynamic games.

In this chapter, we will begin by reviewing a few classical results on zero-sum games with scalar payoffs. We will then introduce vector-valued games and define the problem of finding the minimal simultaneous guarantees on the losses in these games. We will give a few simple examples to illustrate the use of a dynamic programming based approach to characterize these minimal guarantees in dynamic games. We will finally give a brief overview of the organization of the rest of the thesis.

1.1 Scalar zero-sum games: review of results

In its simplest form, a zero-sum game consists of two players, call them Alice (or the player/agent/subject) and Bob (the adversary), who are endowed with a finite set of actions A and B respectively, and for each pair of actions (a, b), an amount of utility $r(a, b)$ is transferred from Alice to Bob. Figure 1.2 shows a matrix representation of an example of a zero-sum game.

There are two quantities of interest: the minmax and maxmin. The minmax is the smallest upper bound that Alice can guarantee on her loss in the game, formally defined as:

$$\bar{v} = \min_{a \in A} \max_{b \in B} r(a, b).$$

		Bob	
		1	2
Alice	1	2	1
	2	0	3

Figure 1.2. A zero-sum game.

The corresponding action that achieves this guarantee is called the minmax strategy. The maxmin is the highest lower bound that Bob can guarantee on his gain, defined as:

$$\underline{v} = \max_{b \in B} \min_{a \in A} r(a, b).$$

The corresponding Bob's action that achieves this guarantee is called the maxmin strategy. The minmax of the game defined above is 2 and is achieved by Alice's action 1, and the maxmin is 1, achieved by Bob's action 2. Observe that the minmax is greater than the maxmin, a fact that is in general true. Indeed, if a^* and b^* are the minmax and maxmin actions respectively, then

$$\bar{v} = \max_{b \in B} r(a^*, b) \geq r(a^*, b^*) \geq \min_{a \in A} r(a, b^*) = \underline{v}.$$

One can consider richer strategy spaces, namely those that result from randomizations over actions. Let $\Delta(A)$ and $\Delta(B)$ denote the simplices of all probability distributions on the actions of Alice and Bob respectively. These strategies are called mixed strategies. The minmax and maxmin can be defined analogously as:

$$\overline{V} = \min_{a \in \Delta(A)} \max_{b \in \Delta(B)} E[r(a, b)]$$

and

$$\underline{V} = \max_{b \in \Delta(B)} \min_{a \in \Delta(A)} E[r(a, b)],$$

where the expectation is over the randomness in the choice of these actions. Again $\overline{V} \geq \underline{V}$. But the celebrated minmax theorem by Von-Neumann states that in this case, the opposite inequality holds as well and hence $\overline{V} = \underline{V}$. This quantity, denoted by V, is then called the value of the game. For the game defined above, one can compute that the both the maxmin and the minmax are $\frac{3}{2}$. The optimal strategy for Alice is to play 1 with probability $\frac{3}{4}$ and

2 with probability $\frac{1}{4}$. Bob's optimal strategy is to play both actions with equal probability. Playing these two strategies comprises an 'equilibrium': one is a best response against the other. Further any such equilibrium gives the same loss (gain) V to Alice (Bob).

The minmax theorem can be stated in a few different ways, nevertheless conveying the same idea. First, in the form that we just presented, it says that the lowest upper bound on the losses that Alice can guarantee is the same as the highest lower bound that Bob can guarantee on his gains. This fact that there is no 'gap' between the maxmin and minmax can also be stated in the following equivalent way. For any $r \in \mathbb{R}$, either Alice has a strategy that guarantees that her loss is not greater than r, or Bob has a strategy that guarantees that his gain is at least r. This is clearly true if r is greater than minmax or if r is smaller than maxmin. But if

$$\underline{V} < r < \overline{V},$$

then for such an r, neither Alice has a strategy that can guarantee that her loss is no more than r, nor Bob has a strategy that can guarantee that his gain is at least r. The fact that the minmax theorem holds implies that there can be no such r, and hence the two statements are equivalent.

Another way of conveying the above notion is the following. Suppose that the payoffs of the matrix game lie in the interval $[m, M]$. Then for Alice, for any sets of the form $[m, r]$, either she has a strategy that ensures that the expected payoff of the game is in the given set, or Bob has a strategy that ensures that the expected payoff is outside that set, in which case we say that Bob can exclude this set. Similarly, for Bob, for any sets of the form $[r, M]$, either he has a strategy that ensures that the expected payoff of the game is in the given set, or Alice has a strategy that ensures that the expected payoff is outside that set. We will call sets $[m, r]$, downward closed (since for such a set S, if $x \in S$ and if $y \in [m, M]$ is such that $y \leq x$, then $y \in S$ as well) and the sets $[r, M]$, upward closed.

Note that although this distinction between the types of sets that Alice can guarantee (or Bob can exclude) and Bob can guarantee (or Alice can exclude) is practically well justified: Alice being the minimizer, is only interested in upper bounds on her losses, while Bob being the maximizer is only interested in lower bounds on his gains, but mathematically, the distinction between these two types of sets in artificial if one thinks about characterizing the extent to which Alice and Bob can control the expected payoffs of the game. Indeed, one can switch roles to have Alice be the payoff maximizer and Bob be the minimizer. The minmax theorem will give the value of this game, which will typically be different from the value of the original game (although in the game defined in Figure 1.2, the value of the game with the roles reversed is the same). This implies an analogous result for Alice and Bob, but with the types of the sets they can achieve reversed. Thus the minmax theorem is equivalent to saying that if the payoffs of the matrix game lie in the interval $[m, M]$, then for any sets of the form $[r, M]$ or $[m, r]$, for any player, either he/she has a strategy that ensures that the expected payoff of the game is in the given set, or other player has a strategy that ensures that the expected payoff is outside that set. Note that this result does not hold for arbitrary sets, even if they are closed and convex, i.e., closed intervals.

1.2 Simultaneous guarantees in vector-valued games

In the simplest model of a game with vector payoffs, given finite action spaces A and B for Alice and Bob respectively, the choice of a pair of actions a and b result in the transfer of a utility vector $\mathbf{r}(a,b)$ with K components from Alice to Bob. The game of 'Bluff' in Figure 1.1 is an example. We will now define the analogue of minmax for this vector-valued version of a zero-sum game (the maxmin is analogous). Alice would like to choose a strategy that would give robust guarantees on her loss irrespective of Bob's behavior. But since the losses are vector-valued, she is interested in ensuring *simultaneous* guarantees on her losses on different components. To be more precise, for a choice of distribution $\bar{\alpha} \in \Delta(A)$, consider the vector:

$$\mathbf{v}(\bar{\alpha}) = \left(\max_{b \in B} \sum_{a \in A} \alpha_a r_k(a,b) \right)_{k=1,\cdots,K}.$$

By this choice of $\bar{\alpha} \in \Delta(A)$, Alice guarantees that her expected loss on component k is no more than $v_k(\bar{\alpha})$ for any k, irrespective of what Bob does. We then say that $\bar{\alpha}$ achieves the simultaneous guarantee vector $\mathbf{v}(\bar{\alpha})$. Alice would then like to determine the minimal such simultaneous guarantees that she can achieve. That is, she would like to determine the *Lower Pareto frontier* of the set of different guarantees that she can achieve using different mixed strategies, which is the set $\mathcal{V}^* = \Lambda(\{\mathbf{v}(\bar{\alpha}) : \bar{\alpha} \in \Delta(A)\})$, where for a set $\mathcal{U}$,

$$\Lambda(\mathcal{U}) \triangleq \{\bar{x} \in \mathcal{U} : \forall\, \bar{x}' \in \mathcal{U} \setminus \{\bar{x}\},\, \exists\, k \text{ s.t. } x_k < x'_k\}.$$

A geometric illustration of this operation for the game defined in Figure 1.1 is given in Figure 1.3. The left figure illustrates the inner maximization by Bob for two different fixed choices $\bar{\alpha}_1$ and $\bar{\alpha}_2$ of Alice. By varying $\bar{\alpha}$, the resulting set of guarantees is given by the union of line segment joining $(0.2, 0.2)$ and $(0.36, -0.04)$ and the segment joining $(0.36, 0.04)$ and $(1, -0.2)$. The Pareto frontier of this set is all the points on the segment, as shown in the figure on the right. This frontier is achieved by choosing $\bar{\alpha}$ to be all the different points on the simplex.

In general, Alice may be interested in ensuring simultaneous guarantees not only on the different components, but on different linear combinations of components. Let $(\gamma_1, \gamma_2, \cdots, \gamma_L)$ be a set of real valued (row) vectors with K components. Then in general Alice is interested in computing:

$$\mathcal{V}^* = \Lambda\left(\left\{ \left(\max_{b \in B} \sum_{a \in A} \alpha_a \gamma_l \mathbf{r}(a,b)^T \right)_{l=1,\cdots,L} : \bar{\alpha} \in \Delta(A) \right\} \right).$$

If $L = 1$, this is the case where the situation reduces to a scalar-valued game. The notion of simultaneous guarantees on linear combinations of the different components of a vector valued game allows us to answer whether Alice has a strategy that the expected payoffs on the different components lie in a given convex polyhedron (possibly infinite). To do so, suppose the convex polyhedron is of the form

$$\mathbf{A}x^T \leq \mathbf{b},$$

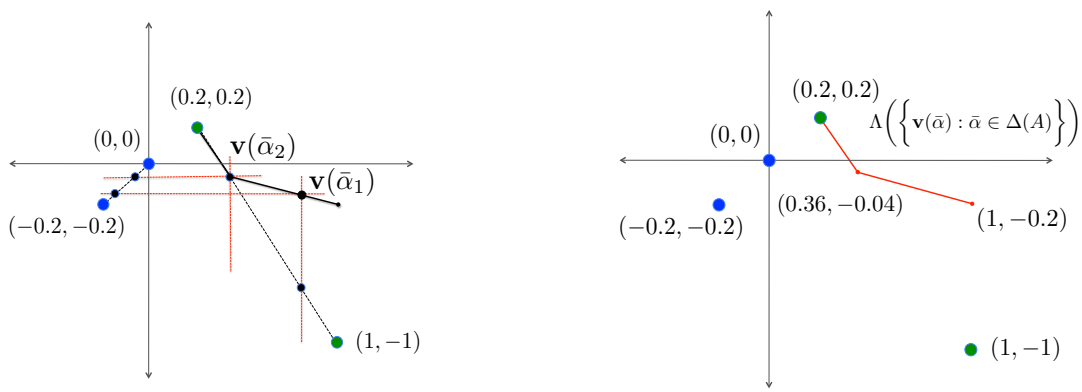

Figure 1.3. Computing the set of simultaneous guarantees (left) and minimal simultaneous guarantees (right)

comprising of L inequalities of the form $A_l x^T \leq b_l$, then if one computes the set

$$\mathcal{V}^* = \Lambda\left(\left\{\left(\max_{b\in B}\sum_{a\in A}\alpha_a \mathbf{A}_l \mathbf{r}(a,b)^T\right)_{l=1,\cdots,L} : \bar{\alpha} \in \Delta(A)\right\}\right),$$

then one can simply verify whether there is a $\bar{u} \in \mathcal{V}^*$ such that $\bar{u} \preceq \mathbf{b}$. If so, then the polyhedron is attainable, otherwise it is not. Since any convex set can be approximated to an arbitrary precision by a convex polyhedron, one can approximately answer the question of attainability for this convex set in a similar way.

1.3 Repeated vector-valued games

A classical model in games is where a pair of players play a particular matrix game repeatedly, while getting the matrix payoffs corresponding to the actions chosen in each stage. When the game is repeated finitely many times, then the payoff of this game is typically defined to be simply the average of the stage payoffs. When the game is played infinitely often, one then needs to define the total payoff of the game. Two payoffs typically considered are the expected discounted payoff and the expected limiting average payoff. For a sequence of actions $(a_1, b_1, a_2, b_2, \cdots)$, for a discount factor $\beta \in (0,1)$, the total discounted payoff is defined as:

$$(1-\beta)\sum_{t=1}^{\infty}\beta^t r(a_t, b_t). \tag{1.1}$$

		Bob 1	Bob 2
Alice	1	(0, 1)	(0, 0)
	2	(0, 0)	(1, 0)

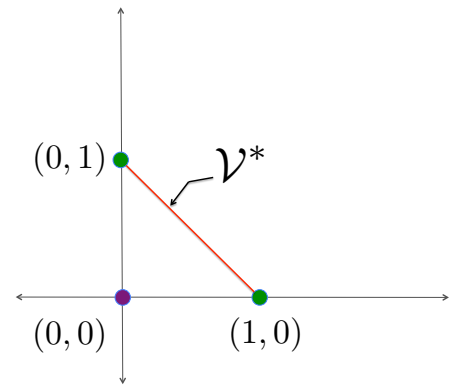

Figure 1.4. A vector valued zero-sum game (left) and the set of minimal one-stage simultaneous guarantees (right).

Note that if the payoffs $r(a, b)$ are bounded, then the quantity $\sum_{t=1}^{\infty} \beta^t r(a_t, b_t)$ is well-defined and it is bounded as well. The limiting average payoff is defined as:

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r(a_t, b_t). \tag{1.2}$$

An adaptive randomized strategy (also called behavioral strategy) for each player prescribes a mapping from the history of observed actions till time t, i.e., $\mathcal{H}_t = (a_1, b_1, \cdots, a_t, b_t)$ to a randomization over their action set. We then ask the standard questions: what are the minmax and maxmin values of the game and are they equal? When the game is scalar zero-sum, the answer is simple. The minmax and the maxmin strategies are exactly playing the minmax and maxmin strategies respectively of the single-shot game in every stage. The value of the finite average, the discounted and the limiting average game is V, the value of the one shot game.

But what about repeated vector-valued games? Let us look at a simple example where the game in Figure 1.4 is repeated twice. Computing such multi-stage optimal guarantees is going to be a recurring theme throughout this thesis and hence we explain it in some detail. Clearly, since the second stage is the last, the set of optimal guarantees that Alice can achieve after the first stage is the set $\mathcal{V}^*$ shown in Figure 1.4 on the right. Now a strategy for Alice consists of choosing some $\bar{\alpha} \in \Delta(A)$ at stage 1 and then, choosing a point $R(b)$ in the set $\mathcal{V}^*$ in stage 2 for every possible action b of Bob in stage 1. These choices are illustrated in Figure 1.5, on the left.

For these choices of Alice, observe that if Bob wants to maximize component 1, he chooses action 2, and if he wants to maximize component 2 then he chooses action 1. Thus the resulting guarantee that Alice can achieve from time 1 onwards is as shown in the figure. By varying the choices of $\bar{\alpha}$, $R(1)$ and $R(2)$, one gets the set of all achievable simultaneous guarantees. The Pareto frontier of this set, $\mathcal{V}_2^*$, is shown in the figure on the right. Observe

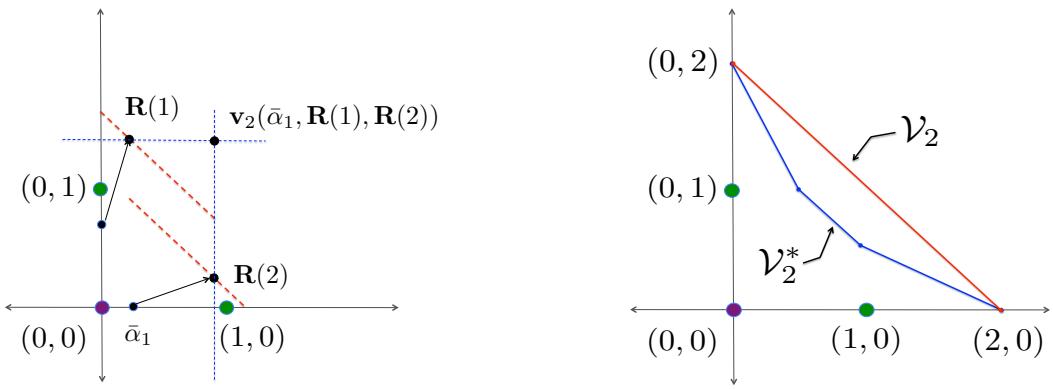

Figure 1.5. Computing the set of two-step simultaneous guarantees (left) and the minimal set of two-stage guarantees $\mathcal{V}_2^*$ (right). Also shown is the set $\mathcal{V}_2$ of minimal guarantees obtained with non-adaptive strategies

that $\mathcal{V}_2^*$ clearly gives (component-wise) better guarantees than those in $\mathcal{V}_2$ obtained by simply choosing different one shot guarantees in each stage. Thus reacting to Bob's actions has a clear advantage. As an example, the point $\bar{u} = (0.5, 1)$ is achieved by first choosing $\bar{\alpha}_1 = (0.5, 0.5)$ in stage 1, then if Bob plays action 1, then play $\bar{\alpha}_2 = (0.5, 0.5)$ in stage 2 as well, otherwise if he plays action 2, then play action 1 in stage 2.

This example is a simple illustration of the fact that the minimal simultaneous guarantees achievable in repeated games can be obtained by backward induction. One starts with the optimal set achievable in the last stage, and constructs the optimal set achievable from the second to last stage and so on till one obtains the optimal guarantees achievable from the beginning of the game. The operator that takes the minimal achievable set for an N stage repeated game and gives the minimal achievable set for the $N+1$ stage game is a set-valued dynamic programming operator defined on the space of Lower Pareto Frontiers of certain sets. This is an illustration of the general approach that will be formally established for different classes of games in the rest of this thesis.

1.3.1 A Minmax theorem due to Blackwell

Although there are no known analogues of the minmax theorem for a single-shot matrix game with vector payoffs, Blackwell, in his seminal paper [4] proved a generalization of the minmax theorem for infinitely repeated vector-valued games. The notion is geometric. Recall that an interpretation of the minmax theorem is that if the payoffs of the matrix game lie in the interval $[m, M]$, then for any sets of the form $[r, M]$ or $[m, r]$, for any player, either he/she has a strategy that ensures that the expected payoff of the game is in the given set, or other player has a strategy that ensures that the expected payoff is outside that set.

Blackwell proved a result in similar vein for infinitely repeated games with vector payoffs. He showed that that for any closed convex set, a player either has a strategy to ensure that the long-run average payoffs of the game lie in that set with a probability approaching 1 as the number of stages goes to infinity, in which case the set is said to be *approachable*, or the other player has a strategy to ensure that the average payoffs are outside that set with a probability approaching 1, in which case the set is said to be *excludable*. He also showed that a necessary and sufficient condition for a closed convex set to be approachable is that every halfspace that contains it is approachable. And a halfspace $\gamma \mathbf{x}^T \leq b$ is approachable if and only if the corresponding scalar-valued game obtained by the projection of the payoff vectors on the halfspace, i.e., the game with payoffs $\{\gamma \mathbf{r}(a,b)^T\}$ has a value that is less than or equal to b. In the case where a closed convex set is approachable, Blackwell also designed a strategy that ensures that the set is 'approached'.

Using Blackwell's results, one can determine the set of minimal simultaneous guarantees that Alice can achieve in a repeated game with limiting average rewards. We have observed, as we will see when we discuss the applications of our results, that the set of minimal guarantees for the discounted case seems to converge to this set (in a sense that would be defined later in Chapter 2) as the discount factor approaches 1. We conjecture that this holds in general. Analogous results have been shown to hold in Markov Decision Processes and zero-sum stochastic games, see [27, 30].

1.4 Organization of the thesis

Chapter 2 begins with the model of repeated games with discounting and formally establishes that the set of minimal achievable simultaneous guarantees is the fixed point of an appropriately defined set-valued dynamic programming operator defined on the space of Lower Pareto Frontiers of convex and compact sets with an appropriately defined metric. It also demonstrates a value-iteration based procedure to approximate this set and extract approximately optimal strategies.

Chapter 3 is devoted to the applications and extensions of the results in Chapter 2. Two applications are considered. First is to the problem of regret minimization in repeated games with discounted losses, where this theory yields the first characterization of the minmax optimal regret and the corresponding policy. The second is to the celebrated model of discounted repeated games with incomplete information on one side due to Aumann and Maschler [3], where this theory resolves a long-standing open problem of characterizing and computing the optimal policy for the uninformed player. We consider two extensions: one where in each stage, the actions are not chosen simultaneously, but either of the players chooses his/her action first and then the other one makes a choice. Both these cases result in different dynamic programming operators and the properties of the optimal policy changes in the case where Alice acts first.

Chapter 4 is concerned with generalizing these ideas to more general models of dynamic games with vector payoffs. We consider two classes of models. First is the class of stochastic

games. These games proceed in stages and in each stage the players are in one of a finite set of states. Depending on the state, their actions result in a vector-valued payoff and the state probabilistically moves to another state in the next stage. A repeated game can thus be seen as a stochastic game in which there is only one state. We show again that for stochastic games with discounted payoffs, the sets of minimal simultaneous guarantees achievable beginning from the different states are a fixed point of a dynamic programming operator defined on the product (one for each state) of the space of lower Pareto frontiers of convex compact sets.

Next, we consider another model of dynamic games called reachability games. In a reachability game, Alice and Bob take turns to deterministically move the state of a system within a finite set of states, according to a set of fixed rules. Given an initial state, a set of states is reachable if Bob has a strategy that guarantees that the state enters this set in a finite time. Similarly, a set of states is excludable if Alice has a strategy that guarantees that the state never reaches this set. An excludable set is maximal if any larger set that contains it is reachable. We present an efficient algorithm to compute the maximal excludable subsets of any specified set of target states and the corresponding strategies. We characterize these maximal subsets as the fixed point of a dynamic programming operator defined on the space of 'maximal' subsets of the set of target sets. Chapter 5 concludes the thesis with a summary and a discussion of possible future directions.

Chapter 2

Simultaneous Guarantees in Repeated games with vector losses

2.1 Model

Consider a two-player vector-valued game $\mathbb{G}$ defined by an action set $A = \{1, \cdots, m\}$ for player 1, who is the decision-maker and whom we will call Alice, and the action set $B = \{1, \cdots, n\}$ for player 2 who is the adversary and whom we will call Bob. For each pair of actions $a \in A$ and $b \in B$, Alice incurs a vector-valued loss $\mathbf{r}(a, b) \in \mathbb{R}^K$. For simplicity, we restrict the discussion to the case where $K = 2$, i.e. $\mathbf{r}(a, b) = \big(r_1(a, b), r_2(a, b)\big)$, although the results hold for any finite K.

The game $\mathbb{G}$ is played repeatedly in stages $t = 1, 2, 3, \cdots$. In each stage t, both Alice and Bob simultaneously pick their actions a_t and b_t respectively, and Alice bears the vector of losses $\mathbf{r}(a_t, b_t) = \big(r_1(a_t, b_t), r_2(a_t, b_t)\big)$. Fix a discount factor $\beta \in [0, 1)$. Then the vector of total discounted losses is defined as:

$$\sum_{t=1}^{\infty} \beta^{t-1} \mathbf{r}(a_t, b_t) = \left(\sum_{t=1}^{\infty} \beta^{t-1} r_1(a_t, b_t), \sum_{t=1}^{\infty} \beta^{t-1} r_2(a_t, b_t) \right). \tag{2.1}$$

An adaptive randomized strategy ϕ_A for Alice specifies for each stage t, a mapping from the set of observations till stage t, i.e., $H_t = (a_1, b_1, \cdots, a_{t-1}, b_{t-1})$, to a probability distribution on the action set A, denoted by $\Delta(A)$. Let Φ_A be the set of all such policies of Alice. Similarly, let Φ_B be the set of all adaptive randomized strategies for Bob. For a pair of strategies ϕ_A and ϕ_B, the expected discounted loss on component k in the repeated game is

given by:

$$R_k(\phi_A, \phi_B) = E_{\phi_A,\phi_B}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_k(a_t, b_t)\right], \tag{2.2}$$

where the expectation is over the randomness in the strategies ϕ_A and ϕ_B. Now consider a fixed policy $\phi_A \in \Phi_A$. If Alice plays this strategy, then irrespective of the strategy chosen by Bob, Alice guarantees that the long term expected losses on different components lie in the '*corner* set' :

$$\mathcal{C}(\phi_A) \triangleq \left\{\bar{x} \in \mathbb{R}^m : x_k \leq \max_{\phi_B^k \in \Phi_B} R_k(\phi_A, \phi_B^k) \text{ for all } k \in \{1, 2\}\right\}$$

defined by the corner point

$$\left(\max_{\phi_B^1 \in \Phi^B} R_1(\phi_A, \phi_B^1), \max_{\phi_B^2 \in \Phi^B} R_2(\phi_A, \phi_B^2)\right).$$

Let the set of all the corner points, or *simultaneous guarantees* that correspond to *all* the strategies $\phi_A \in \Phi_A$ be defined as:

$$\mathcal{W} \triangleq \left\{\left(\max_{\phi_B^1 \in \Phi^B} R_1(\phi_A, \phi_B^1), \max_{\phi_B^2 \in \Phi^B} R_2(\phi_A, \phi_B^2)\right) : \phi_A \in \Phi_A\right\}. \tag{2.3}$$

Our objective is to characterize and compute the *minimal* points in the set $\mathcal{W}$, i.e., its *Lower Pareto frontier*, which is the set

$$\mathcal{U}^* = \Lambda(\mathcal{W}) \triangleq \{\bar{x} \in \mathcal{W} : \forall\, \bar{x}' \in \mathcal{W} \setminus \{x\},\, \exists\, k \text{ s.t. } x_k < x'_k\}, \tag{2.4}$$

and compute policies for Alice in Φ_A that guarantee different points in this set.

2.1.1 Summary of main results in this chapter

- We show that the set $\mathcal{U}^*$ of *minimal* losses that a (loss minimizing) player can simultaneously guarantee in a vector-valued zero-sum repeated game with discounted losses is the fixed point of a set-valued dynamic programming operator defined on the space of Lower Pareto frontiers of closed convex sets with an appropriately defined metric. We then show that the optimal policies that guarantee different points in this set are of the following form. $\mathcal{U}^*$ can be parametrized so that each point corresponds to a 'state' in a compact state space. Each state is associated with an immediate optimal randomized action and a transition rule that depends on the observed action of the adversary. In order to attain a point in $\mathcal{U}^*$, the minimizing player starts with the corresponding state, plays the associated randomized action, transitions into another state depending on the adversary's observed action as dictated by the rule, plays the randomized action associated with the new state and so on. In particular, the strategy does not depend on the past actions of the minimizer and it depends on the past actions of

the adversary only through this state that the minimizing player keeps track of. The compactness of the state space of this strategy is in sharp contrast to the unbounded memory required to implement the Blackwell approachability strategy (and other such optimal strategies) for the average case, since it needs to keep track of the time.

- For the case where $K = 2$, we give a value-iteration based procedure to approximate $\mathcal{U}^*$ and to compute an approximately optimal policy that only uses a coarse finite quantization of the compact state space. This strategy can be simply implemented by a randomized finite-state automaton. Any desired diminishing approximation error can be attained by choosing the appropriate quantization granularity and number of iterations. Our procedure is easily extendable to an arbitrary number of actions.

2.2 Set-valued dynamic programming

In the remainder of the chapter, our goal is to compute $\mathcal{U}^* = \Lambda(\mathcal{W})$ as defined before. We will show that this set is the unique fixed point of a set-valued dynamic programming operator defined on an appropriately defined metric space of Pareto Frontiers. In order to present our results, we first define this space.

2.2.1 Defining the space of Pareto frontiers

We first define Pareto frontiers in $[0,1]^2$.

Definition 2.2.1.

(a) Let $\bar{u}, \bar{v} \in \mathbb{R}^2$. We say that $\bar{u} \preceq \bar{v}$ if $u_1 \leq v_1$ and $u_2 \leq v_2$. Also, we say that $\bar{u} \prec \bar{v}$ if $\bar{u} \preceq \bar{v}$ and $\bar{u} \neq \bar{v}$. If $\bar{u} \prec \bar{v}$, we say that $\bar{v}$ is dominated *by $\bar{u}$.*

(b) A Pareto frontier *in $[0,1]^2$ is a subset $\mathcal{V}$ of $[0,1]^2$ such that no $\bar{v} \in \mathcal{V}$ is dominated by another element of $\mathcal{V}$.*

(c) The Lower Pareto frontier *(or simply* Pareto frontier*) of $\mathcal{S} \subset [0,1]^2$, denoted by $\Lambda(\mathcal{S})$, is the set of elements of $\mathcal{S}$ that do not dominate any another element of $\mathcal{S}$.*

Figure 2.1 shows the lower Pareto frontiers of some sets in $[0,1]^2$: The Pareto frontier of a set may be empty, as is certainly the case when the set is open. The following result is useful.

Lemma 2.2.1. *Suppose that S is a compact subset of $\mathbb{R}^2$. Then $\Lambda(S)$ is non-empty.*

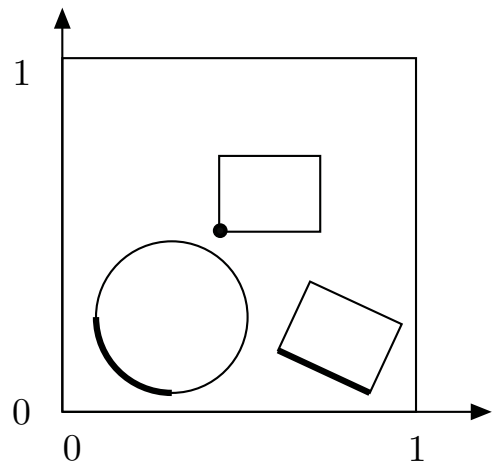

Figure 2.1. Lower Pareto frontiers of some sets in $[0,1]^2$.

Proof. For some $p \in (0,1)$ consider the minimization problem:

$$\min_{x \in S} f(x) = px_1 + (1-p)x_2.$$

Since $f(x)$ is a continuous function defined on a compact set, it achieves this minimum value at some point $\mathbf{x}(p) \in S$. Hence there cannot be any point $\bar{x}' \preceq \mathbf{x}(p)$, which means that $\mathbf{x}(p)$ is on the Pareto frontier of S. $\square$

Note that a set in $\mathbb{R}^2$ is compact iff it is closed and bounded. We define the following space of pareto frontiers:

Definition 2.2.2. *$\mathcal{F}$ is the space of Pareto frontiers of closed and convex subsets of $[0,1]^2$.*

We will now define a metric on this space. We first define the *upset* of a set, illustrated in Figure 2.2.

Definition 2.2.3. *Let $\mathcal{A}$ be a subset of $[0,1]^2$. The* upset *$up(\mathcal{A})$ of $\mathcal{A}$ is defined as $up(\mathcal{A}) = \{\bar{x} \in [0,1]^2 \mid x_1 \geq y_1$ and $x_2 \geq y_2$, for some $\bar{y} \in \mathcal{A}\}$. Equivalently, $up(\mathcal{A}) = \{\bar{x} \in [0,1]^2 \mid \bar{x} = \bar{y} + \bar{v}$, for some $\bar{y} \in \mathcal{A}$ and $\bar{v} \succeq 0\}$.*

It is immediate that the upset of the Pareto frontier of a closed convex set in $[0,1]^2$ is closed and convex.

We recall the definition of *Hausdorff* distance induced by the ∞-norm.

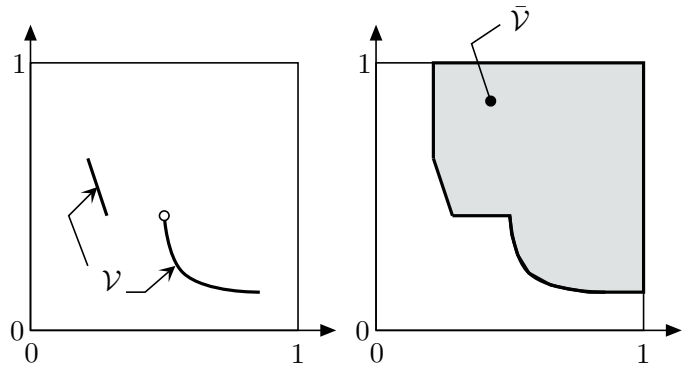

Figure 2.2. A Pareto frontier $\mathcal{V}$ and its upset $up(\mathcal{V})$.

Definition 2.2.4. *Let $\mathcal{A}$ and $\mathcal{B}$ be two subsets of $\mathbb{R}^2$. The Hausdorff distance $h(\mathcal{A}, \mathcal{B})$ between the two sets is defined as*

$$h(\mathcal{A}, \mathcal{B}) = \max\{\sup_{\bar{x}\in\mathcal{A}} \inf_{\bar{y}\in\mathcal{B}} ||\bar{x} - \bar{y}||_\infty, \sup_{\bar{y}\in\mathcal{B}} \inf_{\bar{x}\in\mathcal{A}} ||\bar{x} - \bar{y}||_\infty\}.$$

We now define the distance between two pareto frontiers in $\mathcal{F}$ as the Hausdorff distance between their upsets.

Definition 2.2.5. *For two pareto frontiers $\mathcal{U}$ and $\mathcal{V}$ in $\mathcal{F}$, we define the distance $d(\mathcal{U}, \mathcal{V})$ between them as $d(\mathcal{U}, \mathcal{V}) = h(up(\mathcal{U}), up(\mathcal{V}))$.*

We can then show that $\mathcal{F}$ is complete in the metric d. This follows from the completeness of the Hausdorff metric for closed convex subsets of $[0, 1]^2$.

Proposition 2.2.1. *Let $\mathcal{V}_n$ be a sequence in $\mathcal{F}$. Suppose that $\sup_{m,k>n} d(\mathcal{V}_m, \mathcal{V}_k) \to 0$. Then there exists a unique $\mathcal{V} \in \mathcal{F}$ such that $d(\mathcal{V}_n, \mathcal{V}) \to 0$.*

In order to prove the result, we need the following set of results about the Hausdorff distance:

Lemma 2.2.2. *a) h is a metric on the space of closed subsets of $\mathbb{R}^2$.*

b) Assume that $\{\mathcal{A}_n\}$ is a Cauchy sequence of closed subsets of $[0, 1]^2$. Then there is a unique closed subset $\mathcal{A}$ of $[0, 1]^2$ such that $h(\mathcal{A}_n, \mathcal{A}) \to 0$. This set $\mathcal{A}$ is defined as follows:

$$\mathcal{A} = \{\bar{x} \in [0, 1]^2 \mid \exists \bar{x}_n \in \mathcal{A}_n \text{ s.t. } \bar{x}_n \to \bar{x}\}.$$

c) If the sets $\{\mathcal{A}_n\}$ in b) are convex , then $\mathcal{A}$ is convex.

d) $h(up(\mathcal{A}), up(\mathcal{B})) \leq h(\mathcal{A}, \mathcal{B})$.

Proof. a)-b) This is the well-known completeness property of the Hausdorff metric; see [28].

c) Say that $\bar{x}, \bar{y} \in \mathcal{A}$. Then $\bar{x} = \lim_n \bar{x}_n$ and $\bar{y} = \lim_n \bar{y}_n$ for $\bar{x}_n \in \mathcal{A}_n$ and $\bar{y}_n \in \mathcal{A}_n$. By convexity of each $\mathcal{A}_n$, $\bar{z}_n := \lambda\bar{x}_n + (1-\lambda)\bar{y}_n \in \mathcal{A}_n$. But then, $\bar{z}_n \to \bar{z} := \lambda\bar{x} + (1-\lambda)\bar{y}$. It follows that $\bar{z} \in \mathcal{A}$, so that $\mathcal{A}$ is convex.

d) Let $\epsilon := h(\mathcal{A}, \mathcal{B})$. Pick $\bar{x} \in up(\mathcal{A})$. Then $\bar{x} = \bar{y} + \bar{v}$ for some $\bar{y} \in \mathcal{A}$ and $v \succeq 0$. There is some $\bar{y}' \in \mathcal{B}$ with $\|\bar{y} - \bar{y}'\|_\infty \leq \epsilon$. Then $\bar{x}' = \min\{\bar{y}' + v, \mathbf{1}\} \in up(\mathcal{B})$, where the minimization is component-wise. We claim that $\|\bar{x}' - \bar{x}\|_\infty \leq \epsilon$.

If $\bar{y}' + \bar{v} \in [0,1]^2$, this is clear. Assume $y_1' + v_1 > 1$. Then,

$$x_1' = 1 < y_1' + v_1 \text{ and } x_1 = y_1 + v_1 \leq 1.$$

Thus,

$$0 \leq x_1' - x_1 < y_1' + v_1 - y_1 - v_1 = y_1' - y_1.$$

Hence, $|x_1' - x_1| \leq |y_1' - y_1|$. Similarly, $|x_2' - x_2| \leq |y_2' - y_2|$. Thus, one has

$$\|\bar{x}' - \bar{x}\|_\infty \leq \|\bar{y}' - \bar{y}\|_\infty \leq \epsilon.$$

□

Now we can prove the proposition.

Proof. Under the assumptions of the proposition, $\{up(\mathcal{V})_n, n \geq 1\}$ is Cauchy in the Hausdorff metric, so that, by Lemma 2.2.2, there is a unique closed convex set such that $h(up(\mathcal{V})_n, \mathcal{A}) \to 0$. But since $h(up(\mathcal{V})_n, up(\mathcal{A})) \leq h(up(\mathcal{V})_n, \mathcal{A})$ (from Lemma 2.2.2), we have that $h(up(\mathcal{V})_n, up(\mathcal{A})) \to 0$ and hence $up(\mathcal{A}) = \mathcal{A}$. Thus the Pareto frontier $\mathcal{V}$ of $\mathcal{A}$ is then such that $d(\mathcal{V}_n, \mathcal{V}) \to 0$.

To show uniqueness of $\mathcal{V}$, assume that there is some $\mathcal{U} \in \mathcal{F}$ such that $d(\mathcal{V}_n, \mathcal{U}) \to 0$. Then, the closed convex set $up(\mathcal{U})$ is such that $h(up(\mathcal{V})_n, up(\mathcal{U})) \to 0$. By Lemma 2.2.2, this implies that $up(\mathcal{U}) = up(\mathcal{V})$, so that $\mathcal{U} = \mathcal{V}$. $\square$

2.2.2 A dynamic programming operator and the existence of a fixed point

By scaling and shifting the losses, we assume without loss of generality that $r_k(a,b) \in [0, 1-\beta]$ for all (a,b,k). Accordingly, the total discounted losses of the game take values in $[0,1]$. Now, for a closed set $\mathcal{S} \subseteq [0,1]^2$, define the following operator Ψ that maps $\mathcal{S}$ to a subset of $\mathbb{R}^2$:

$$\Psi(\mathcal{S}) = \Bigg\{ \Bigg(\max_b \Bigg\{ \sum_{a=1}^{m} \alpha_a \big[r_1(a,b) + \beta R_1(a,b) \big] \Bigg\}, \max_b \Bigg\{ \sum_{a=1}^{m} \alpha_a \big[r_2(a,b) + \beta R_2(a,b) \big] \Bigg\} \Bigg) \\ : (R_1(a,b), R_2(a,b)) \in \mathcal{S}, \bar{\alpha} \in \Delta(A) \Bigg\}. \tag{2.5}$$

This operator can be interpreted as follows. Assuming that $\mathcal{S}$ is the set of pairs of expected guarantees on losses that Alice can ensure from stage $t+1$ onwards, $\Psi(\mathcal{S})$ is the set of pairs of expected guarantees that she can ensure beginning from time t.

We will show that if $\mathcal{S}$ is closed then $\Psi(\mathcal{S})$ is closed as well. But if $\mathcal{S}$ is convex then $\Psi(\mathcal{S})$ is not necessarily convex. Nevertheless we can show that the Pareto frontier of $\Psi(\mathcal{S})$ is the Pareto frontier of some convex and compact set. Following is the result.

Lemma 2.2.3. *Let $\mathcal{S} \subseteq [0,1]^2$ be a closed set. Then $\Psi(\mathcal{S}) \subseteq [0,1]^2$ is closed. If in addition, $\mathcal{S}$ is convex, then:*

1. *Any point $\bar{u}$ in $\Lambda(\Psi(\mathcal{S}))$ is of the form:*

$$\bar{u} = \Bigg(\max_b \Bigg\{ \sum_{a=1}^{m} \alpha_a \big[r_1(a,b) \big] + \beta Q_1(b) \Bigg\}, \max_b \Bigg\{ \sum_{a=1}^{m} \alpha_a \big[r_2(a,b) \big] + \beta Q_2(b) \Bigg\} \Bigg)$$

   for some $Q(b) \in \Lambda(\mathcal{S})$, for each $b \in B$.

2. $\Lambda(\Psi(\mathcal{S})) \in \mathcal{F}$.

The first point calls for some explanation. Recall the interpretation that $\mathcal{S}$ is the set of guarantees that Alice can achieve from time $t+1$ onwards. The result says that if $\mathcal{S}$ is closed and convex, then to achieve the minimal points in the set $\Psi(\mathcal{S})$ beginning from time t, the guarantees in $\mathcal{S}$ beginning from time $t+1$ chosen by Alice need only depend on the action of Bob at time t and not on her own action at time t. In order to prove this lemma, we need a few intermediate results. First, we need the following fact:

Lemma 2.2.4. *Let $\mathcal{V}$ be the Lower Pareto frontier of a closed convex set. Then $\mathcal{V}$ is closed.*

Proof. Suppose that $\{\bar{x}^n\}$ is a sequence of points in $\mathcal{V}$ that converge to some point $\bar{x}$. Then since S is closed, $\bar{x} \in S$. We will show that $\bar{x} \in \mathcal{V}$. Suppose not. Then there is some $\bar{u} \in \mathcal{V}$ such that $\bar{u} \preceq \bar{x}$. Suppose first that $u_1 < x_1$ and $u_2 < x_2$. Then let $\epsilon = \frac{\min(x_1-u_1, x_2-u_2)}{2}$ and consider the ball of radius ϵ around $\bar{x}$, i.e.

$$B_{\bar{x}}(\epsilon) = \{\bar{y} \in \mathbb{R}^2 : \|\bar{y} - \bar{x}\|_2 \leq \epsilon\}.$$

Then for any point $\bar{y}$ in $B_{\bar{x}}(\epsilon)$, we have that $\bar{u} \preceq \bar{y}$. But since $\{\bar{x}^n\}$ converges to $\bar{x}$, there exists some point in the sequence that is in $B_{\bar{x}}(\epsilon)$, and $\bar{u}$ is dominated by this point, which is a contradiction. Hence either $u_1 = x_1$ or $u_2 = x_2$. Suppose w.l.o.g. that $u_1 < x_1$ and $u_2 = x_2$. See Figure 2.3. Let $\delta = \frac{x_1-u_1}{2}$ and consider the ball of radius δ centered at $\bar{x}$, i.e. $B_{\bar{x}}(\delta)$. Let $\bar{x}^n$ be a point in the sequence such that $\bar{x}^n \in B_{\bar{x}}(\delta)$. Now $x_1^n > u_1$ and hence it must be that $x_2^n < u_2$.

Now some $\lambda \in (0,1)$, consider a point $\bar{r} = \lambda\bar{u} + (1-\lambda)\bar{x}^n$ such that $r_1 = u_1 + \frac{\delta}{2}$. It is possible to pick such a point since $x_1 = u_1 + 2\delta$ and $|x_1^n - x_1| \leq \delta$, and hence $x_1^n > u_1 + \frac{\delta}{2}$ (please see the figure). Now $\bar{r} \in S$ since S is convex. Now $r_1 = x_1 - \frac{3\delta}{2} < x_1$ and also $r_2 < u_2 = x_2$ since $\lambda > 0$ and $x_2^n < u_2$. Let $\delta' = \frac{x_2-r_2}{2}$. Then consider the ball $B_{\bar{x}}(\delta')$ centered at $\bar{x}$. Clearly $\bar{r} \preceq \bar{y}$ for any $\bar{y} \in B_{\bar{x}}(\delta')$. But since $\{\bar{x}^n\}$ converges to $\bar{x}$, there exist some point in the sequence that is in $B_{\bar{x}}(\delta')$, and $\bar{r}$ is dominated by this point, which is again a contradiction. Thus $\bar{x} \in \mathcal{V}$ and hence $\mathcal{V}$ is closed. □

Note that the Pareto frontier of a closed set need not be a closed set, as the example in Figure 2.4 shows.

Next we define the following notion of convexity of pareto frontiers.

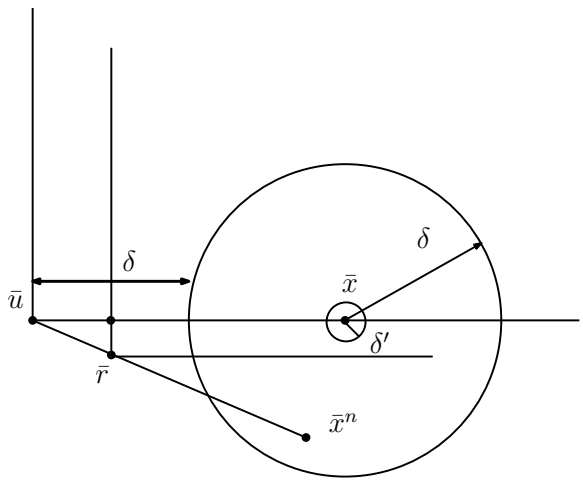

Figure 2.3. Construction in the proof of Lemma 2.2.4.

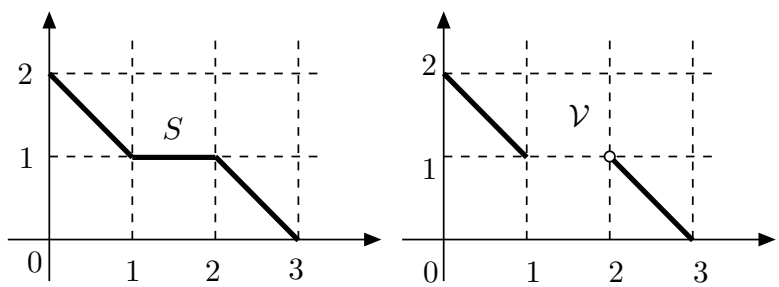

Figure 2.4. A closed set S whose Pareto frontier $\mathcal{V}$ is not closed.

Definition 2.2.6. *A Pareto frontier $\mathcal{V}$ is p-convex if for any $\bar{v}$, $\bar{u} \in \mathcal{V}$ and for each $\lambda \in [0,1]$, there exists a point $\bar{r}(\lambda) \in \mathcal{V}$ such that $\bar{r}(\lambda) \preceq \lambda\bar{v} + (1-\lambda)\bar{u}$.*

We then show the following equivalences.

Lemma 2.2.5. *For a Pareto frontier $\mathcal{V} \subset [0,1]^2$, the following statements are equivalent:*

1. *$\mathcal{V}$ is p-convex and a closed set.*
2. *$\mathcal{V}$ is p-convex and the lower Pareto frontier of a closed set $\mathcal{S} \subseteq [0,1]^2$.*
3. *$\mathcal{V}$ is the lower Pareto frontier of a closed convex set $H \subseteq [0,1]^2$.*

Proof. 1 is a special case of 2 and hence 1 implies 2. To show that 2 implies 3, if $\mathcal{S}$ is convex

then there is nothing to prove. So assume $\mathcal{S}$ is not convex. Then let $\mathcal{H}$ be the convex hull of $\mathcal{S}$. First, since $[0,1]\times[0,1]$ is convex, $H \subseteq [0,1]\times[0,1]$. Then since $\mathcal{S}$ is closed and bounded, it is also compact. Hence $\mathcal{H}$ is the convex hull of a compact set, which is compact and hence closed and bounded. Now we will show that $\mathcal{V}$ is the Lower Pareto frontier of $\mathcal{H}$. To see this, any $\bar{u} \in \mathcal{S}$ is of the form $\bar{u} = \lambda\bar{x} + (1-\lambda)\bar{y}$ where $\bar{x}, \bar{y} \in \mathcal{S}$. But then there are points $\bar{x}', \bar{y}' \in \mathcal{V}$ such that $\bar{x}' \preceq \bar{x}$ and $\bar{y}' \preceq \bar{y}$. Thus we have that

$$\lambda\bar{x}' + (1-\lambda)\bar{y}' \preceq \lambda\bar{x} + (1-\lambda)\bar{y} = \bar{u}.$$

But since $\mathcal{V}$ is convex, there exists some $\bar{r}(\lambda) \in \mathcal{V}$ such that

$$\bar{r}(\lambda) \preceq \lambda\bar{x}' + (1-\lambda)\bar{y}' \preceq \lambda\bar{x} + (1-\lambda)\bar{y} = \bar{u}.$$

Thus the Pareto frontier of $\mathcal{H}$ is a subset of $\mathcal{V}$, but since $\mathcal{V} \in \mathcal{H}$ and it is a Pareto frontier, $\mathcal{V}$ is the Lower Pareto frontier of $\mathcal{H}$. Finally Lemma 2.2.4 shows that 3 implies that $\mathcal{V}$ is closed. To show it is convex, suppose that $\bar{u}$ and $\bar{v}$ are two points in $\mathcal{V}$. Since they also belong to H, which is convex, for each $\lambda \in [0,1]$, $\lambda\bar{u} + (1-\lambda)\bar{v} \in \mathcal{S}$ and thus there is some $\bar{r}(\lambda) \in \mathcal{V}$ such that $\bar{r}(\lambda) \preceq \lambda\bar{u} + (1-\lambda)\bar{v}$. Thus $\mathcal{V}$ is convex. □

We can now finally prove the lemma.

Proof. Note that $\Psi(\mathcal{S})$ is the image of the continuous function f from the product space $\mathcal{S}^{m\times n} \times \Delta(A)$ to a point in $\mathbb{R}^2$, which is a Hausdorff space. Since $\mathcal{S}$ is closed and bounded, it is compact. Also the simplex $\Delta(A)$ is compact. Thus by Tychonoff's theorem, the product space $\mathcal{S}^{m\times n} \times \Delta(A)$ is compact. Hence by the closed map Lemma, f is a closed map and hence $\Psi(\mathcal{S})$ is closed.

Now assume that $\mathcal{S}$ is a closed convex set. Then $\Lambda(\mathcal{S})$ exists by Lemma 3.1 and further it is p-convex by Lemma 2.2.5. Let $\mathcal{U} = \Lambda(\mathcal{S})$. Clearly, $\Lambda(\Psi(\mathcal{S})) = \Lambda(\Psi(\mathcal{U}))$. Recall that any point $\bar{u}$ in $\Lambda(\Psi(\mathcal{U}))$ is of the form:

$$\bar{u} = \left(\max_b \left\{ \sum_{a=1}^{m} \alpha_a \big[r_1(a,b) + \beta R_1(a,b) \big] \right\}, \max_b \left\{ \sum_{a=1}^{m} \alpha_a \big[r_2(a,b) + \beta R_2(a,b) \big] \right\} \right)$$

for some $\bar{\alpha} \in \Delta(A)$ and $R(a,b) \in \mathcal{U}$. But since $\mathcal{U}$ is p-convex, for each $b \in B$, there exists some $Q(b) \in \mathcal{U}$ such that $Q(b) \preceq \sum_{a=1}^{m} \alpha_a R(a,b)$. Hence statement 1 follows.

Now let

$$\bar{u} = \left(\max_b \left\{ \sum_{a=1}^{m} \alpha_a \big[r_1(a,b)\big] + \beta Q_1(b) \right\}, \max_b \left\{ \sum_{a=1}^{m} \alpha_a \big[r_2(a,b)\big] + \beta Q_2(b) \right\} \right)$$

and

$$\bar{v} = \left(\max_b \left\{ \sum_{a=1}^{m} \eta_a \big[r_1(a,b)\big] + \beta R_1(b) \right\}, \max_b \left\{ \sum_{a=1}^{m} \eta_a \big[r_2(a,b)\big] + \beta R_2(b) \right\} \right)$$

be two points in $\Lambda(\Psi(\mathcal{U}))$, where $\bar{\alpha}$, $\bar{\eta} \in \Delta(A)$ and $Q(b)$, $R(b) \in \mathcal{V}$.

For a fixed $\lambda \in [0,1]$, let $\kappa_a = \alpha_a \lambda + \eta_a(1-\lambda)$. Then

$$\lambda\bar{u}+(1-\lambda)\bar{v} = \left(\lambda \max_b \left\{ \sum_{a=1}^{m} \alpha_a \big[r_1(a,b)\big] + \beta Q_1(b) \right\} + (1-\lambda) \max_b \left\{ \sum_{a=1}^{m} \eta_a \big[r_1(a,b)\big] + \beta R_1(b) \right\}, \right.$$

$$\left. \lambda \max_b \left\{ \sum_{a=1}^{m} \alpha_a \big[r_2(a,b)\big] + \beta Q_2(b) \right\} + (1-\lambda) \max_b \left\{ \sum_{a=1}^{m} \eta_a \big[r_2(a,b)\big] + \beta R_2(b) \right\} \right)$$

$$\succeq \left(\max_b \left\{ \sum_{a=1}^{m} \kappa_a \big[r_1(a,b)\big] + \beta\lambda Q_1(b) + (1-\lambda) R_1(b) \right\}, \right.$$

$$\left. \max_b \left\{ \sum_{a=1}^{m} \kappa_a \big[r_2(a,b)\big] + \beta\lambda Q_2(b) + (1-\lambda) R_2(b) \right\} \right)$$

$$\succeq \left(\max_b \left\{ \sum_{a=1}^{m} \kappa_a \big[r_1(a,b)\big] + \beta L_1(b) \right\}, \max_b \left\{ \sum_{a=1}^{m} \kappa_a \big[r_2(a,b)\big] + \beta L_2(b) \right\} \right)$$

The first inequality holds since max is a convex function and the second follows since $\mathcal{U}$ is p-convex, and hence $L(b) = (L_1(b), L_2(b)) \in \mathcal{U}$ that satisfy the given relation exist. Thus $\Lambda(\Psi(\mathcal{S}))$ is p-convex. And hence from Lemma 2.2.5, it is the Lower Pareto frontier of a closed convex set in $[0,1]^2$, i.e. it is in $\mathcal{F}$. $\square$

Define the following dynamic programming operator Φ on $\mathcal{F}$. We will call it the generalized Shapley operator, since Shapley [29] was the first to propose a dynamic programming operator to compute the minmax in zero-sum Stochastic games.

Definition 2.2.7. *(Generalized Shapley (GS) operator) For $\mathcal{V} \in \mathcal{F}$, we define $\Phi(\mathcal{V}) = \Lambda(\Psi(\mathcal{V}))$.*

Now since $\mathcal{V}$ is the Lower Pareto frontier of some closed convex subset of $\mathbb{R}^2$, say $\mathcal{S}$, and since $\Lambda(\Psi(\mathcal{V})) = \Lambda(\Psi(\mathcal{S}))$, from Lemma 2.2.3, we know that $\Phi(\mathcal{V}) \in \mathcal{F}$ whenever $\mathcal{V} \in \mathcal{F}$. Next, we claim that Φ is a contraction in the metric d.

Theorem 2.2.1.

$$d(\Phi(\mathcal{U}), \Phi(\mathcal{V})) \leq \beta d(\mathcal{U}, \mathcal{V}). \tag{2.6}$$

In order to prove this lemma, we first define another metric on the space $\mathcal{F}$ that is equivalent to d.

Definition 2.2.8. *For two Pareto frontiers $\mathcal{A}$ and $\mathcal{B}$ of $[0,1]^2$, we define*

$$e(\mathcal{A}, \mathcal{B}) \triangleq$$

$$\inf\{\epsilon \geq 0 : \forall\, \bar{u} \in \mathcal{A},\, \exists\, \bar{v} \in \mathcal{B} \text{ s.t. } \bar{v} \preceq \bar{u} + \epsilon\mathbf{1} \text{ and } \forall\, \bar{v} \in \mathcal{B},\, \exists\, \bar{u} \in \mathcal{A} \text{ s.t. } \bar{u} \preceq \bar{v} + \epsilon\mathbf{1}\}. \tag{2.7}$$

Here $\mathbf{1} = (1,1)$.

We show that the two metrics are equivalent.

Lemma 2.2.6.

$$e(\mathcal{A}, \mathcal{B}) = d(\mathcal{A}, \mathcal{B}).$$

Proof. Suppose that $e(\mathcal{A}, \mathcal{B}) \leq \epsilon$. Consider a point $\bar{x} \in up(\mathcal{A})$ such that $\bar{x} = \bar{y} + \bar{v}$ where $\bar{y} \in \mathcal{A}$ and $\bar{v} \succeq 0$. Suppose that there is no $\bar{x}' \in up(\mathcal{B})$ such that $\|\bar{x} - \bar{x}'\|_\infty \leq \epsilon$. This means that $up(\mathcal{B})$ is a subset of the region $\mathcal{S}$ shown in the Figure 2.5. But since $\bar{y} = \bar{x} - \bar{v}$, $\bar{y}$ is in region $\mathcal{S}'$. But for any $\bar{u} \in \mathcal{S}$ and $\bar{w} \in \mathcal{S}'$, $\|\bar{u} - \bar{w}\|_\infty \geq \epsilon$. This contradicts the fact that for $\bar{y}$ there is some $\bar{y}' \in \mathcal{B}$, such that $\bar{y} + \epsilon\mathbf{1} \succeq \bar{y}'$. Thus $d(\mathcal{A}, \mathcal{B}) \leq \epsilon$. Now suppose that $d(\mathcal{A}, \mathcal{B}) \leq \epsilon$. Then for any $\bar{x} \in \mathcal{A}$, there is a $\bar{x}' \in up(\mathcal{B})$ such that $\|\bar{x} - \bar{x}'\|_\infty \leq \epsilon$ where $\bar{x}' = \bar{y} + \bar{v}$ for $\bar{y} \in \mathcal{B}$ and $\bar{v} \succeq 0$. Thus $\bar{x} + \epsilon\mathbf{1} \succeq \bar{x}' = \bar{y} + \bar{v}$. Thus $e(\mathcal{A}, \mathcal{B}) \leq \epsilon$. $\square$

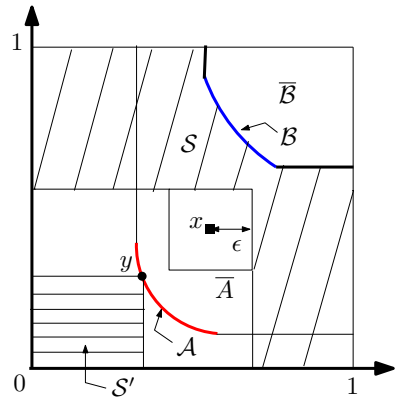

Figure 2.5. Construction for the proof of Lemma 2.2.6

We can now prove Lemma 3.3.

Proof. Suppose $e(\mathcal{U}, \mathcal{V}) = \epsilon$. Let

$$\left(\max_b \left\{ \sum_{a=1}^{m} \alpha_a r_1(a,b) + \beta R_1(b) \right\}, \max_b \left\{ \sum_{a=1}^{m} \alpha_a r_2(a,b) + \beta R_2(b) \right\} \right)$$

be some point in $\Phi(\mathcal{V})$, where $\bar{\alpha} \in \Delta(A)$. Then for each $R(a,b)$, for each a and b, we can choose $Q(b) \in \mathcal{U}$ such that $Q(b) \preceq R(b) + \epsilon \mathbf{1}$. We then have

$$\begin{aligned}
&\max_b \left\{ \sum_{a=1}^{m} \alpha_a r_1(a,b) + \beta Q_1(b) \right\} \\
&= \max_b \left\{ \sum_{a=1}^{m} \alpha_a r_1(a,b) + \beta R_1(b) + \beta(Q_1(b) - R_1(b)) \right\} \\
&\leq \max_b \left\{ \sum_{a=1}^{m} \alpha_a r_1(a,b) + \beta R_1(b) + \beta\epsilon \right\} \\
&= \max_b \left\{ \sum_{a=1}^{m} \alpha_a r_1(a,b) + \beta R_1(b) \right\} + \beta\epsilon.
\end{aligned}$$

Similarly, we can show that

$$\max_b \left\{ \sum_{a=1}^{m} \alpha_a r_2(a,b) + \beta Q_2(b) \right\} \leq \max_b \left\{ \sum_{a=1}^{m} \alpha_a r_2(a,b) + \beta R_2(b) \right\} + \beta\epsilon. \tag{2.8}$$

Thus

$$\begin{aligned}&\left(\max_b\left\{\sum_{a=1}^m \alpha_a r_1(a,b)+\beta Q_1(b)\right\},\max_b\left\{\sum_{a=1}^m \alpha_a r_2(a,b)+\beta Q_2(b)\right\}\right)\\ \preceq\ &\left(\max_b\left\{\sum_{a=1}^m \alpha_a r_1(a,b)+\beta R_1(b)\right\},\max_b\left\{\sum_{a=1}^m \alpha_a r_2(a,b)+\beta R_2(b)\right\}\right)+\beta\epsilon\mathbf{1}.\end{aligned}$$

But since $\left(\max_b\left\{\sum_{a=1}^m \alpha_a r_1(a,b)+\beta Q_1(b)\right\},\max_b\left\{\sum_{a=1}^m \alpha_a r_2(a,b)+\beta Q_2(b)\right\}\right)\in\Psi(\mathcal{U})$, and since $\Phi(\mathcal{U})=\Lambda(\Psi(\mathcal{U}))$, there exists some $(L_1,L_2)\in\Phi(\mathcal{U})$ such that

$$(L_1,L_2)\preceq\left(\max_b\left\{\sum_{a=1}^m \alpha_a r_1(a,b)+\beta Q_1(b)\right\},\max_b\left\{\sum_{a=1}^m \alpha_a r_2(a,b)+\beta Q_2(b)\right\}\right).$$

Thus

$$(L_1,L_2)\preceq\left(\max_b\left\{\sum_{a=1}^m \alpha_a r_1(a,b)+\beta R_1(b)\right\},\max_b\left\{\sum_{a=1}^m \alpha_a r_2(a,b)+\beta R_2(b)\right\}\right)+\beta\epsilon\mathbf{1}.$$

We can show the other direction (roles of $\Phi(\mathcal{U})$ and $\Phi(\mathcal{V})$ reversed) similarly and thus we have that

$$e(\Phi(\mathcal{U}),\Phi(\mathcal{V}))\leq\beta\epsilon=\beta e(\mathcal{U},\mathcal{V}). \tag{2.9}$$

□

Finally we show that the GS operator has a unique fixed point and starting from a Pareto frontier in $\mathcal{F}$, the sequence of frontiers obtained by a repeated application of this operator converges to the unique fixed point.

Theorem 2.2.2. *Let $\mathcal{V}\in\mathcal{F}$. Then the sequence $(\mathcal{A}_n=\Phi^n(\mathcal{V}))_{n\in\mathbb{N}}$ converges in the metric d to a Pareto frontier $\mathcal{V}^*\in\mathcal{F}$, which is the unique fixed point of the operator Φ, i.e., the unique solution of $\Phi(\mathcal{V})=\mathcal{V}$.*

Proof. Since Φ is a contraction in the metric d, the sequence $\{\mathcal{A}_n\}$ is Cauchy in $\mathcal{F}$. Hence by Lemma 2.2.2, $\{\mathcal{A}_n\}$ converges to a Pareto frontier $\mathcal{V}^*\in\mathcal{F}$. The continuity of the operator further implies that

$$\mathcal{V}^*=\Phi(\mathcal{V}^*).$$

To show uniqueness, observe that if there are two fixed points $\mathcal{U}$ and $\mathcal{V}$, then we have

$$d(\mathcal{U}, \mathcal{V}) = d(\Phi(\mathcal{U}), \Phi(\mathcal{V})) \leq \beta d(\mathcal{U}, \mathcal{V}),$$

which implies that $d(\mathcal{U}, \mathcal{V}) = 0$ and hence $\mathcal{U} = \mathcal{V}$. □

We can then show that $\mathcal{V}^*$ is indeed the optimal set $\mathcal{U}^*$ defined in (3.7) that we are looking for.

Theorem 2.2.3. $\mathcal{U}^* = \mathcal{V}^*$.

Proof. Fix $N \geq 1$ and consider a truncated game where Alice can guarantee the cumulative losses in $\beta^{N+1}\mathcal{V}^*$ after time $N+1$. Then the minimal losses that she can guarantee after time N is the set:

$$\Lambda\Bigg(\Bigg\{\max_{b\in B} \beta^N \sum_{a\in A} \alpha_a r_1(a,b) + \beta^{N+1} Q_1(b), \max_{b\in B} \beta^N \sum_{a\in A} \alpha_a r_2(a,b) + \beta^{N+1} Q_2(b)$$

$$\mid \bar{\alpha} \in \Delta(A), b \in B, Q(b) \in \mathcal{V}^*, \forall\, b \in B\Bigg\}\Bigg).$$

This set is $\beta^N \mathcal{V}^*$. By induction, this implies that the set of minimal losses that she can guarantee after time 0 is V^*.

The losses of the truncated game and of the original game differ only after time $N+1$. Since the losses at each step are bounded by $(1-\beta)$, the cumulative losses after time $N+1$ are bounded by $\frac{\beta^{N+1}(1-\beta)}{1-\beta} = \beta^{N+1}$. Consequently, the minimal losses of the original game must be in the set

$$\{(\bar{u} \in [0,1]^2 : u_1 \in [x_1 - \beta^{N+1}, x_1 + \beta^{N+1}], u_2 \in [x_2 - \beta^{N+1}, x_2 + \beta^{N+1}],\ x \in \mathcal{V}^*\}.$$

Since $N \geq 1$ is arbitrary, the minimal losses that Alice can guarantee in the original game must be in $\mathcal{V}^*$. □

2.2.3 Optimal policies: Existence and Structure

For a Pareto frontier $\mathcal{V} \in \mathcal{F}$, one can define a one-to-one function from a parameter set $\mathcal{P}$ to $\mathcal{V}$. Such a function parameterizes the Pareto frontier. For instance, consider the function $F^{\mathcal{V}}(p) : [0,1] \to \mathcal{V}$, where one defines

$$F^{\mathcal{V}} = \arg\min_{x \in \mathcal{V}} \{px_1^2 + (1-p)x_2^2\}. \tag{2.10}$$

This function is indeed one-to-one because there is only one ellipse $px_1^2 + (1-p)x_2^2 = c$ that shares a tangent at a particular point of $\mathcal{V}$. We now express the GS operator in the form of such a parametrization. Assume that $\mathcal{V}^*$ is such that $\mathcal{V}^* = \Phi(\mathcal{V}^*)$. For $p \in \mathcal{P}$, choose $\bar{\alpha}(p) \in \Delta(A)$ and $q(b,p) \in \mathcal{P}$ for each $b \in B$ such that for $k \in \{1,2\}$,

$$F_k^{\mathcal{V}^*}(p) = \max_b \{\sum_{a \in A} \alpha_a(p) r_k(a,b) + \beta F_k^{\mathcal{V}^*}(q(b,p))\}.$$

Then we have the following result.

Theorem 2.2.4. *For any $p_0 \in \mathcal{P}$, the pair of upper bounds on losses $x = F^{\mathcal{V}^*}(p_0)$ in $\mathcal{V}^*$ is guaranteed by Alice first choosing action $a_0 \in A$ with probability $\alpha_a(p_0)$. Then if Bob chooses an action $b_0 \in B$, the optimal guarantees to choose from the second step on are then $\beta F^{\mathcal{V}^*}(p_1)$ in $\beta\mathcal{V}^*$, where $p_1 = q(b_0, p_0)$, which can be guaranteed by Alice by choosing action $a_1 \in A$ with probability $\alpha_a(p_1)$, and so on.*

Proof. Assume that Alice can guarantee every pair $\beta^{N+1}\bar{u}$ of cumulative losses with $\bar{u} \in \mathcal{V}^*$ after time $N+1$ by choosing some continuation strategy in Φ_A. Let $\bar{x} = F(p, \mathcal{V}^*)$. We claim that after time N, Alice can guarantee a loss of no more than $\beta^N \bar{x}$ on each component by first choosing $a_N = a$ with probability $\alpha_a(p)$ and then if Bob chooses $b \in B$, choosing a continuation strategy that guarantees her $F(p', \mathcal{V}^*)$, where $p' = q(b,p)$. Indeed by following this strategy, her expected loss after time N is then

$$\{\beta^N \sum_a \alpha_a(p) r_k(a,b) + \beta^{N+1} F_k^{\mathcal{V}^*}(q(b,p))\} \leq \beta^N F_k^{\mathcal{V}^*}(p) = \beta^N x_k$$

when the game is G_k. Thus, this strategy for Alice guarantees that her loss after time N is no more than $\beta^N \mathcal{V}^*$. Hence by induction, following the indicated strategy (in the statement of the theorem) for the first N steps and then using the continuation strategy from time

$N+1$ onwards, guarantees that her loss is not more than $F(p_0, \mathcal{V}^*)$ after time 0. Now, even if Alice plays arbitrarily after time $N+1$ after following the indicated strategy for the first N steps, she still guarantees that her loss is no more than $F(p_0, \mathcal{V}^*) + \beta^{N+1}(1,1)^T$. Since this is true for arbitrarily large values of N, playing the indicated policy indefinitely guarantees that her loss is no more than $F(p_0, \mathcal{V}^*)$. □

This implies that $\mathcal{P}$ can be thought of as an *information state space*. Each state is associated with an immediate optimal randomized action and a transition rule that depends on the observed action of Bob. In order to attain a point in $\mathcal{V}^*$, Alice starts with the corresponding state, plays the associated randomized action, transitions into another state depending on Bob's observed action as dictated by the rule, plays the randomized action associated with the new state and so on. *In particular, the policy does not depend on the past actions of Alice and it depends on the past actions of Bob only through this information state that Alice keeps track of.* Since Alice's optimal policy itself does not depend on her own past actions, Bob's optimal response does not depend on them either. Hence one can see that Bob has an *oblivious* best response to any optimal policy of Alice.

2.3 Approximating the optimal frontier

We now proceed to propose a computational procedure to approximate the optimal pareto frontier in $\mathbb{R}^2$ and devise approximately optimal policies. In order to do so, we need to define an approximation of a Pareto frontier. Consider the following approximation scheme for a Pareto frontier $\mathcal{V} \in \mathcal{F}$. For an integer N, choose $2N+1$ lines defined as: $\{y = x \pm \frac{k}{N} : k = 1, 2, \cdots, N\}$. Now for a pareto frontier $\mathcal{V}$ in $\mathcal{F}$, define the vector valued function $F_N^{\mathcal{V}}(p) : \{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\} \to \mathbb{R}^2$, where for each p,

$$F_N^{\mathcal{V}}(p) = \arg\min x + y \tag{2.11}$$

$$\text{s.t. } x \geq x_1,\ y \geq x_2,\ (x_1, x_2) \in \mathcal{V},\ y = x + p.$$

$F_N^{\mathcal{V}}(p)$ is essentially the point of intersection of the line $y = x + p$ with the upset of $\mathcal{V}$ in $[0,2]^2$ (see Figure 2.6). Define the approximation operator to be $\Gamma_N(\mathcal{V}) = \Lambda\Big(\text{ch}\Big(\big\{F_N^{\mathcal{V}}(p) : p \in \{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\}\big\}\Big)\Big)$. Here ch denotes the convex hull of a set. Now suppose that $\mathcal{V}$ is the pareto frontier of a convex and compact set. Then we know that $\Phi(\mathcal{V})$ is also the Pareto frontier of a convex and compact set, and thus we can express the compound operator $\Gamma_N(\Phi(\mathcal{V}))$ via a set of explicit optimization problems as in 2.11. Consider the value-iteration based approximation procedure defined in Algorithm 1. We then we have the following result:

Algorithm 1: A procedure for approximating $\mathcal{V}^*$:

- Fix integer N and number of iterations n.
- Initialize $F^0(p) = (0,0)$ for all $p \in \{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\}$.
- For $i = 0 : n-1$ and for $p \in \{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\}$, solve

$$F^{i+1}(p) = \arg\min x_1 + x_2$$

$$\text{s.t. } x_1 \geq \sum_{a \in A} \alpha_a(p) r_1(a,b) + \beta Q_1(b,p), \;\; x_2 \geq \sum_{a \in A} \alpha_a(p) r_2(a,b) + \beta Q_2(b,p),$$

$$x_2 = x_1 + p, \;\; \bar{\alpha}(p) \in \Delta(A), \;\; Q_1(b,p) \geq F_1^i(1), \;\; Q_2(b,p) \geq F_2^i(-1).$$

$$\text{If } F_1^i(q) - F_1^i(q') \neq 0 \text{ then}$$

$$Q_2(b,p) - F_2^i(q) \geq \frac{F_2^i(q) - F_2^i(q')}{F_1^i(q) - F_1^i(q')}(Q_1(b,p) - F_1^i(q)) \, \forall \, q, q' \text{ s.t. } |q - q'| = \frac{1}{N},$$

$$\text{else if } F_1^i(q) - F_1^i(q') = 0 \text{ then}$$

$$Q_1(b,p) \geq F_1^i(q).$$

- $\mathcal{G}_n = \Lambda\Big(\text{ch}\Big(\big\{F^n(p) : p \in \{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\}\big\}\Big)\Big)$.

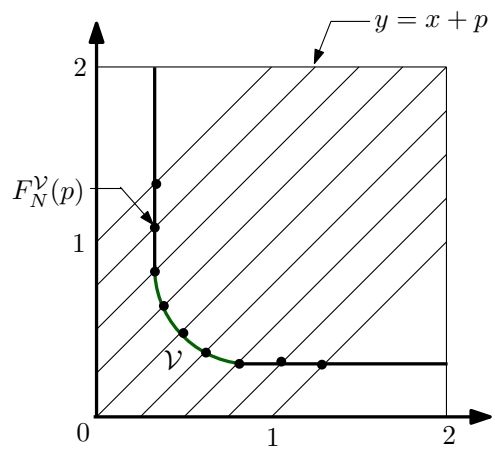

Figure 2.6. Approximating $\mathcal{V}$.

Theorem 2.3.1.

$$d(\mathcal{V}^*, \mathcal{G}_n) \leq \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right) + \beta^n. \tag{2.12}$$

We first need the following lemma about the approximation operator.

Lemma 2.3.1. *Consider a $\mathcal{V} \in \mathcal{F}$. Then*

$$d(\mathcal{V}, \Gamma_N(\mathcal{V})) \leq \frac{1}{N}.$$

Proof. Any point in $\Gamma_N(\mathcal{V})$ is of the form $\lambda\bar{u} + (1-\lambda)\bar{v}$ where $\bar{u}, \bar{v} \in \mathcal{V}$. By the p-convexity of $\mathcal{V}$, there is some $\bar{r}(\lambda) \in \mathcal{V}$, such that $\bar{r}(\lambda) \preceq \lambda\bar{u} + (1-\lambda)\bar{v}$. Also clearly for any $\bar{u} \in \mathcal{V}$,

$$\min\left\{||\bar{u} - \bar{v}||_\infty : \bar{v} \in \Gamma_{N,M}(\mathcal{V})\right\} \leq \max\left\{||F_{M,N}^{\mathcal{V}}(p) - F_{M,N}^{\mathcal{V}}(p')||_\infty : |p' - p| = \frac{1}{N}\right\} = \frac{1}{N}.$$

□

Next, consider the sequence of functions (F^n) generated by the procedure. Define

$$\mathcal{G}_n = \Lambda\left(\text{ch}\left(\left\{F^n(p) : p \in \{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\}\right\}\right)\right)$$

and consider the corresponding sequence $(\mathcal{G}_n)$. From the definition of F^n, one can see that

$$\mathcal{G}_n = \Gamma_N(\Phi(\mathcal{G}_{n-1})) \tag{2.13}$$

Note that $\mathcal{G}_0 = \{(0,0)\}$. Now consider another sequence of pareto frontiers

$$\Big(\mathcal{A}_n = \Phi^n(\mathcal{G}_0)\Big)_{n\in\mathbb{N}} \tag{2.14}$$

Then we have

$$\begin{aligned} d(\mathcal{A}_n, \mathcal{G}_n) &= d(\Phi(\mathcal{A}_{n-1}), \Gamma_N(\Phi(\mathcal{G}_{n-1}))) && (2.15)\\ &\overset{(a)}{\leq} d(\Phi(\mathcal{A}_{n-1}), \Phi(\mathcal{G}_{n-1})) + d(\Phi(\mathcal{G}_{n-1}), \Gamma_N(\Phi(\mathcal{G}_{n-1}))) && (2.16)\\ &\overset{(b)}{\leq} \beta d(\mathcal{A}_{n-1}, \mathcal{G}_{n-1}) + \frac{1}{N} && (2.17) \end{aligned}$$

where inequality (a) is the triangle inequality and (b) follows from (3.27) and Lemma 2.3.1. Coupled with the fact that $d(\mathcal{A}_0, \mathcal{G}_0) = 0$, we have that

$$\begin{aligned} d(\mathcal{A}_n, \mathcal{G}_n) &\leq \frac{1}{N}\Big(1 + \beta + \beta^2 + \cdots \beta^{n-1}\Big) && (2.18)\\ &= \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right) && (2.19) \end{aligned}$$

Since Φ is a contraction, the sequence $\{\mathcal{A}_n\}$ converges to some pareto frontier $\mathcal{V}^*$. Suppose that we stop the generation of the sequences $\{\mathcal{A}_n\}$ and $\{\mathcal{G}_n\}$ at some n. Now since $\mathcal{A}_0 = \mathcal{G}_0 = \{(0,0)\}$, and since the stage payoffs $r_k(a,b) \in [0, 1-\beta]$, we have that $d(\mathcal{A}_1, \mathcal{A}_0) \leq 1-\beta$. Using the contraction property of the GS operator, this implies that $d(\mathcal{V}^*, \mathcal{A}_n) \leq \frac{\beta^n(1-\beta)}{1-\beta} = \beta^n$ and thus by triangle inequality we have

$$d(\mathcal{V}^*, \mathcal{G}_n) \leq \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right) + \beta^n. \tag{2.20}$$

Hence for any ϵ, there is a pair (N, n) such that $d(\mathcal{V}^*, \mathcal{G}_n) \leq \epsilon$.

2.3.1 Extracting an approximately optimal policy

A *2N+1-mode policy* γ is a mapping from each $p \in \{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\}$ to the pair

$$\Bigg(\bar{\alpha}(p), \Bigg\{\Big(q(b,p),\, q'(b,p),\, \kappa(b,p)\Big) : q(b,p),\, q'(b,p) \in \{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\}$$

$$\text{s.t. } |q(b,p) - q'(b,p)| = \frac{1}{N},\, \kappa(b,p) \in [0,1],\, \forall\, b \in B\Bigg\}\Bigg). \tag{2.21}$$

The interpretation is that if the current '*mode*' is p, then Alice first chooses action $a \in A$ with probability $\alpha_a(p)$. Then if Bob plays action $b \in B$, Alice considers the new mode to be $q(b,p)$ with probability $\kappa(b,p)$ and $q'(b,p)$ with probability $1-\kappa(b,p)$ and plays accordingly thereafter.

Now consider the optimization problem (2.3) that corresponds to $i = n$, i.e., the problem the procedure would have solved if it was allowed to continue for one more iteration. Now one can extract a $2N+1$-mode policy γ_n from the solution of this optimization problem as follows. Defining $\bar{\alpha}(p)$ is immediate. Now note that the optimal $Q(b,p)$ is such that either $Q(b,p) = F^n(1)$ or $Q(b,p) = F^n(-1)$, or $Q(b,p) = \kappa(b,p)F^n(q) + (1-\kappa(b,p))F^n(q')$ for some $\kappa(b,p) \in [0,1]$ and some q, q' such that $|q-q'| = \frac{1}{N}$. These define $\kappa(b,p)$, $q(b,p)$ and $q'(b,p)$ in our policy. If $Q(b,p) = F^n(1)$, then $\kappa(b,p) = 1$ and $q(b,p) = 1$, where as if $Q(b,p) = F^n(-1)$ then $\kappa(b,p) = 0$ and $q'(b,p) = -1$.

Let $\mathcal{V}^{\gamma_n}$ be the corresponding Pareto frontier that is attained by the policy γ_n by choosing different possible initial randomizations over the $2N+1$ modes. We have the following result.

Theorem 2.3.2.

$$d(\mathcal{V}^{\gamma_n}, \mathcal{V}^*) \leq \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right) + 2\beta^n + \frac{1}{N}\left(\frac{2-\beta^n-\beta^{n+1}}{(1-\beta)^2}\right). \tag{2.22}$$

In order to prove this we need a few intermediate definitions and results. First, we need to characterize the losses guaranteed by any $2N+1$-mode policy. Such a policy γ defines the following operator on a function $F : \{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\} \to \mathbb{R}^2$.

$$\Delta_N^\gamma(F)(p) = \Bigg(\max_{b\in B} \Big\{ \sum_{a\in A} \alpha_a(p) r_1(a,b) + \kappa(b,p)\beta F_1(q(b,p)) + (1-\kappa(b,p))\beta F_1(q'(b,p)) \Big\},$$

$$\max_{b\in B} \Big\{ \sum_{a\in A} \alpha_a(p) r_2(a,b) + \kappa(b,p)\beta F_2(q(b,p)) + (1-\kappa(b,p))\beta F_2(q'(b,p)) \Big\} \Bigg). \tag{2.23}$$

For a function $F : \{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\} \to \mathbb{R}^2$, define the following norm:

$$\|F\| = \max_{p\in\{0,\pm\frac{1}{N},\pm\frac{2}{N},\cdots,\pm\frac{N-1}{N},\pm 1\}} \|F(p)\|_\infty.$$

We can easily show that Δ_N^γ is a contraction in the norm.

Lemma 2.3.2.

$$\|\Delta_N^\gamma(F) - \Delta_N^\gamma(G)\| \leq \beta\|F-G\|. \tag{2.24}$$

We can then show the following result.

Lemma 2.3.3. *Consider a $2N+1$-mode policy γ. Then there is a unique function F^γ : $\{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\} \to \mathbb{R}^2$ such that*

$$\Delta_N^\gamma(F^\gamma) = F^\gamma.$$

The policy γ initiated at mode p where $p \in \{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\}$, guarantees the vector of losses $F^\gamma(p)$.

The first part of the result follows from the fact that the operator is a contraction and the completeness of the space of vector-valued functions with a finite domain for the given norm. The second part follows from arguments similar to those in the proof of Theorem 3.3.

Now let $\mathcal{V}^{\gamma_n} = \Lambda(ch(\{F^{\gamma_n}(p) : p \in \{0, \pm\frac{1}{N}, \pm\frac{2}{N}, \cdots, \pm\frac{N-1}{N}, \pm 1\}\}))$, where F^{γ_n} is the fixed point of the operator $\Delta_N^{\gamma_n}$.

We then have that

$$\begin{aligned} d(\mathcal{V}^{\gamma_n}, \mathcal{V}^*) &\leq d(\mathcal{V}^{\gamma_n}, \mathcal{G}_n) + d(\mathcal{G}_n, \mathcal{V}^*) && (2.25)\\ &\leq d(\mathcal{V}^{\gamma_n}, \mathcal{G}_n) + \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right) + \beta^n. && (2.26) \end{aligned}$$

The following result is immediate.

Lemma 2.3.4.

$$d(\mathcal{V}^{\gamma_n}, \mathcal{G}_n) \leq \|F^{\gamma_n} - F^n\|. \tag{2.27}$$

Next we have

$$\begin{aligned} \|F^{\gamma_n} - F^n\| &\leq \|F^{\gamma_n} - \Delta_N^{\gamma_n}(F^n)\| + \|\Delta_N^{\gamma_n}(F^n) - F^n\| && (2.28)\\ &\overset{(a)}{=} \|\Delta_N^{\gamma_n}(F^{\gamma_n}) - \Delta_N^{\gamma_n}(F^n)\| + \|F^{n+1} - F^n\| && (2.29)\\ &\overset{(b)}{\leq} \beta\|F^{\gamma_n} - F^n\| + \|F^{n+1} - F^n\|. && (2.30)\\ & && (2.31) \end{aligned}$$

Here (a) holds because $\Delta_N^{\gamma_n}(F^n) = F^{n+1}$ by the definition of the policy γ_n, and also because F^{γ_n} is a fixed point of the operator $\Delta_N^{\gamma_n}$. (b) holds because $\Delta_N^{\gamma_n}$ is a contraction. Thus we have

$$d(\mathcal{V}^{\gamma_n}, \mathcal{G}_n) \leq \|F^{\gamma_n} - F^n\| \leq \frac{\|F^{n+1} - F^n\|}{1-\beta}. \tag{2.32}$$

And finally we have:

$$d(\mathcal{V}^{\gamma_n}, \mathcal{V}^*) \leq \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right) + \beta^n + \frac{\|F^{n+1} - F^n\|}{1-\beta}. \tag{2.33}$$

To finish up, we need the following result:

Lemma 2.3.5.

$$\|F^{n+1} - F^n\| \leq d(\mathcal{G}_{n+1}, \mathcal{G}_n).$$

Proof. Let $\bar{u} = F^{n+1}(p)$ and $\bar{v} = F^n(p)$ for some p. Now $\bar{u}$ is the point of intersection of $\mathcal{G}_{n+1}$ and the line $y = x + p$. $\bar{v}$ is the point of intersection of the frontier $\mathcal{G}_n$ and the line $y = x + p$. Now suppose that $\|\bar{u} - \bar{v}\|_\infty > d(\mathcal{G}_{n+1}, \mathcal{G}_n)$. Then either for $\bar{u}$, there is no $\bar{r} \in \mathcal{G}_n$ such that $\bar{r} \preceq \bar{u} + \mathbf{1}d(\mathcal{G}_{n+1}, \mathcal{G}_n)$ or for $\bar{v}$, there is no $\bar{r} \in \mathcal{G}_{n+1}$ such that $\bar{r} \preceq \bar{v} + \mathbf{1}d(\mathcal{G}_{n+1}, \mathcal{G}_n)$. Either of the two cases contradict the definition of $d(\mathcal{G}_{n+1}, \mathcal{G}_n)$. Thus $\|\bar{u} - \bar{v}\|_\infty \leq d(\mathcal{G}_{n+1}, \mathcal{G}_n)$. □

Finally, by the triangle inequality we have

$$\begin{aligned} d(\mathcal{G}_{n+1}, \mathcal{G}_n) &\leq d(\mathcal{A}_{n+1}, \mathcal{A}_n) + d(\mathcal{G}_{n+1}, \mathcal{A}_{n+1}) + d(\mathcal{G}_n, \mathcal{A}_n) && (2.34) \\ &\leq (1-\beta)\beta^n + \frac{1}{N}\left(\frac{1-\beta^{n+1}}{1-\beta}\right) + \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right). && (2.35) \end{aligned}$$

Combining with (2.33) we have the result.

2.3.2 Remarks

Note that the procedure to approximate the frontier and extract an approximately optimal policy is not optimized for complexity: it is mainly presented to illustrate that our characterization of the minmax optimal policy via the fixed point of a dynamic programming operator opens up the possibility of using several dynamic programming based approximation procedures. In particular, we have not tried to determine an algorithm that achieves the optimal error-complexity tradeoff. For fixed (N, n), in order to approximate the optimal frontier, the procedure needs to solve nN linear programs, each with $O(N)$ variables and constraints to give the corresponding error bound in the theorem. One can split the error into two terms: the first term is the quantization error which is bounded by $\frac{1}{N(1-\beta)}$ and the iteration error which is bounded by β^n. The second term is relatively benign but the first term requires $N = \frac{1}{(1-\beta)\epsilon}$, which grows rapidly when β is close to 1. For finding an approximately optimal policy, the scaling is like $\frac{1}{(1-\beta)^2\epsilon}$, which is even worse. Nevertheless, note that all of this computation can be done offline. The resulting approximately optimal policy is very simple to implement, and requires a small memory.

Chapter 3

Applications and Extensions

In this chapter, we first discuss two applications of the results in the previous chapter. Then we will also discuss some extensions of the core ideas. In all of our discussion so far we have restricted ourselves to games with simultaneous moves, i.e., the two players Alice and Bob choose their actions simultaneously at each stage. We will now relax this assumption and consider two related dynamic programming operators: one for the case where Alice chooses her action first in each step, and then Bob chooses his action after having observed Alice's action, and vice versa. We will see that the structure of the optimal policy for Alice changes if she moves first: in this case, the information state transitions not only depend on Bob's actions, but also on her own actions.

3.1 Application 1: Regret minimization in discounted repeated games

Several types of sequential decision-making problems in an adversarial environment can be modeled as a repeated game between an agent and the environment (the adversary), where at each time step, the agent chooses one of several available actions and the environment simultaneously chooses the loss incurred by the agent for each action. An important example that has been particularly studied due to its numerous applications (e.g., in financial decision making) is the problem of combining expert advice (a variant of the path selection problem that we introduced briefly in Chapter 1). In this problem on each day, the decision-maker chooses to act based on the recommendations made by a set of experts (see e.g., [9] for a survey). In these settings, the notion of 'regret' is of central importance: it measures the difference between the player's actual loss and the loss that she would have incurred if she had always chosen the single best action against the realized sequence of loss vectors in

hindsight. It is desirable to use a *no-regret* strategy, that is a strategy that ensures that the average regret vanishes as the number of time steps increases regardless of the environment's behavior.

A powerful method to obtain no-regret strategies is to transform the game into a vector-valued repeated game, where the different components keep track of the additional loss incurred relative to the loss incurred if each of the possible actions were always chosen in the past, and use Blackwell approachability theory [4] to obtain a strategy that guarantees a zero average loss on each component. As we discussed in Chapter 1, theory of approachability gives sufficient conditions for a set to be *approachable* by a player in a repeated game with vector losses, which means that there exists a strategy for a player that ensures that the average loss approaches this set regardless of the adversary's actions. Moreover, it explicitly defines an adaptive randomized strategy that ensures this.

Blackwell approachability is an elegant theory that addresses a fundamental problem in multi-objective decision-making, as a result of which it has found applications in various online learning problems (see [18], [9, Section 7.8], [26]). However, an important drawback of this theory is that it is only applicable to average losses. In the real world, losses incurred in the near future are more damaging than those incurred later. This is usually captured by introducing a discount factor $\beta \in (0, 1)$ and weighting the t-th stage loss by $(1-\beta)\beta^{t-1}$ (note that $(1-\beta)$ is just a normalizing factor, which ensures that the sum of the weights is 1). This weighted average of losses is called the discounted average. When current losses are more important than future ones, we cannot expect losses incurred in the initial stages to have an increasingly negligible contribution to the total loss as the number of stages increases, and hence we cannot expect to achieve a vanishing average discounted regret. A straightforward application of Blackwell approachability strategy in the discounted case gives an expected regret bound of $O(\sqrt{1-\beta})$ [25] that can be large if β is not close to one.

Curiously, computing the exact minimal expected worst-case regret for the discounted case, or computing strategies that guarantee this regret has been an open problem. But now that we can compute the set of minimal simultaneous upper bounds on the losses that can be achieved by a player in a discounted repeated game with vector losses, we can finally resolve this problem. In this section, we will demonstrate this.

3.1.1 Related work

The first study of regret minimization in repeated games dates back to the pioneering work of Hannan [17] who first proposed an algorithm that achieves vanishing average regret. Since then, numerous other algorithms have been proposed for regret minimization, particularly in the experts setting [24, 33, 8, 15]. Other settings with more limited feedback have been considered, most notably the multi-armed bandit setting [2, 7]. Stronger notions of regret such as internal regret, have also been defined, and corresponding minimizing strategies have been proposed [14, 10, 6, 31].

Approachability theory was introduced by Blackwell [4] to study achievable guarantees in vector-valued repeated games. It was also Blackwell who first noticed that approachabil-

ity can be used to obtain no-regret strategies [5]. This theory was subsequently extended in various ways [32, 23], and stronger connections with learning problems such as regret minimization or calibration were shown [1, 26], always for the case of average rewards.

The idea of regret minimization with non-uniformly weighted losses has been considered before in the literature. These works derive bounds on the regret of natural extensions of no-regret algorithms ([10], Thm 2.8; [26], Prop. 6]. While these bounds are useful (average regret goes to zero) if the weights satisfy a non-summability condition, they can be quite crude if this condition is not satisfied (in particular for the natural discounting of losses that we consider where weights decrease exponentially) because the average regret in this case does not vanish. Also, despite the existence of lower bounds ([10], Thm 2.7), the minimax regret was unknown. In contrast to deriving such bounds, we derive the exact optimal minmax regret and policy for the natural discounting with any given discount factor. Also, a few works (see [11] in particular) derive better bounds for the case where future losses are given a higher weight that current ones, but such a weighting goes against the notion of time value of money. The difference between our results and prior literature is similar to the difference between the optimal policy for the discounted multi-armed bandit problem due to Gittins [16] and the regret minimization for the stochastic multi-armed bandit problem due to Lai and Robbins [22]: the first analysis gives exact optimality using dynamic programming while the second is concerned with rate optimality and uses techniques that are quite different.

3.1.2 Repeated games with discounted losses

Let G be a two player game with m actions $A = \{1, \ldots, m\}$ for player 1, who is assumed to be the minimizer and who we will call Alice (the decision-maker), and n actions $B = \{1, \ldots, n\}$ for player 2, who is the adversary and who we will call Bob. For each pair of actions $a \in A$ and $b \in B$, the corresponding loss for Alice is $l(a, b) \in \mathbb{R}$. The losses for different pairs of actions are known to Alice. The game G is played repeatedly in stages $t = 1, 2, \cdots$. In each stage, both Alice and Bob simultaneously pick their actions $a_t \in A$ and $b_t \in B$ and Alice incurs the corresponding loss $l(a_t, b_t)$. The loss of the repeated game is defined to be the total discounted loss given by $\sum_{t=1}^{\infty} \beta^{t-1} l(a_t, b_t)$. We define the total discounted regret of Alice as:

$$\sum_{t=1}^{\infty} \beta^{t-1} l(a_t, b_t) - \min_{a \in A} \sum_{t=1}^{\infty} \beta^{t-1} l(a, b_t), \tag{3.1}$$

which is the difference between her actual discounted loss and the loss corresponding to the single best action against the sequence of actions chosen by Bob in hindsight. An adaptive randomized strategy ϕ_A for Alice specifies for each stage t, a mapping from the set of observations till stage t, i.e., $H_t = (a_1, b_1, \cdots, a_{t-1}, b_{t-1})$, to a probability distribution on the action set A, denoted by $\Delta(A)$. Let Φ_A be the set of all such policies of Alice.

The adversary Bob is assumed to choose a deterministic oblivious strategy, i.e., his choice is simply a sequence of actions $\phi_B = (b_1, b_2, b_3, \cdots)$ chosen before the start of the game. Let Φ_B be the set of all such sequences. We would like to compute the worst case or minmax

expected discounted regret which is defined as:

$$\min_{\phi_A \in \Phi_A} \max_{\phi_B \in \Phi_B} E_{\phi_A}\left[\sum_{t=1}^{\infty} \beta^{t-1} l(a_t, b_t)\right] - \min_{a \in A} \sum_{t=1}^{\infty} \beta^{t-1} l(a, b_t), \tag{3.2}$$

and the strategy for Alice that guarantees this value. Here the expectation is over the randomness in Alice's strategy. We can equivalently write this as:

$$\min_{\phi_A \in \Phi_A} \max_{\phi_B \in \Phi_B} \max_{a \in A} E_{\phi_A}\left[\sum_{t=1}^{\infty} \beta^{t-1} (l(a_t, b_t) - l(a, b_t))\right]. \tag{3.3}$$

In order to address this objective, it is convenient to define a vector-valued game $\mathbb{G}$, in which, for a pair of actions $a \in A$ and $b \in B$, the vector of losses is $r(a, b)$ with m components (recall that $|A| = m$), where

$$r_k(a, b) = l(a, b) - l(k, b) \tag{3.4}$$

for $k = 1, \cdots, m$. $r_k(a, b)$ is the single-stage additional loss that Alice bears by choosing action a instead of action k, when Bob chooses b: the so called single-stage regret with respect to action k. For a choice of strategies $\phi_A \in \Phi_A$ and $\phi_B \in \Phi_B$ of the two players, the expected loss on component k in this vector-valued repeated game is given by

$$R_k(\phi_A, \phi_B) \;=\; E_{\phi_A}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_k(a_t, b_t)\right], \tag{3.5}$$

where the expectation is over the randomness in Alice's strategy. Now observe that by playing a fixed policy $\phi_A \in \Phi_A$, irrespective of the strategy chosen by Bob, Alice guarantees that the long term expected losses on different components lie in the '*corner* set' :

$$\mathcal{C}(\phi_A) \triangleq \left\{\bar{x} \in \mathbb{R}^m : x_k \leq \max_{\phi_B^k \in \Phi_B} R_k(\phi_A, \phi_B^k) \text{ for all } k \in \{1, \cdots, m\}\right\}$$

defined by the corner point $\left(\max_{\phi_B^k \in \Phi^B} R_k(\phi_A, \phi_B^k)\right)_{k=1,\cdots,m}$. Suppose that we determine the set of all the corner points, or *simultaneous guarantees* that correspond to *all* the strategies $\phi_A \in \Phi_A$, defined as:

$$\mathcal{W} \triangleq \left\{\left(\max_{\phi_B^k \in \Phi_B} R_k(\phi_A, \phi_B^k)\right)_{k=1,\cdots,m} : \phi_A \in \Phi_A\right\}. \tag{3.6}$$

Then it is easy to see that $\min_{\phi_A \in \Phi_A} \max_{\phi_B \in \Phi_B} \max_{a \in A} E_{\phi_A}\left[\sum_{t=1}^{\infty} \beta^{t-1}(l(a_t, b_t) - l(a, b_t))\right] = \min_{\bar{x} \in \mathcal{W}} \max_k x_k$. In fact, we are only interested in finding the *minimal* points in the set $\mathcal{W}$, i.e., its *Lower Pareto frontier*, which is the set

$$\mathcal{U}^* = \Lambda(\mathcal{W}) \triangleq \{x \in \mathcal{W} : \forall\, x' \in \mathcal{W} \setminus \{x\},\, \exists\, k \text{ s.t. } x_k \leq x'_k\}. \tag{3.7}$$

Our results help us exactly characterize this set as we will now argue. In fact, $\mathcal{U}^*$ is exactly the fixed point $\mathcal{V}^*$ of the GS operator in Definition 2.2.7. The only problem in showing this seems to be that in our discussion on repeated games with vector losses, we endowed the adversary Bob with a richer strategy space than being able to just choose a sequence of actions offline and deterministically. We will show that this does not make any difference. First note that the best response to the optimal strategy of Alice that achieves different points on the frontier $\mathcal{V}^*$ is deterministic and offline. This is because Alice does not use Bob's actions to determine the information state transitions. In fact, if Alice is restricted to use strategies that do not depend on her own actions chosen in the past, then the best response to such policy is always an offline deterministic policy, and hence the minimal achievable frontier if Bob is restricted to use offline deterministic policies is $\mathcal{V}^*$. So all we need to show is that Alice does not gain by using policies that depend on her own past actions, when Bob is restricted to using only offline deterministic strategies. To see this, suppose that $\mathcal{V}_t$ is the set of guarantees that Alice can achieve from time $t+1$ onwards by using general randomized adaptive strategies, assuming that Bob is restricted to using deterministic offline policies. Then the guarantees that she can achieve at time t are given by first choosing a distribution over her actions $\bar{\alpha}$ and then a mapping from (a, b) to some continuation (randomized adaptive) policy $\phi(a, b)$ in response to the realized action a and Bob's action b. But since Bob's responses that maximize the losses on the different components cannot depend on the realization of Alice's action a, and can only depend on $\bar{\alpha}$, his best responses from time $t+1$ onwards would effectively be against the strategy $\phi'(b)$ of Alice that chooses the policy $\phi(a, b)$ with probability α_a for each realized action a. Note that such policy guarantees a point in $\mathcal{V}_t$. Thus the guarantees that Alice can achieve from time t onwards is given by the set:

$$\mathcal{V}_{t+1} = \Lambda\Bigg(\Bigg\{ \max_{b\in B} \beta^N \sum_{a\in A} \alpha_a r_1(a,b) + \beta^{N+1} Q_1(b), \max_{b\in B} \beta^N \sum_{a\in A} \alpha_a r_2(a,b) + \beta^{N+1} Q_2(b)$$

$$\mid \bar{\alpha} \in \Delta(A), b \in B, Q(b) \in \mathcal{V}_t, \forall\, b \in B \Bigg\}\Bigg).$$

But this is exactly the dynamic programming operator in Definition 2.2.7. Hence we can conclude from the Theorems 2.2.3 and 2.2.4 that $\mathcal{V}^*$ is indeed the set of minimal guarantees, even if Bob is restricted to using deterministic offline policies.

3.1.3 Example: Repeated path selection

Consider the following problem that we introduced in Chapter 1. Alice can choose one of two paths to go to work each day. Each path may be congested, leading to a loss of 1 for Alice, or it may be clear, leading to no loss. Each day, without knowing the state of the paths, Alice chooses some path, possibly randomly. After having taken this path, she learns the congestion levels of the two paths in hindsight. We reproduce the matrix of losses in Table 3.1. The matrix of single-stage regrets is reproduced in Table 3.2.

Table 3.1. Possible loss scenarios.

Path 1	1	0	1	0
Path 2	0	1	1	0

Table 3.2. Single-stage regret.

Path 1	(0,1)	(0,-1)	(0,0)	(0,0)
Path 2	(-1,0)	(1,0)	(0,0)	(0,0)

In Figure 3.1, the computed approximately optimal Pareto frontiers for a range of values of β are shown with the corresponding (theoretical) approximation errors as given by Theorem 2.3.1. Note that these frontiers appear to converge to the optimal frontier for the average case, i.e. the single point $(0, 0)$, as could be expected. In Table 3.3, for illustration purposes we compute an approximately optimal $11 - mode$ policy ($N = 5$) for $\beta = 0.8$. The second column contains the probability of choosing Path 1 in each of the different modes and columns 3 and 4 give the transition rules to the different modes if Path 1 incurs a loss and if Path 2 incurs a loss respectively. If both the experts incur a loss or incur no loss then one stays in the same mode as before (although its is sub-optimal for the adversary to choose that option).

For $\beta \leq 0.5$ we can argue that the following trivial policy is regret-optimal: choose either of the experts with equal probability in the first stage, and from the next stage onwards always choose the expert that incurred no loss in the first stage (repeat if both experts incurred the same loss). The reason is that, because the discounting is so high, once an expert incurs no loss in the first stage, even if he always incurs loss in the future stages, he is still the best expert in the long run (since $\beta = 0.5$, the first stage is as important as all the future stages). Thus the optimal policy is to just choose this expert forever.

3.2 Application 2: Repeated games with incomplete information on one side

One of the most celebrated and well studied models of dynamic games with incomplete information is a model introduced by Aumann and Maschler of zero-sum repeated games with incomplete information on one side, see [3]. It is described as follows.

There are K two person zero-sum games $G_1, \cdots, G_K$, each with m actions, $A = \{1, \cdots, m\}$ for player 1, who is the minimizer (Alice), and n actions $B = \{1, \cdots, n\}$ for player 2, who is the maximizer (Bob). For simplicity, consider the case where $K = 2$. Let the payoff corresponding to actions a and b of players 1 and 2 respectively be denoted by $r_1(a, b)$ in game G_1 and $r_2(a, b)$ in game G_2. We further restrict ourselves to the setting where in each of the games G_1 and G_2, Alice and Bob play their actions simultaneously.

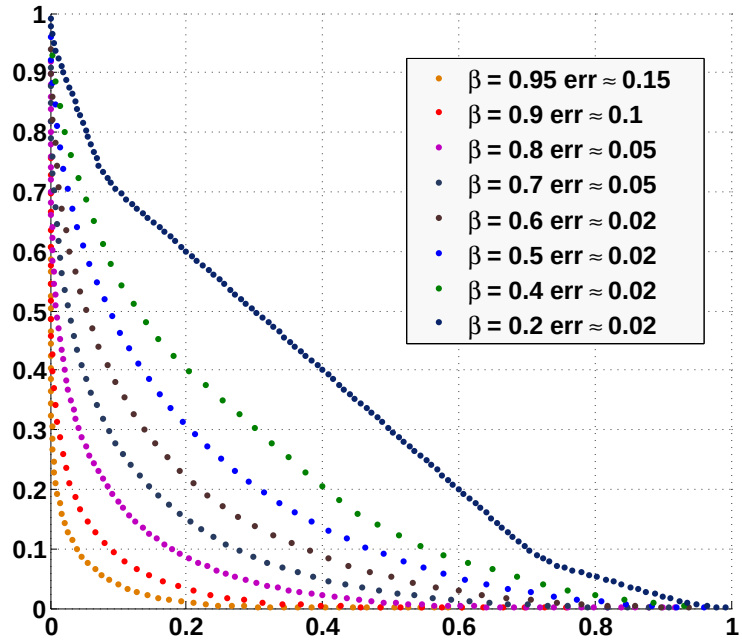

Figure 3.1. Approximations of $(1-\beta)\mathcal{V}^*$ for different β values with corresponding errors

Table 3.3. An approximately optimal 11-mode policy for $\beta = 0.8$.

Modes	Pr(Path 1)	Transition if Path 1 incurs loss	Transition if Path 2 incurs loss
+5	0.994	+4 w.p. 0.63 and +5 w.p. 0.37	Stay in +5
+4	0.97	+3 w.p. 0.88 and +4 w.p. 0.12	Go to +5
+3	0.9113	+1 w.p. 0.13 and +2 w.p. 0.87	Go to +5
+2	0.8082	+1 w.p 0.62 and 0 w.p. 0.38	+5 w.p. 0.37 and +4 w.p. 0.63
+1	0.6656	-1 w.p. 0.62 and 0 w.p. 0.38	+4 w.p. 0.12 and +3 w.p. 0.88
0	0.5	-2 w.p. 0.87 and -1 w.p. 0.13	+2 w.p. 0.87 and +1 w.p. 0.13
-1	0.3344	-4 w.p. 0.12 and -3 w.p. 0.88	+1 w.p. 0.62 and 0 w.p. 0.38
-2	0.1918	-5 w.p. 0.37 and -4 w.p. 0.63	-1 w.p 0.62 and 0 w.p. 0.38
-3	0.0887	Go to -5	-1 w.p. 0.13 and -2 w.p. 0.87
-4	0.03	Go to -5	-3 w.p. 0.88 and -4 w.p. 0.12
-5	0.006	Stay in -5	-4 w.p. 0.63 and -5 w.p. 0.37

We define the game G^∞ as follows. One of the two games G_1 and G_2 is chosen by nature with probability $(p, 1-p)$ respectively. This distribution is known to both the players but the actual choice of the game is informed to Bob and not to Alice. Let the chosen game be denoted by G. Then this game G is played infinitely often in stages $t = 1, \cdots, \infty$. At each stage t, Alice and Bob play their actions simultaneously. The payoff that is incurred by the players is not observed by Alice at any stage, but she observes Bob's actions. An adaptive randomized strategy (also called a behavioral strategy) ϕ^1 for Alice specifies for each time t, a mapping from her set of observations till time t, i.e. $H_t^1 = (a_1, b_1, \cdots, a_{t-1}, b_{t-1})$, to $\Delta(A)$. A behavioral strategy ϕ^2 for Bob specifies for each time t, a mapping from his set of observations till time t *and the choice of the game* G, i.e. $H_t^2 = (G, a_1, b_1, \cdots, a_{t-1}, b_{t-1})$, to $\Delta(B)$. We will express the behavioral strategy ϕ^2 of Bob as $\phi^2 = (\phi_1^2, \phi_2^2)$, where ϕ_i^2 is his strategy conditioned on the event $\{G = G_i\}$.

One needs to specify the objectives of the two players in G^∞. For a discount factor $\beta \in (0, 1)$ and for a choice of strategies ϕ_1 and ϕ_2 of the two players, the ex-ante expected payoff is given by

$$R(\phi^1, \phi^2) = E_{\phi^1,\phi^2,G}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_G(a_t, b_t)\right] \tag{3.8}$$

$$= pE_{\phi^1,\phi_1^2}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_1(a_t, b_t)\right] + (1-p)E_{\phi^1,\phi_2^2}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_2(a_t, b_t)\right] \tag{3.9}$$

Alice's objective is to minimize this payoff while Bob's objective is to maximize it. The minmax or the upper value of the game is given by

$$\overline{\mathbf{V}} = \min_{\phi^1} p \max_{\phi_1^2} E_{\phi^1,\phi_1^2}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_1(a_t, b_t)\right] + (1-p) \max_{\phi_2^2} E_{\phi^1,\phi_2^2}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_2(a_t, b_t)\right]. \tag{3.10}$$

The minimizing strategy in the outer minimization problem is the minmax strategy for Alice and it will simply be called her optimal strategy. Similarly, the maxmin or the lower value of the game is given by

$$\underline{\mathbf{V}} = \max_{\phi_1^2,\phi_2^2} \min_{\phi^1} \left(pE_{\phi^1,\phi_1^2}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_1(a_t, b_t)\right] + (1-p)E_{\phi^1,\phi_2^2}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_2(a_t, b_t)\right]\right). \tag{3.11}$$

The optimal strategy for Bob is similarly defined as his maxmin strategy, i.e. the maximizing strategy in the outer maximization problem. In general, we have that $\overline{\mathbf{V}} \geq \underline{\mathbf{V}}$, but in this case one can show that a minmax theorem holds and $\overline{\mathbf{V}} = \underline{\mathbf{V}}$, see [30, 34].

One would to determine the structure of the optimal policy for the uninformed player (Alice) and provide a computationally efficient algorithm to compute this policy. In order to see why this is a difficult problem, it is instructive to think about the corresponding question for Bob, the informed player.

Computing the maxmin policy for Bob : To compute the maxmin policy for Bob (see [30, 34]) one can use a dynamic programming approach that exploits the structural

relationship between the original game and the game after one stage has elapsed. Suppose $V(p)$ is a function that assigns to every prior probability p of choosing game G_1, the maxmin value of the associated infinitely repeated game. Then one can show that the maxmin value is the fixed point of the following dynamic programming operator defined on the function V:

$$\Psi(V)(p) = \max_{q_1^2, q_2^2 \in \Delta A^2} \min_{q^1 \in \Delta A^1} pE_{q_1^2, q^1}[r_1(a,b)] + (1-p)E_{q_2^2, q^1}[r_2(a,b)] \tag{3.12}$$

$$+ \sum_{b \in A^2} \beta(pq_1^2(b) + (1-p)q_2^2(b))V\left(\frac{pq_1^2(b)}{pq_1^2(b) + (1-p)q_2^2(b)}\right). \tag{3.13}$$

To see this intuitively, notice that in the first stage, any probability distribution over Bob's actions, chosen by him for the two games G_1 and G_2 as a part of his strategy ϕ^2, makes his realized action an informative signal of the true game chosen by nature. Since Alice is assumed to know this strategy in her inner minimization, she can perform a Bayesian update of her belief about the chosen game. Thus once the randomization of Bob in the first stage is fixed, there is a one-stage expected reward that is minimized by Alice, and then every realized action of Bob results in a new game, which has a maxmin value corresponding to the computed posterior distribution, weighted by β. Bob thus chooses a randomization (for the two choices) that maximizes the sum of these two values. Consistency then requires that the function $V(p)$ has to be the fixed point of this resulting operator. One can show that the operator is a contraction and that a fixed point exists. Further one can also show that the optimal policy for Bob is a stationary policy that depends only on the posterior p_t at stage t and does not depend on the actions of player 2.

Now what is the problem with using a similar approach to computing the minmax policy for Alice, the uninformed player? The problem is that in order to perform the Bayesian update as a part of her policy ϕ^1, Alice needs to know Bob's policy ϕ^2, which means that ϕ^1 presupposes the knowledge of ϕ^2, which contradicts the fact that the maxmin policy is 'universal': it guarantees that her loss is no more than $\overline{V}$ irrespective of the strategy chosen by Bob. Even if Bob's optimal strategy is unique, the best response strategy of Alice that computes the posterior updates at each stage and plays optimally accordingly is vulnerable to bluffing by Bob. Thus the optimal strategy of Alice cannot rely on the computation of these posterior distributions and must depend instead on Bob's actions and the corresponding losses incurred in the different possible choices of games.

The computation of Alice's optimal policy has been an open problem. Structurally, it is known (see [30]) that in the optimal policy, Alice's decision at stage t depends on Bob's actions till time t and not on her own actions. This also suggests the possibility that any dynamic programming based procedure that may be developed to compute this policy may suffer from the curse of dimensionality, i.e., the state may include the entire history of actions.

The key step that resolves this problem is the following. Instead of computing the upper value $\overline{\mathbf{V}}$ corresponding to the prior distribution p, suppose that one computes the following set:

$$\mathcal{W} = \left\{ \left(\max_{\phi_1^2} E\left[\sum_{t=1}^{\infty} \beta^{t-1} r_1(a_t, b_t)\right], \max_{\phi_2^2} E\left[\sum_{t=1}^{\infty} \beta^{t-1} r_2(a_t, b_t)\right] \right) : \phi^1 \in \Phi^1 \right\}. \tag{3.14}$$

This is the set of upper guarantees that Alice can simultaneously achieve on the two components of the vector of the long term discounted payoffs, by playing all the possible strategies in Φ^1. If we determine this set, then one can simply choose a point $\bar{r}(p) \in \overline{\mathcal{V}}$ such that

$$\bar{r}(p) = \arg \min_{(r_1,r_2)\in\overline{\mathcal{V}}} pr_1 + (1-p)r_2. \tag{3.15}$$

The corresponding strategy of Alice that results in the simultaneous guarantee $\bar{r}$ is then the optimal policy in the original game. In fact, one need not compute the entire set $\mathcal{W}$, but just its lower Pareto frontier, since in any case the optimal point that solves (3.15) has to be on this frontier. Then we are interested in characterizing the set $\mathcal{U}^* = \Lambda(\overline{\mathcal{V}})$. But this is exactly the set we characterized in Chapter 2.

Note that we solve a harder problem than the one we set out to solve, since instead of computing the minmax value corresponding to one prior p, we are trying to simultaneously compute the minmax values corresponding to all the possible priors. But it turns out that this harder objective makes this problem suddenly become amenable to a dynamic programming based approach. This should not be too surprising, since as we have seen for the case of the informed player, in order to solve for the lower value corresponding to a prior p and to compute the optimal strategy, one needs to simultaneously solve for games starting from all possible priors $p \in [0,1]$.

3.3 Games with alternating moves

In the previous discussion, we had restricted ourselves to the case where Alice and Bob take actions simultaneously in each stage. In this section, we relax this assumption and consider two possibilities: one where Alice moves first and the one where Bob (the adversary) moves first. Both these cases result in different dynamic programming operators.

3.3.1 Case 1: Alice plays first

In this case, adaptive randomized strategy ϕ_A for Alice specifies for each stage t, a mapping from the set of observations till stage t, i.e., $H_t = (a_1, b_1, \cdots, a_{t-1}, b_{t-1})$ to $\Delta(A)$. For Bob, his strategy ϕ_B specifies for each stage t, a mapping from the set of observations till stage t and Alice's action at time t, i.e., $H_t = (a_1, b_1, \cdots, a_{t-1}, b_{t-1}, a_t)$ to $\Delta(B)$. Consider the following operator defined on frontier in $\mathcal{F}$:

Definition 3.3.1. *For $\mathcal{V} \in \mathcal{F}$, define*

$$\Phi(\mathcal{V}) = \Lambda\left\{\left(\left\{\sum_{a=1}^{m} \alpha_a \max_b \left[r_1(a,b) + \beta R_1(a,b)\right]\right\}, \left\{\sum_{a=1}^{m} \alpha_a \max_b \left[r_2(a,b) + \beta R_2(a,b)\right]\right\}\right)\right.$$

$$: (R_1(a,b), R_2(a,b)) \in \mathcal{V}, \bar{\alpha} \in \Delta(A)\Big\}. \tag{3.16}$$

The interpretation is as before: suppose that $\mathcal{V}$ is the set of minimal guarantees that Alice can achieve from time $t+1$ onwards, then $\Phi(\mathcal{V})$ is the set of guarantees that she can achieve from time t onwards. Similar to the simultaneous moves case, Alice first chooses an action with a probability distribution in $\Delta(A)$ and then, depending on the action chosen by Bob, chooses a point in $\mathcal{V}$. But in this case, in contrast to the simultaneous moves case, her action is observed by Bob before he acts and thus in the dynamic programming operator, the order of the *max* and expectation (over the randomness in Alice's action) is interchanged. We can then show the following:

Lemma 3.3.1. $\Phi(\mathcal{V}) \in \mathcal{F}$

Proof. Since $\mathcal{V} \in \mathcal{F}$, it is clear from the closed map lemma that $\Phi(\mathcal{V})$ is Pareto-frontier of a closed and compact set. All that remains to be shown is that $\Phi(\mathcal{V})$ is p-convex. Let

$$\bar{u} = \Big(\sum_{i=1}^{m} \alpha_i \max_j\{r_1(i,j) + \beta Q_1(i,j)\}, \sum_{i=1}^{m} \alpha_i \max_j\{r_2(i,j) + \beta Q_2(i,j)\}\Big)$$

and

$$\bar{v} = \Big(\sum_{i=1}^{m} \eta_i \max_j\{r_1(i,j) + \beta R_1(i,j)\}, \sum_{i=1}^{m} \eta_i \max_j\{r_2(i,j) + \beta R_2(i,j)\}\Big)$$

be two points in $\Phi(\mathcal{V})$, where $\bar{\alpha}$, $\bar{\eta} \in \Sigma_{A^1}$ and $\bar{Q}(i,j)$, $\bar{R}(i,j) \in \mathcal{V}$. For a fixed $\lambda \in [0,1]$, let $\kappa_i = \alpha_i\lambda + \eta_i(1-\lambda)$. Then

$$\lambda\bar{u} + (1-\lambda)\bar{v} = \Big(\sum_{i=1}^{m} \kappa_i \Big[\frac{\lambda\alpha_i}{\kappa_i} \max_j\{r_1(i,j) + \beta Q_1(i,j)\} + \frac{(1-\lambda)\eta_i}{\kappa_i} \max_j\{r_1(i,j) + \beta R_1(i,j)\}\Big],$$

$$\sum_{i=1}^{m} \kappa_i \Big[\frac{\lambda\alpha_i}{\kappa_i} \max_j\{r_2(i,j) + \beta Q_2(i,j)\} + \frac{(1-\lambda)\eta_i}{\kappa_i} \max_j\{r_2(i,j) + \beta R_2(i,j)\}\Big]\Big)$$

$$\preceq \Big(\sum_{i=1}^{m} \kappa_i \max_j\{r_1(i,j) + \beta\big(\frac{\lambda\alpha_i}{\kappa_i} Q_1(i,j) + \frac{(1-\lambda)\eta_i}{\kappa_i} R_1(i,j)\big)\},$$

$$\sum_{i=1}^{m} \kappa_i \max_j\{r_2(i,j) + \beta\big(\frac{\lambda\alpha_i}{\kappa_i} Q_2(i,j) + \frac{(1-\lambda)\eta_i}{\kappa_i} R_2(i,j)\big)\}\Big)$$

$$\preceq \Big(\sum_{i=1}^{m} \kappa_i \max_j\{r_1(i,j) + \beta S_1(i,j)\},$$

$$\sum_{i=1}^{m} \kappa_i \max_j \{r_2(i,j) + \beta S_2(i,j)\}\Bigg).$$

The first inequality follows since max is a convex function and the second follows since $\mathcal{V}$ is convex, and hence $\bar{S}(i,j) = (S_1(i,j), S_2(i,j)) \in \mathcal{V}$ that satisfy the given relation exist. $\square$

Note that in this case, contrary to the operator for simultaneous moves, it is not true that Alice's choice of points in $\mathcal{V}$ from time $t+1$ onwards needs to only depend on Bob's observed action at time t. This choice must also depend on her realized action at time t. We next show that this operator is a contraction:

Theorem 3.3.1.

$$d(\Phi(\mathcal{U}), \Phi(\mathcal{V})) \leq \beta d(\mathcal{U}, \mathcal{V}). \tag{3.17}$$

Proof. Suppose $e(\mathcal{U}, \mathcal{V}) = \epsilon$. Let $\Bigg(\sum_{i=1}^{m} \alpha_i \max_j \{r_1(i,j) + \beta R_1(i,j)\}, \sum_{i=1}^{m} \alpha_i \max_j \{r_2(i,j) + \beta R_2(i,j)\}\Bigg)$ be some point in $\Phi(\mathcal{V})$, where $\bar{\alpha} \in \Delta A$. Then for each $R(i,j)$, for each i and j, we can choose $R'(i,j) \in \mathcal{U}$ such that $R'(i,j) \preceq R(i,j) + \epsilon \mathbf{1}$. We then have

$$\sum_{i=1}^{m} \alpha_i \max_j \{r_1(i,j) + \beta R_1'(i,j)\} \tag{3.18}$$

$$= \sum_{i=1}^{m} \alpha_i \max_j \{r_1(i,j) + \beta R_1(i,j) + \beta(R'(i,j) - R(i,j))\} \tag{3.19}$$

$$\leq \sum_{i=1}^{m} \alpha_i \max_j \{r_1(i,j) + \beta R_1(i,j) + \beta\epsilon\} \tag{3.20}$$

$$= \sum_{i=1}^{m} \alpha_i \max_j \{r_1(i,j) + \beta R_1(i,j)\} + \beta\epsilon. \tag{3.21}$$

Similarly, we can show that

$$\sum_{i=1}^{m} \alpha_i \max_j \{r_2(i,j) + \beta R_2'(i,j)\} \leq \sum_{i=1}^{m} \alpha_i \max_j \{r_2(i,j) + \beta R_2(i,j)\} + \beta\epsilon. \tag{3.22}$$

Thus

$$\Bigg(\sum_{i=1}^{m} \alpha_i \max_j \{r_1(i,j) + \beta R_1'(i,j)\}, \sum_{i=1}^{m} \alpha_i \max_j \{r_2(i,j) + \beta R_2'(i,j)\}\Bigg) \tag{3.23}$$

$$\preceq \Bigg(\sum_{i=1}^{m} \alpha_i \max_j \{r_1(i,j) + \beta R_1(i,j)\}, \sum_{i=1}^{m} \alpha_i \max_j \{r_2(i,j) + \beta R_2(i,j)\}\Bigg) + \beta\epsilon\mathbf{1}. \tag{3.24}$$

But there exists some $(L_1, L_2) \in \Phi(\mathcal{U})$ such that

$$(L_1, L_2) \preceq \left(\sum_{i=1}^{m} \alpha_i \max_j \{r_1(i,j) + \beta R_1'(i,j)\}, \sum_{i=1}^{m} \alpha_i \max_j \{r_2(i,j) + \beta R_2'(i,j)\} \right) \quad (3.25)$$

Thus

$$(L_1, L_2) \preceq \left(\sum_{i=1}^{m} \alpha_i \max_j \{r_1(i,j) + \beta R_1(i,j)\}, \sum_{i=1}^{m} \alpha_i \max_j \{r_2(i,j) + \beta R_2(i,j)\} \right) + \beta\epsilon \quad (3.26)$$

We can show the other direction (roles of $\Phi(\mathcal{U})$ and $\Phi(\mathcal{V})$ reversed) similarly and thus we have that

$$d(\Phi(\mathcal{U}), \Phi(\mathcal{V})) \leq \beta\epsilon = \beta d(\mathcal{U}, \mathcal{V}). \quad (3.27)$$

□

The following three results immediately follow. We omit the proofs here in order to refrain from reproducing the same arguments.

Theorem 3.3.2. *Let $\mathcal{V} \in \mathcal{F}$. Then the sequence $(\mathcal{A}_n = \Phi^n(\mathcal{V}))_{n\in\mathbb{N}}$ converges in the metric d to a Pareto frontier $\mathcal{V}^* \in \mathcal{F}$, which is the unique fixed point of the operator Φ, i.e., the unique solution of $\Phi(\mathcal{V}) = \mathcal{V}$.*

Theorem 3.3.3. *Let $\mathcal{V}^*$ be the fixed point of the GS operator Φ. Then*

$$\mathcal{V}^* = \Lambda\left(\left\{ \left(\max_{\Phi^2} E\left[\sum_{t=1}^{\infty} \beta^{t-1} r_1(a_t^1, a_t^2) \right], \max_{\bar{\Phi}^2} E\left[\sum_{t=1}^{\infty} \beta^{t-1} r_2(a_t^1, a_t^2) \right] \right) : \Phi^1 \in \Phi^1 \right\}\right). \quad (3.28)$$

Finally, we can derive the structure of the optimal policy. For a Pareto frontier $\mathcal{V} \in \mathcal{F}$, consider a one-to-one function $F^{\mathcal{V}}(p) : \mathcal{P} \to \mathcal{V}$. Assume $\mathcal{V}^*$ is such that $\mathcal{V}^* = \Phi(\mathcal{V}^*)$. Then for a fixed $p \in \mathcal{P}$, choose $\bar{\alpha}(p) \in \Delta(A)$ and a function $q(a, b, p) \in \mathcal{P}$ for each $p \in \mathcal{P}$, $a \in A$ and $b \in B$ that satisfies

$$F_k^{\mathcal{V}^*}(p) = \left\{ \sum_{a=1}^{m} \alpha_a \max_{b\in B} \left[r_k(s,a,b) + \beta F_k^{\mathcal{V}^*}(q(a,b,p)) \right] \right\}.$$

Then the following policy is optimal:

Theorem 3.3.4. *For any $p_0 \in \mathcal{P}$, the pair of upper bounds on losses $x = F^{\mathcal{V}^*}(p_0)$ in $\mathcal{V}^*$ is guaranteed by Alice first choosing action $a_0 \in A$ with probability $\alpha_a(p_0)$. Then if Bob chooses*

an action $b_0 \in B$, the optimal guarantees to choose from the second step on are then $\beta F^{\mathcal{V}^*}(p_1)$ in $\beta\mathcal{V}^*$, where $p_1 = q(a_0, b_0, p_0)$, which can be guaranteed by Alice by choosing action $a_1 \in A$ with probability $\alpha_a(p_1)$, and so on.

Thus in this case, the transitions of the information state not only depend on Bob's actions but on Alice's actions as well.

3.3.2 Case 2: Bob plays first

In this case, adaptive randomized strategy ϕ_B for Bob specifies for each stage t, a mapping from the set of observations till stage t, i.e., $H_t = (b_1, a_1, \cdots, b_{t-1}, a_{t-1})$ to $\Delta(B)$. For Alice, her strategy ϕ_A specifies for each stage t, a mapping from the set of observations till stage t and Bob's action at time t, i.e., $H_t = (b_1, a_1, \cdots, b_{t-1}, a_{t-1}, b_t)$ to $\Delta(B)$. Consider the following operator defined on a frontier in $\mathcal{F}$:

Definition 3.3.2. *For $\mathcal{V} \in \mathcal{F}$, define*

$$\Phi(\mathcal{V}) = \Lambda\Bigg\{\Bigg(\max_b \Bigg\{\sum_{a=1}^{m} \alpha_a(b)\big[r_1(a,b)+\beta R_1(a,b)\big]\Bigg\}, \max_b \Bigg\{\sum_{a=1}^{m} \alpha_a(b)\big[r_2(a,b)+\beta R_2(a,b)\big]\Bigg\}\Bigg)$$

$$: (R_1(a,b), R_2(a,b)) \in \mathcal{V}, \bar{\alpha}(b) \in \Delta(A)\ \forall\ b \in B\Bigg\}. \tag{3.29}$$

The interpretation is again similar. Suppose that $\mathcal{V}$ is the set of guarantees that Alice can achieve from time $t+1$ onwards, then $\Phi(\mathcal{V})$ is the set of guarantees she can achieve beginning from time t. She does so by choosing a randomization over her actions $\bar{\alpha}(b)$ for each possible action b that Bob chooses and then chooses a point in $\mathcal{V}$ depending on actions b and her realized action a. It turns out that analogous to the case of simultaneous moves, her choice of a point in $\mathcal{V}$ need not depend on her won action at time t.

Lemma 3.3.2. *Suppose that $\mathcal{V} \in F$. Then*

1. *Any point $\bar{u} \in \Phi(\mathcal{V})$ is of the form*

$$\bar{u} = \Bigg(\max_b \Bigg\{\sum_{a=1}^{m} \alpha_a(b) r_1(a,b) + \beta Q_1(b)\Bigg\}, \max_b \Bigg\{\sum_{a=1}^{m} \alpha_a(b) r_2(a,b) + \beta Q_2(b)\Bigg\}\Bigg)$$

where $Q(b) \in \mathcal{V}$ for each $b \in B$.

Proof. The first part follows from the fact that, by p-convexity of $\mathcal{V}$, there exists a $Q(b) \in \mathcal{V}$ such that $Q(b) \preceq \sum_{a\in A} \alpha_a(b)R(a,b)$. For the second claim, since $\mathcal{V} \in \mathcal{F}$, we know from the closed map lemma that $\Phi(\mathcal{V})$ is the Pareto frontier of a compact set, and thus all that is needed to be shown is that $\Phi(\mathcal{V})$ is p-convex. Let

$$\bar{u} = \left(\max_b \left\{ \sum_{a=1}^{m} \alpha_a(b)\big[r_1(a,b)\big] + \beta Q_1(b) \right\}, \max_b \left\{ \sum_{a=1}^{m} \alpha_a(b)\big[r_2(a,b)\big] + \beta Q_2(b) \right\} \right)$$

and

$$\bar{v} = \left(\max_b \left\{ \sum_{a=1}^{m} \eta_a(b)\big[r_1(a,b)\big] + \beta R_1(b) \right\}, \max_b \left\{ \sum_{a=1}^{m} \eta_a(b)\big[r_2(a,b)\big] + \beta R_2(b) \right\} \right)$$

be two points in $\Phi(\mathcal{V})$, where $\bar{\alpha}(b)$, $\bar{\eta}(b) \in \Delta(A)$ and $Q(b)$, $R(b) \in \mathcal{V}$ for all $b \in B$.

For a fixed $\lambda \in [0,1]$, let $\kappa_a(b) = \alpha_a(b)\lambda + \eta_a(b)(1-\lambda)$. Then

$$\lambda\bar{u} + (1-\lambda)\bar{v} = \left(\lambda \max_b \left\{ \sum_{a=1}^{m} \alpha_a(b)\big[r_1(a,b)\big] + \beta Q_1(b) \right\} \right.$$

$$+(1-\lambda) \max_b \left\{ \sum_{a=1}^{m} \eta_a(b)\big[r_1(a,b)\big] + \beta R_1(b) \right\},$$

$$\left. \lambda \max_b \left\{ \sum_{a=1}^{m} \alpha_a(b)\big[r_2(a,b)\big] + \beta Q_2(b) \right\} + (1-\lambda) \max_b \left\{ \sum_{a=1}^{m} \eta_a(b)\big[r_2(a,b)\big] + \beta R_2(b) \right\} \right)$$

$$\succeq \left(\max_b \left\{ \sum_{a=1}^{m} \kappa_a(b)\big[r_1(a,b)\big] + \beta\lambda Q_1(b) + (1-\lambda)R_1(b) \right\}, \right.$$

$$\left. \max_b \left\{ \sum_{a=1}^{m} \kappa_a(b)\big[r_2(a,b)\big] + \beta\lambda Q_2(b) + (1-\lambda)R_2(b) \right\} \right)$$

$$\succeq \left(\max_b \left\{ \sum_{a=1}^{m} \kappa_a(b)\big[r_1(a,b)\big] + \beta L_1(b) \right\}, \max_b \left\{ \sum_{a=1}^{m} \kappa_a(b)\big[r_2(a,b)\big] + \beta L_2(b) \right\} \right)$$

The first inequality holds since max is a convex function and the second follows since $\mathcal{V}$ is p-convex, and hence $L(b) = (L_1(b), L_2(b)) \in \mathcal{V}$ that satisfy the given relation exist. Thus $\Phi(\mathcal{V})$ is p-convex. $\square$

Next we can show that the operator is a contraction in the metric d. We omit the proof since it is almost identical to the proof of Theorem 2.2.1.

Theorem 3.3.5.

$$d(\Phi(\mathcal{U}), \Phi(\mathcal{V})) \leq \beta d(\mathcal{U}, \mathcal{V}). \tag{3.30}$$

The following three results immediately follow. Again we omit the proofs in order to refrain from producing the same arguments.

Theorem 3.3.6. *Let $\mathcal{V} \in \mathcal{F}$. Then the sequence $(\mathcal{A}_n = \Phi^n(\mathcal{V}))_{n\in\mathbb{N}}$ converges in the metric d to a Pareto frontier $\mathcal{V}^* \in \mathcal{F}$, which is the unique fixed point of the operator Φ, i.e., the unique solution of $\Phi(\mathcal{V}) = \mathcal{V}$.*

Theorem 3.3.7. *Let $\mathcal{V}^*$ be the fixed point of the GS operator Φ. Then*

$$\mathcal{V}^* = \Lambda(\left\{ \left(\max_{\Phi^2} E\left[\sum_{t=1}^{\infty} \beta^{t-1} r_1(a_t^1, a_t^2) \right], \max_{\bar{\Phi}^2} E\left[\sum_{t=1}^{\infty} \beta^{t-1} r_2(a_t^1, a_t^2) \right] \right) : \Phi^1 \in \Phi^1 \right\}). \tag{3.31}$$

Finally, we can derive the structure of the optimal policy. For a Pareto frontier $\mathcal{V} \in \mathcal{F}$, consider a one-to-one function $F^{\mathcal{V}}(p) : \mathcal{P} \to \mathcal{V}$. Assume $\mathcal{V}^*$ is such that $\mathcal{V}^* = \Phi(\mathcal{V}^*)$. Then for a $p \in \mathcal{P}$ and $b \in B$, choose $\bar{\alpha}(b, p) \in \Delta(A)$ and a function $q(b, p) \in \mathcal{P}$ for each $p \in \mathcal{P}$ and $b \in B$ that satisfies

$$F_k^{\mathcal{V}^*}(p) = \max_{b\in B} \left\{ \sum_{a=1}^{m} \alpha_a(b, p) r_k(s, a, b) + \beta F_k^{\mathcal{V}^*}(q(b, p)) \right\}.$$

Then the following policy is optimal:

Theorem 3.3.8. *For any $p_0 \in \mathcal{P}$, the pair of upper bounds on losses $x = F^{\mathcal{V}^*}(p_0)$ in $\mathcal{V}^*$ is guaranteed by Alice by first choosing action $a_0 \in A$ with probability $\alpha_a(b_0, p_0)$, after having observed Bob's action $b_0 \in B$. Then the optimal guarantees to choose from the second step on are $\beta F^{\mathcal{V}^*}(p_1)$ in $\beta\mathcal{V}^*$, where $p_1 = q(b_0, p_0)$, which can be guaranteed by Alice by choosing action $a_1 \in A$ with probability $\alpha_a(b_1, p_1)$ after having observed Bob's action b_1, and so on.*

Chapter 4

Generalizations to games with multiple states

4.1 Stochastic games with vector losses

We define a vector-valued generalization of a stochastic game $\mathbb{G}$ with a finite state space $S = \{1, \cdots, \Omega\}$. These games were first introduced by Shapley for the scalar case in [29]. Each state s is associated with action spaces $A^s = \{1, \cdots, m^s\}$ and $B^s = \{1, \cdots, n^s\}$ for Alice and Bob respectively. The game is played in stages $t = 1, 2, \cdots$, starting from some state $s_1 \in S$. In each stage, Alice and Bob simultaneously play one of the actions that are available to them in the current state, as a result of which they get a vector of losses and the state of the game probabilistically transitions into another state in the next stage. Suppose the game is in some state s, then for the pair of actions $a \in A^s$ and $b \in B^s$, the vector of instantaneous losses is given by $\mathbf{r}(s, a, b) \in \mathbb{R}^l$, and let $p(s'|s, a, b)$ denote the probability that the next state is s' for each $s' \in S$. For simplicity, we assume that the vector of instantaneous payoffs has only two components, i.e. $\mathbf{r}(s, a, b) = (r_1(s, a, b), r_1(s, a, b))$.

Let s_t denote the state, and let a_t and b_t denote the actions of Alice and Bob at time t. Fix a discount factor $\beta \in [0, 1)$. Then the vector of total discounted losses is defined as:

$$\sum_{t=1}^{\infty} \beta^{t-1} \mathbf{r}(s_t, a_t, b_t) = \left(\sum_{t=1}^{\infty} \beta^{t-1} r_1(s_t, a_t, b_t), \sum_{t=1}^{\infty} \beta^{t-1} r_2(s_t, a_t, b_t) \right). \tag{4.1}$$

An adaptive randomized strategy ϕ_A for Alice specifies for each stage t, a mapping from the set of observations till stage t, i.e., $H_t = (s_1, a_1, b_1, \cdots, s_{t-1}, a_{t-1}, b_{t-1}, s_t)$, to a probability distribution on the action set A^{s_t}, denoted by $\Delta(A^{s_t})$. Let Φ_A be the set of all such policies of

Alice. Similarly, let Φ_B be the set of all adaptive randomized strategies for Bob. Beginning with a state $s_1 = s$, for a pair of strategies ϕ_A and ϕ_B, the expected discounted loss on component k in the stochastic game is given by:

$$R_k(s, \phi_A, \phi_B) = E_{s,\phi_A,\phi_B}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_k(s_t, a_t, b_t)\right], \tag{4.2}$$

where the expectation is over the randomness in the strategies ϕ_A and ϕ_B, and in the state transitions. Now consider a fixed policy $\phi_A \in \Phi_A$ and a fixed initial state $s \in S$. If Alice plays this strategy, then irrespective of the strategy chosen by Bob, Alice guarantees that the long term expected losses on different components lie in the '*corner* set' :

$$\mathcal{C}(s, \phi_A) \triangleq \left\{\bar{x} \in \mathbb{R}^m : x_k \leq \max_{\phi_B^k \in \Phi_B} R_k(s, \phi_A, \phi_B^k) \text{ for all } k \in \{1, 2\}\right\}$$

defined by the corner point

$$\left(\max_{\phi_B^1 \in \Phi^B} R_1(s, \phi_A, \phi_B^1), \max_{\phi_B^2 \in \Phi^B} R_2(s, \phi_A, \phi_B^2)\right).$$

For a game starting from state s, let the set of all the corner points, or *simultaneous guarantees* that correspond to *all* the strategies $\phi_A \in \Phi_A$ be defined as:

$$\mathcal{W}_s \triangleq \left\{\left(\max_{\phi_B^1 \in \Phi^B} R_1(s, \phi_A, \phi_B^1), \max_{\phi_B^2 \in \Phi^B} R_2(s, \phi_A, \phi_B^2)\right) : \phi_A \in \Phi_A\right\}. \tag{4.3}$$

Our objective is to characterize and compute the *minimal* points in the set $\mathcal{W}_s$, i.e., its *Lower Pareto frontier*

$$\mathcal{U}_s^* = \Lambda(\mathcal{W}_s) \tag{4.4}$$

for each $s \in S$, and compute policies for Alice in Φ_A that guarantee different points in this set.

4.1.1 The Generalized Shapley operator

As in chapter 2, let $\mathcal{F}$ be the space of Pareto frontiers of convex and closed subsets of $[0, 1]^2$ endowed with the metric d, which is defined as the Hausdorff distance between the upsets of the frontiers.

Consider the space $\mathcal{F}^\Omega$ with each element of the form $\overline{\mathcal{V}} = (\mathcal{V}_s)_{s \in S}$ and define

$$\bar{d}(\overline{\mathcal{V}}, \overline{\mathcal{U}}) = \max_{s \in S} d(\mathcal{U}_s, \mathcal{V}_s). \tag{4.5}$$

Since $\mathcal{F}$ is complete in metric d, it is clear that $\mathcal{F}^\Omega$ is complete in the metric $\bar{d}$.

Now again assume that by scaling and shifting the losses $r_k(s,a,b) \in [0, 1-\beta]$ for all $s \in S$, $a \in A^s$, $b \in B^s$ and $k = 1, 2$. Thus the total discounted losses of the game lie in the set $[0,1]^2$. Now consider a collection of sets $\big(\mathcal{S}_s\big)_{s\in S}$ in $[0,1]^2$, one associated with each state $s \in S$. We will denote this collection by $\overline{\mathcal{S}}$. Define the following operator on the space of such collections of sets in $[0,1]^2$.

$$\begin{aligned}\Psi(\overline{\mathcal{S}})_s \;=\; \Bigg\{\Bigg(& \max_b \Bigg\{ \sum_{a=1}^m \alpha_a \big[r_1(s,a,b) + \beta \sum_{s'\in S} p(s'|s,a,b) R_1(s',a,b)\big]\Bigg\}, \\ & \max_b \Bigg\{ \sum_{a=1}^m \alpha_a \big[r_2(s,a,b) + \beta \sum_{s'\in S} p(s'|s,a,b) R_2(s',a,b)\big]\Bigg\}\Bigg) \\ & : (R_1(s',a,b), R_2(s',a,b)) \in \mathcal{S}_{s'}, \bar{\alpha} \in \Delta(A^s)\Bigg\}. \end{aligned} \tag{4.6}$$

This operator can be interpreted in the following way. Suppose that $\mathcal{S}_{s'}$ is the set of simultaneous guarantees that Alice can achieve starting from state s' at time $t+1$ for all $s' \in S$. Then $\Psi(\overline{\mathcal{S}})_s$ is the set of simultaneous guarantees that she can achieve starting from state s at time t. A point in this set corresponds to Alice first choosing a probability distribution over her available actions at time t in state s, and then choosing a mapping from 1) the action of Bob at time t and 2) the realized state s' at time $t+1$, to a point in $\mathcal{S}_{s'}$. We will next prove that if each $\mathcal{S}_s$ for $s \in S$ is closed, the each $\Psi(\overline{\mathcal{S}})_s$ is closed as well. Further, if each $\mathcal{S}_s$ is convex, then the Pareto frontier $\Lambda(\Psi(\overline{\mathcal{S}})_s)$ is the Pareto frontier of some closed convex set.

Lemma 4.1.1. *Let $\mathcal{S}_s \subseteq [0,1]^2$ for each $s \in S$ be closed sets. Then $\Psi(\overline{\mathcal{S}})_s \subseteq [0,1]^2$ is closed for each $s \in S$. If in addition, $\mathcal{S}_s$ is convex for each $s \in S$, then:*

1. *Any point $\bar{u}$ in $\Lambda(\Psi(\overline{\mathcal{S}})_s)$ is of the form:*

$$\Bigg(\max_b \Bigg\{ \sum_{a=1}^m \alpha_a r_1(s,a,b) + \beta \sum_{s'\in S} \big(\sum_{a=1}^m \alpha_a p(s'|s,a,b)\big) Q_1(s',b)\Bigg\},$$

$$\max_b \Bigg\{ \sum_{a=1}^m \alpha_a r_2(s,a,b) + \beta \sum_{s'\in S} \big(\sum_{a=1}^m \alpha_a p(s'|s,a,b)\big) Q_2(s',b)\Bigg\}\Bigg)$$

   for some $Q(s',b) \in \Lambda(\mathcal{S}_{s'})$, for each $s' \in S$ and $b \in B$.

2. $\Lambda(\Psi(\overline{\mathcal{S}})_s) \in \mathcal{F}$.

Proof. The first claim of the Lemma follows from the closed map lemma as in the proof of Lemma 2.2.3. Now assume that $\mathcal{S}_s$ is a closed convex set for each $s \in S$. Then $\Lambda(\mathcal{S}_s)$

exists by Lemma 3.1 and further it is p-convex by Lemma 2.2.5. Let $\mathcal{U}_s = \Lambda(\mathcal{S}_s)$. Clearly, $\Lambda(\Psi(\overline{\mathcal{S}})_s) = \Lambda(\Psi(\overline{\mathcal{U}})_s)$ where $\overline{\mathcal{U}} = (\mathcal{U}_s)_{s\in S}$. Recall that any point $\bar{u}$ in $\Lambda(\Psi(\overline{\mathcal{U}})_s)$ is of the form:

$$\Big(\max_b \Big\{ \sum_{a=1}^m \alpha_a \big[r_1(s,a,b) + \beta \sum_{s'\in S} p(s'|s,a,b) R_1(s',a,b)\big] \Big\},$$
$$\max_b \Big\{ \sum_{a=1}^m \alpha_a \big[r_2(s,a,b) + \beta \sum_{s'\in S} p(s'|s,a,b) R_2(s',a,b)\big] \Big\}\Big)$$

for some $\bar{\alpha} \in \Delta(A^s)$ and $R(s',a,b) \in \mathcal{U}_{s'}$. But since $\mathcal{U}_{s'}$ is p-convex, for each $b \in B^s$, there exists some $Q(s',b) \in \mathcal{U}_{s'}$ such that $Q(s',b) \preceq \frac{\sum_{a=1}^m \alpha_a p(s'|s,a,b) R(s',a,b)}{\sum_{a=1}^m \alpha_a p(s'|s,a,b)}$. Hence statement 1 follows.

Now let

$$\bar{u} = \Big(\max_b \Big\{ \sum_{a=1}^m \alpha_a r_1(s,a,b) + \beta \sum_{s'\in S} \big(\sum_{a=1}^m \alpha_a p(s'|s,a,b)\big) Q_1(s',b) \Big\},$$
$$\max_b \Big\{ \sum_{a=1}^m \alpha_a r_2(s,a,b) + \beta \sum_{s'\in S} \big(\sum_{a=1}^m \alpha_a p(s'|s,a,b)\big) Q_2(s',b) \Big\}\Big)$$

and let

$$\bar{v} = \Big(\max_b \Big\{ \sum_{a=1}^m \eta_a r_1(s,a,b) + \beta \sum_{s'\in S} \big(\sum_{a=1}^m \eta_a p(s'|s,a,b)\big) R_1(s',b) \Big\},$$
$$\max_b \Big\{ \sum_{a=1}^m \eta_a r_2(s,a,b) + \beta \sum_{s'\in S} \big(\sum_{a=1}^m \eta_a p(s'|s,a,b)\big) R_2(s',b) \Big\}\Big)$$

for some $\bar{\alpha}, \bar{\eta} \in \Delta(A^s)$ and $Q(s',b), R(s',b) \in \mathcal{U}_{s'}$ for each $s' \in S$ and $b \in B$. Now for a fixed $\lambda \in [0,1]$, let $\kappa_a = \lambda\alpha_a + (1-\lambda)\eta_a$.

Then we have

$$\lambda\bar{u} + (1-\lambda)\bar{v}$$
$$= \Big(\lambda \max_b \Big\{ \sum_{a=1}^m \alpha_a r_1(s,a,b) + \beta \sum_{s'\in S} \big(\sum_{a=1}^m \alpha_a p(s'|s,a,b)\big) Q_1(s',b) \Big\}$$
$$+(1-\lambda) \max_b \Big\{ \sum_{a=1}^m \eta_a r_1(s,a,b) + \beta \sum_{s'\in S} \big(\sum_{a=1}^m \eta_a p(s'|s,a,b)\big) R_1(s',b) \Big\},$$
$$\lambda \max_b \Big\{ \sum_{a=1}^m \alpha_a r_2(s,a,b) + \beta \sum_{s'\in S} \big(\sum_{a=1}^m \alpha_a p(s'|s,a,b)\big) Q_2(s',b) \Big\}$$

$$+(1-\lambda)\max_b\Big\{\sum_{a=1}^m \eta_a r_2(s,a,b)+\beta\sum_{s'\in S}\big(\sum_{a=1}^m \eta_a p(s'|s,a,b)\big)R_2(s',b)\Big\}\Big)$$

$$\succeq\Big(\max_b\Big\{\sum_{a=1}^m \kappa_a r_1(s,a,b)+\beta\sum_{s'\in S}\big(\sum_{a=1}^m \kappa_a p(s'|s,a,b)\big)$$

$$\times\frac{\lambda(\sum_{a=1}^m \alpha_a p(s'|s,a,b))Q_1(s',b)+(1-\lambda)(\sum_{a=1}^m \eta_a p(s'|s,a,b))R_1(s',b)}{\sum_{a=1}^m \kappa_a p(s'|s,a,b)}\Big\},$$

$$\max_b\Big\{\sum_{a=1}^m \kappa_a r_2(s,a,b)+\beta\sum_{s'\in S}\big(\sum_{a=1}^m \kappa_a p(s'|s,a,b)\big)$$

$$\times\frac{\lambda(\sum_{a=1}^m \alpha_a p(s'|s,a,b))Q_2(s',b)+(1-\lambda)(\sum_{a=1}^m \eta_a p(s'|s,a,b))R_2(s',b)}{\sum_{a=1}^m \kappa_a p(s'|s,a,b)}\Big\}\Big)$$

$$\succeq\Big(\max_b\Big\{\sum_{a=1}^m \kappa_a r_1(s,a,b)+\beta\sum_{s'\in S}\big(\sum_{a=1}^m \kappa_a p(s'|s,a,b)\big)L_1(s',b)\Big\},$$

$$\max_b\Big\{\sum_{a=1}^m \kappa_a r_2(s,a,b)+\beta\sum_{s'\in S}\big(\sum_{a=1}^m \kappa_a p(s'|s,a,b)\big)L_2(s',b)\Big\}\Big).$$

The first inequality holds since max is a convex function and second holds since $\mathcal{U}_s$ is p-convex, and thus $L(s',b)=(L_1(s',b),L_2(s',b))$ that satisfy the given relation exist. Thus $\Lambda(\Psi(\overline{\mathcal{U}})_s)$ is p-convex. And thus by Lemma 2.2.5, it is the Pareto frontier of a closed convex set, i.e., it is in $\mathcal{F}$. □

We now define the following Generalized Shapley operator on the space $\mathcal{F}^\Omega$.

Definition 4.1.1. *For $\overline{\mathcal{V}}=\big(\mathcal{V}_s\big)_{s\in S}$ where $\mathcal{V}_s\in\mathcal{F}$, define $\Phi(\overline{\mathcal{V}})_s=\Lambda(\Psi(\overline{\mathcal{V}})_s)$.*

From Lemma 4.1.1 we know that $\Phi(\overline{\mathcal{V}})$ is in $\mathcal{F}^\Omega$ if $\overline{\mathcal{V}}$ is in $\mathcal{F}^\Omega$. Next we show that the operator Φ is a contraction in the metric $\bar{d}$.

Lemma 4.1.2.

$$\bar{d}(\Phi(\overline{\mathcal{U}}),\Phi(\overline{\mathcal{V}}))\le\beta\bar{d}(\overline{\mathcal{U}},\overline{\mathcal{V}}).$$

Proof. Suppose that $\bar{d}(\overline{\mathcal{U}},\overline{\mathcal{V}})=\epsilon$. By Lemma 2.2.6, this means that $e(\mathcal{U}_s,\mathcal{V}_s)\le\epsilon$ for each $s\in S$. Now let

$$\bar{u}=\Big(\max_b\Big\{\sum_{a=1}^m \alpha_a r_1(s,a,b)+\beta\sum_{s'\in S}\big(\sum_{a=1}^m \alpha_a p(s'|s,a,b)\big)R_1(s',b)\Big\},$$

$$\max_b \left\{ \sum_{a=1}^{m} \alpha_a r_2(s,a,b) + \beta \sum_{s' \in S} \big(\sum_{a=1}^{m} \alpha_a p(s'|s,a,b) \big) R_2(s',b) \right\} \Big)$$

be some point in $\Phi(\overline{\mathcal{U}})_s$, where $\bar{\alpha} \in \Delta(A^s)$ and $R(s',b) \in \mathcal{U}_{s'}$. Now we can choose some point $Q(s',b) \in \mathcal{V}_{s'}$ for each $s' \in S$ and $b \in B^s$, such that $Q(s',b) \preceq R(s',b) + \epsilon \mathbf{1}$. This in turn implies that

$$\Big(\max_b \left\{ \sum_{a=1}^{m} \alpha_a r_1(s,a,b) + \beta \sum_{s' \in S} \big(\sum_{a=1}^{m} \alpha_a p(s'|s,a,b) \big) Q_1(s',b) \right\},$$

$$\max_b \left\{ \sum_{a=1}^{m} \alpha_a r_2(s,a,b) + \beta \sum_{s' \in S} \big(\sum_{a=1}^{m} \alpha_a p(s'|s,a,b) \big) Q_2(s',b) \right\} \Big)$$

$$\preceq \Big(\max_b \left\{ \sum_{a=1}^{m} \alpha_a r_1(s,a,b) + \beta \sum_{s' \in S} \big(\sum_{a=1}^{m} \alpha_a p(s'|s,a,b) \big) R_1(s',b) \right\},$$

$$\max_b \left\{ \sum_{a=1}^{m} \alpha_a r_2(s,a,b) + \beta \sum_{s' \in S} \big(\sum_{a=1}^{m} \alpha_a p(s'|s,a,b) \big) R_2(s',b) \right\} \Big) + \beta \epsilon \mathbf{1}.$$

But there exists some $(L_1, L_2) \in \Lambda(\Psi(\overline{\mathcal{U}})_s)$ such that

$$(L_1, L_2) \preceq \Big(\max_b \left\{ \sum_{a=1}^{m} \alpha_a r_1(s,a,b) + \beta \sum_{s' \in S} \big(\sum_{a=1}^{m} \alpha_a p(s'|s,a,b) \big) Q_1(s',b) \right\},$$

$$\max_b \left\{ \sum_{a=1}^{m} \alpha_a r_2(s,a,b) + \beta \sum_{s' \in S} \big(\sum_{a=1}^{m} \alpha_a p(s'|s,a,b) \big) Q_2(s',b) \right\} \Big)$$

. We can similarly show the other direction with the roles of $\Phi(\overline{\mathcal{U}})_s$ and $\Phi(\overline{\mathcal{V}})_s$ reversed. Thus

$$e(\Phi(\overline{\mathcal{U}})_s, \Phi(\overline{\mathcal{V}})_s) \leq \beta \epsilon$$

and thus $\bar{d}(\Phi(\overline{\mathcal{U}}), \Phi(\overline{\mathcal{V}})) \leq \beta \bar{d}(\overline{\mathcal{U}}, \overline{\mathcal{V}})$. □

We finally show that the GS operator has a unique fixed point and starting from any initial point in $\mathcal{F}^{\Omega}$, the sequence of sets of Pareto frontiers obtained by repeated application of this operator converges to this fixed point.

Theorem 4.1.1. *Let $\overline{\mathcal{V}} \in \mathcal{F}^{\Omega}$. Then the sequence $(\overline{\mathcal{A}}^n = \Phi^n(\overline{\mathcal{V}}))_{n \in \mathbb{N}}$ converges in the metric $\bar{d}$ to a Pareto frontier $\overline{\mathcal{V}}^* \in \mathcal{F}^{\Omega}$, which is the unique fixed point of the operator Φ, i.e., the unique solution of $\Phi(\overline{\mathcal{V}}) = \overline{\mathcal{V}}$.*

Proof. Since Φ is a contraction in the metric $\bar{d}$, the sequence $\{\overline{\mathcal{A}}_n\}$ is Cauchy in $\mathcal{F}^{\Omega}$. Hence by Lemma 2.2.2, $\{\overline{\mathcal{A}}_n\}$ converges to a Pareto frontier $\overline{\mathcal{V}}^* \in \mathcal{F}^{\Omega}$. The continuity of the operator further implies that

$$\overline{\mathcal{V}}^* = \Phi(\overline{\mathcal{V}}^*).$$

To show uniqueness, observe that if there are two fixed points $\overline{\mathcal{U}}$ and $\overline{\mathcal{V}}$, then we have

$$\bar{d}(\overline{\mathcal{U}}, \overline{\mathcal{V}}) = \bar{d}(\Phi(\overline{\mathcal{U}}, \Phi(\overline{\mathcal{V}})) \leq \beta \bar{d}(\overline{\mathcal{U}}, \overline{\mathcal{V}}),$$

which implies that $\bar{d}(\overline{\mathcal{U}}, \overline{\mathcal{V}}) = 0$ and hence $\overline{\mathcal{U}} = \overline{\mathcal{V}}$. □

Following two theorems are immediate. We omit their proofs since they are very similar to those for the case of repeated games. First we show that the fixed point $\overline{\mathcal{V}}^*$ of the GS operator Φ is exactly the set of sets $(\mathcal{U}_s)_{s \in S}$ defined in equation 4.4.

Theorem 4.1.2. $\overline{\mathcal{U}}^* = \overline{\mathcal{V}}^*$.

The next result gives the structure of the optimal policy. For a Pareto frontier $\mathcal{V} \in \mathcal{F}$, consider a one-to-one function $F^{\mathcal{V}}(p) : \mathcal{P} \to \mathcal{V}$. Assume $\overline{\mathcal{V}}^*$ is such that $\overline{\mathcal{V}}^* = \Phi(\overline{\mathcal{V}}^*)$. Then for a fixed $s \in S$ and $p \in \mathcal{P}$, choose $\bar{\alpha}^s(p) \in \Delta(A^s)$ and a function $q(s', b, p.s) \in \mathcal{P}$ for each $s, s' \in S$, $p \in \mathcal{P}$ and $b \in B^s$ that satisfies

$$F_k^{\mathcal{V}_s^*}(p) = \max_{b \in B^s} \left\{ \sum_{a=1}^{m} \alpha_a^s r_k(s, a, b) + \beta \sum_{s' \in S} \big(\sum_{a=1}^{m} \alpha_a^s p(s'|s, a, b) \big) F_k^{\mathcal{V}_{s'}^*}(q(s', b, p, s)) \right\}$$

Then we have the following result:

Theorem 4.1.3. *For any starting state $s_0 \in S$, for any $p_0 \in \mathcal{P}$, the pair of rewards $x = F^{\mathcal{V}_{s_0}^*}(p_0)$ in $\mathcal{V}_{s_0}^*$ is guaranteed by Alice first choosing action $a_0 \in A^{s_0}$ with probability $\alpha_{a_0}^{s_0}(p_0)$. Then if Bob chooses an action $b_0 \in B$ and the state transitions into another state s_1, then the optimal guarantees to choose from the second stage onwards, beginning from the new state s_1 are then $\beta F^{\mathcal{V}_{s_1}^*}(p_1)$ in $\beta \mathcal{V}_{s_1}^*$, where $p_1 = q(s_1, b_0, p_0, s_0)$, which can be guaranteed by Alice by choosing action $a_1 \in A^{s_1}$ with probability $\alpha_{a_1}^{s_1}(p_1)$, and so on.*

4.2 Maximal guarantees in Reachability games

Several types of pursuit-evasion games that arise in defense operations can be modeled as a reachability game (see [12]) after discretization. In a reachability game, two players Alice and Bob take turns to (deterministically) move the state of the system within a finite set of states, governed by a set of specified rules. Given an initial state, a set of states is reachable if Bob has a strategy that guarantees that the state enters this set in a finite time. Similarly, a set of states is excludable if Alice has a strategy that guarantees that the state never reaches this set. An excludable set is maximal if any larger set that contains it is reachable. In this chapter, we present an efficient algorithm to compute the maximal excludable subsets of any specified set of target states and the corresponding strategies. This objective is practically motivated: for example, if in some scenario a defender is unable to protect all of a given set of sensitive targets from an attacker, then being able to efficiently compute the largest subsets of targets that it can simultaneously protect can be of critical importance. The core idea is similar to the approaches in the previous chapters. We define an appropriate dynamic programming operator on the space of maximal subsets of a finite set, which can be seen as the abstract space of 'Pareto frontiers' of collections of subsets of a finite set.

4.2.1 Model

We consider two-person reachability games of alternating moves on a finite state space S. From each state $s \in S$, there is a set of one-step reachable states for each of the two players and the players take turns to (deterministically) move the state of the system according to a given set of rules. One can think of the states as representing nodes in a graph, and each player has a different set of directed edges in this graph that capture the transitions that are allowed for the player from the different states. Figure 4.1 shows a reachability game with the state space $\{A, B, C, D, E\}$, with the two sets of directed edges, one for Alice and one for Bob. For example, from state D, Alice can move the system to state C or E, but Bob can only move it to state C. The players take turns to move. So suppose that the system is in state D and Alice has the first move, and suppose she moves the state to C, then in the next step it is Bob's turn to move and he can move the state to either A or E, and then Alice moves again, and so on.

Given an initial state $s \in S$, and the specification of the player who moves first, a subset $T \subseteq S$ of states is reachable if Bob has a policy that guarantees that the state enters this subset in a finite time. Similarly, T is excludable if Alice has a policy that guarantees that the state never reaches this set. An excludable set is maximal if any other set that contains it (i.e., a strict superset) is reachable by Bob. We focus on the computation of the maximal excludable subsets of a given set $GOAL \subset S$ from an arbitrary initial state.

At any given stage, we can define the 'state' of this dynamic game to be the pair that includes the state of the system and the player with the next move. In order to avoid confusion, we will refer to the state of the system as the 'position'. The state space is

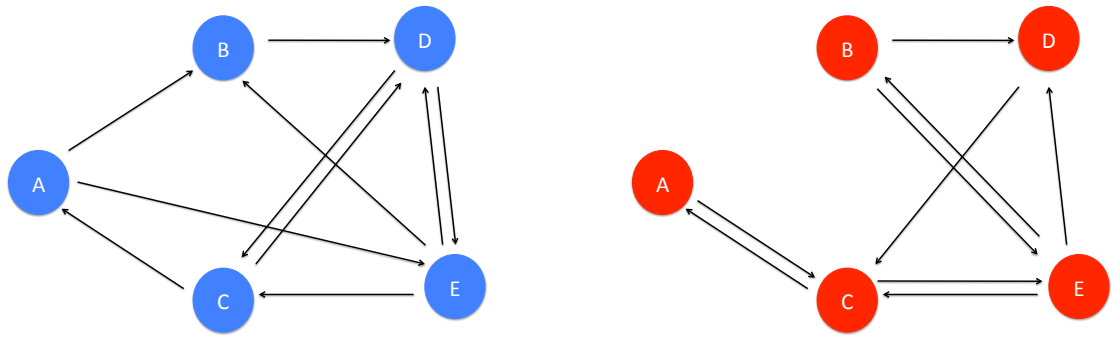

Figure 4.1. The directed graph showing allowed state transitions for Alice (left) and Bob (right)

denoted as $W = S \times \{alice, bob\}$, with any $w \in W$ of the form (s, j) where $s \in S$ is the position of the system and $j \in \{alice, bob\}$ specifies the player.

One can show that starting from any initial state in W, any subset of S is either reachable or excludable. Thus a given subset of S partitions the state space of the dynamic game into two sets: the initial states from which this subset is reachable and those from which it is excludable. Further this partition can be computed efficiently using a simple backward inductive algorithm. For each position $s \in S$, the set of one-step reachable positions from s by player j is denoted by $R(s, j) = R(w)$. Similarly, $L(w)$ denotes the set of positions from where s is one-step reachable by player j, where $w = (s, j) \in W$. For a given set $T \subseteq S$, Algorithm 1 computes the set of initial states from which T is reachable.

To see why the algorithm converges till we have $V_k(w) = V_{k-1}(w)$ for each $w \in W$, observe that if $V_k(w) = 0$ for some k and w, then $V_{k'}(w) = 0$ for all $k' > k$. Thus since the state space is finite the algorithm must converge. It is straightforward to see that $\{w : V^*(w) = 1\}$ are the states from where T is excludable and $\{w : V^{T*}(w) = 0\}$ are the states from where T is reachable.

Now one straightforward way to compute the maximal excludable subsets of $GOAL$ is to consider all the possible subsets of $GOAL$ and compute the partition for each one of them. But this requires solving a prohibitively high number of instances. The question is if one can do better. Intuitively, answering the reachability question for each possible subset of $GOAL$ has severe redundancies: all the information is captured by the maximal excludable sets of $GOAL$, which, in most cases would be a much smaller object to deal with. So can we perhaps come up with an algorithm similar to the one above, that inductively performs appropriate manipulations of these maximal excludable sets in every iteration?

Algorithm 1: Computing the set of states from where T is reachable.

- **Initialize:** For each $w \in W$, set $V_0^T(w) = 1$.
- **Do:** For $k > 0$, for each $w = (s, j) \in W$,
  - If $s \in T$, then $V_k^T(w) = 0$,
  - Otherwise, if $j = B$, then

$$V_k(w) = \min_{s' \in R(w)} V_{k-1}^T((s', A)),$$

  - Else if $j = A$, then

$$V_k(w) = \max_{s' \in R(w)} V_{k-1}^T((s', B)).$$

- **Until:** $V_k^T(w) = V_{k-1}^T(w) = V^{T*}(w)$ for each $w \in W$.
- $\{w : V^{T*}(w) = 1\}$ are the states from where T is excludable and $\{w : V^{T*}(w) = 0\}$ are the states from where T is reachable.

4.2.2 One-step optimality conditions

Let $\mathcal{G}$ be the power set of $GOAL$. Consider a finite collection $\mathcal{U}$ of sets in $\mathcal{G}$. This collection is called maximal if for no two $A, A' \in \mathcal{U}$, $A \subset A'$ (a collection with a single set is maximal). Let $\mathcal{F}$ be the space of maximal collections of elements in $\mathcal{G}$. Let $|GOAL| = G$.

First, we define an operation that takes a collection of subsets in $\mathcal{G}$ and outputs the largest maximal sub-collection. This is analogous to the operation of extracting the Pareto frontier of a set in the real vector space.

Definition 4.2.1. *For a finite collection of sets $\mathcal{U} = \{A_1, \cdots, A_N\}$ in $\mathcal{G}$, define*

$$\Lambda(\mathcal{U}) = \{A \in \mathcal{U} : \forall B \neq A \in \mathcal{U}, A \setminus B \neq \phi\}.$$

Note that $\Lambda \in \mathcal{F}$. For example let $\mathcal{U} = \big\{\{A, B, C\}, \{B\}, \{C, D\}, \{A, C, D\}, \{B, D\}\big\}$. Then $\Lambda(\mathcal{U}) = \big\{\{A, B, C\}, \{A, C, D\}, \{B, D\}\big\}$. This operation can be performed in $O(N^2G)$ time using a simple algorithm: for each set, one can iterate through all the other sets to check if it is a subset (checking whether a set is a subset of the other takes at the most $O(G)$ time assuming the elements are ordered).

Our approach is going to be the following. For every state w, we will associate a maximal collection of subsets of $GOAL$, $\mathcal{V}(w)$, that Alice can exclude, starting from that state. These $\mathcal{V}(w)$ must satisfy certain local one-step optimality conditions that we will identify. These will then help us formulate an iterative algorithm to actually compute these $\mathcal{V}(w)$ for all w.

One-step optimality conditions for a state with Alice's move: Consider a state $w = (s, alice)$ where it is Alice's turn to move. Let $K(w)$ be the set of states that are one-step reachable from this state, i.e., $K(w) = K(s, alice) = \{(s', bob) : s' \in R(w)\}$. Suppose that for each $u \in K(w)$, one has already computed $\mathcal{V}(u)$. What should then $\mathcal{V}(w)$ be?

Consider the situation shown in the Figure 4.2. The set of GOAL states is $\{A, B, C, D, E\}$ and the system is in the state $w = (C, alice)$. $V(u)$ for $u \in K(w)$ is as shown. Now note that since Alice can choose to enter any of the states in $K(w)$, she can effectively exclude any set present in any of the $\mathcal{V}(u)$ for $u \in K(w)$. But since she is already in position C, she cannot exclude C. Hence $\mathcal{V}(w)$ is the maximal sub-collection of the collection of all the sets present in $\mathcal{V}(u)$ for $u \in K(w)$, with C removed from each set. We thus define the following operator:

Definition 4.2.2. *For a list of collections $\{\mathcal{V}_1, \cdots, \mathcal{V}_K\}$, where each $\mathcal{V}_k \in \mathcal{F}$, consider the collection of all sets in all these collections: $\mathcal{U} = \{A : \exists\, k \in \{1, \cdots, K\}\, s.t.\, A \in \mathcal{V}_k\}$. Then $UPPER(\{\mathcal{V}_1, \cdots, \mathcal{V}_K\}) = \Lambda(\mathcal{U})$.*

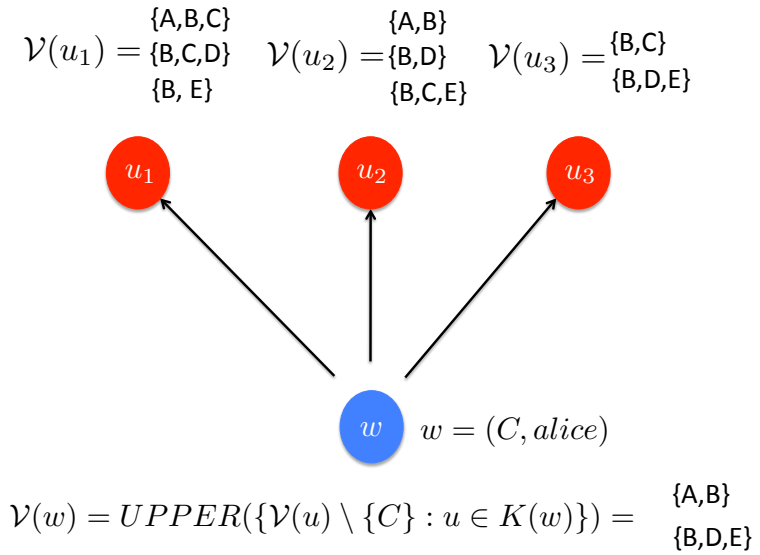

Figure 4.2. One-step optimality condition for a state with Alice's move

Thus the operation $UPPER$ takes a list of maximal collections of sets in $\mathcal{G}$ and outputs the largest maximal sub-collection of the collection of all sets in all of the collections in the list. For example, if $\mathcal{V}_1 = \big\{\{A, B, C\}, \{A, C, D\}, \{B, D\}\big\}$ and $\mathcal{V}_2 = \big\{\{A, B, C\}, \{A, D\}, \{B, C, D\}\big\}$, then

$$UPPER(\{\mathcal{V}_1, \mathcal{V}_2\}) = \big\{\{A, B, C\}, \{A, C, D\}, \{B, C, D\}\big\}.$$

Now if $|\mathcal{V}_i| \leq Q$, then $|\mathcal{U}| \leq KQ$, and this operation can be performed in $O(K^2Q^2G)$ time using the algorithm for computing $\Lambda(\mathcal{U})$ given above. If $w = (s, alice)$ for some $s \in S$, then $\mathcal{V}(w)$ is then given by

$$\mathcal{V}(w) = UPPER(\{\mathcal{V}(u) \setminus \{s\} : u \in K(w)\}), \tag{4.7}$$

where for a collection of sets $\mathcal{U} = (A_1, \cdots, A_N)$ and some set $A \in \mathcal{G}$, by $\mathcal{U} \setminus A$ we mean the collection $\{A_1 \setminus A, \cdots, A_N \setminus A\}$. In the example in the figure, $\mathcal{V}(w) = \big\{\{A, B\}, \{B, D, E\}\big\}$.

One-step optimality conditions for a state with Bob's move: Consider a state $w = (s, bob)$ where it is Bob's turn to move and let $K(w) = K(s, bob) = \{(s', alice) : s' \in R(w)\}$ be the set of states that are one-step reachable by Bob from w. Suppose that for each $u \in K(w)$, one has already computed $\mathcal{V}(u)$. Again, we would like to compute $\mathcal{V}(w)$. Consider

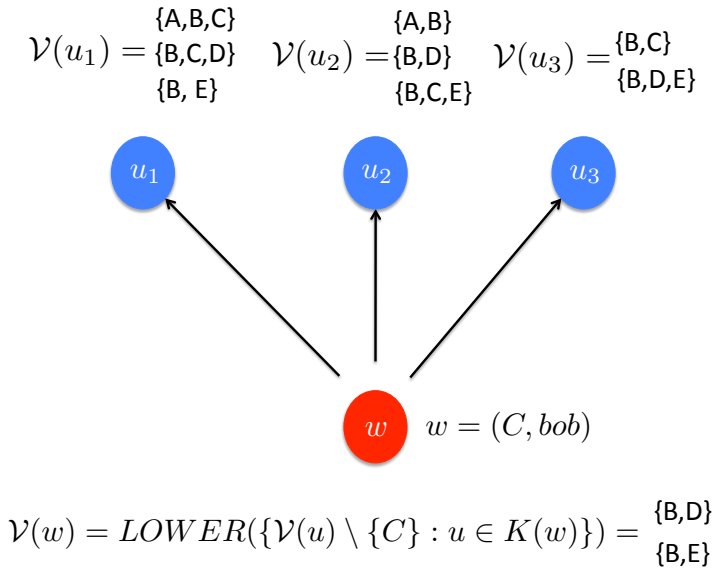

Figure 4.3. One-step optimality condition for a state with Bob's move

the situation shown in the Figure 4.3. The system is in the state $w = (C, bob)$. Clearly Alice cannot exclude C since the system is in position C. Now in each of the states $u \in K(w)$, it is Alice's turn to move, and depending on the state u that Bob chooses to enter, she can respond by choosing any of the set of states in $\mathcal{V}(u) \setminus \{C\}$ to exclude. Such a response plan of Alice consists of a choice of sets $\{A(u) : A(u) \in \mathcal{V}(u) \setminus \{C\}, u \in K(w)\}$, one for each $u \in K(w)$. This response plan guarantees that Alice is able to exclude exactly the set of positions: $\cap_{u \in K(w)} A(u)$, because for any superset of this intersection, Bob can guarantee that he can reach this set (given Alice's response plan). Now by varying over all the possible Alice's response plans, one gets a collection of subsets of $GOAL$ that Alice can exclude. The maximal sub-collection of this collection is thus the collection of maximal excludable subsets starting from state w. We define the following operator:

Definition 4.2.3. *For a list of collections $\{\mathcal{V}_1, \cdots, \mathcal{V}_K\}$, where each $\mathcal{V}_k \in \mathcal{F}$, consider the collection of subsets $\mathcal{U} = \{A_1 \cap A_2 \cap \cdots \cap A_K : A_k \in \mathcal{V}_k\}$. Then $LOWER(\{\mathcal{V}_1, \cdots, \mathcal{V}_K\}) = \Lambda(\mathcal{U})$.*

For $\mathcal{V}_1 = \big\{\{A, B, C\}, \{A, C, D\}, \{B, D\}\big\}$ and $\mathcal{V}_2 = \big\{\{A, B, C\}, \{A, D\}, \{B, C, D\}\big\}$ given above, $LOWER(\{\mathcal{V}_1, \mathcal{V}_2\}) = \big\{\{A, B, C\}, \{A, D\}, \{C, D\}, \{B, D\}\big\}$. Now if $|\mathcal{V}_i| \leq Q$, then $|\mathcal{U}| \leq Q^K$. Computing each intersection takes $O(KG)$ time and hence the total time for com-

puting $\mathcal{U}$ is $O(Q^K KG)$. Then computing the maximal sub-collection $\Lambda(\mathcal{U})$ takes $O(Q^{2K}G)$ time. Thus the total time is $O(Q^{2K}KG)$. The collection of maximal excludable subsets starting from state $w = (s, bob)$ is

$$\mathcal{V}(w) = LOWER(\{\mathcal{V}(u) \setminus \{s\} : u \in R(w)\}). \tag{4.8}$$

In the example given in the figure, $\mathcal{V}(w) = \big\{\{B, D\}, \{B, E\}\big\}$.

These local optimality conditions suggest the following algorithm for computing the maximal excludable subsets of $GOAL$ starting from any state w.

Algorithm 2: Computing the set of maximal excludable subsets of GOAL, starting from each state w.

- **Initialize:** For each $w \in W$, set $\mathcal{V}_0(w) = GOAL$.
- **Do:** For $k > 0$, for each $w = (s, j) \in W$,
  - If $j = bob$, then

$$\mathcal{V}_k(w) = LOWER(\{\mathcal{V}_{k-1}((s', alice)) \setminus \{s\} : s' \in R(w)\}),$$

  - Else if $j = alice$, then

$$\mathcal{V}_k(w) = UPPER(\{\mathcal{V}_{k-1}((s', bob)) \setminus \{s\} : s' \in R(w)\}),$$

- **Until:** $\mathcal{V}_k(w) = \mathcal{V}_{k-1}(w) = \mathcal{V}^*(w)$ for each $w \in W$.
- $\mathcal{V}^*(w)$ is then the maximal collection of excludable subsets of GOAL from the initial state w.

To see why this algorithm converges and gives the correct solution, one only needs to check that there is an equivalence between each iteration of this algorithm and the iteration that one would have performed under the naive approach, which computes the excludability answer for each possible subset of $GOAL$ for each set $w \in W$. Clearly, for a state w, if one is given whether or not one can exclude each possible subset of $GOAL$, then from that information one can extract the maximal excludable subsets. In the other direction, suppose that one is given the maximal excludable subsets of $GOAL$ starting from a state w. Then one can determine excludability for each subset A of goal as follows: if A is contained in some maximal subset (not necessarily strictly), then it is excludable, otherwise it is not.

Now consider the one-step optimality condition defined for the problem of computing maximal excludable sets for a state $w = (s, alice)$, in which it is Alice's turn to move (equation 4.7). For a subset H of $GOAL$, if $s \in H$, then one concludes that it is not excludable by both approaches. If $s \notin H$, then H is excludable only if there is at least one $u \in K(w)$ such that for this state u, $V^H(u) = 1$. But if that is the case, then there is some subset A such that $H \subseteq A$ and $A \in \mathcal{V}(u)$, i.e., it is in the collection of maximal excludable subsets starting from state u. By the one-step optimality condition, there is some maximal set $H' \in \mathcal{V}(w)$ such that $A \subseteq H'$. And thus H is excludable since $H \subseteq H'$. Thus the two approaches give the same answer if H is excludable. Similarly, H is not excludable only if for every $u \in K(w)$, $V^H(u) = 0$. This means that $H \setminus A \neq \phi$ for any A such that $A \in \mathcal{V}(u)$ for any $u \in K(w)$. But then by the one-step optimality condition, this means that $H \setminus A \neq \phi$ for any $A \in \mathcal{V}(w)$. Thus the two approaches give the same answer if H is not excludable.

Next consider the one-step optimality condition defined for the problem of computing maximal excludable sets for a state $w = (s, bob)$, in which it is Bob's turn to move (equation 4.8). For a subset H of $GOAL$, if $s \in H$, then one concludes that it is not excludable by both approaches. If $s \notin H$, then H is excludable only if for all $u \in K(w)$, $V^H(u) = 1$. But if that is the case, then there are sets $\{A(u) : u \in K(w)\}$ such that $H \subseteq A(u)$ for each u. Hence $H \subseteq \cap_{u \in R(w)} A(u)$. Now by the one-step optimality condition, there is some maximal set $H' \in \mathcal{V}(w)$ such that $\cap_{u \in R(w)} A(u) \subseteq H'$. And thus H is excludable since $H \subseteq H'$. Thus the two approaches give the same answer if H is excludable. Similarly, H is not excludable only if there is some $u' \in K(w)$, such that $V^H(u') = 0$. This means that $H \setminus A \neq \phi$ for any A such that $A \in \mathcal{V}(u')$. But then for any choice of sets $\{A(u) : u \in K(w)\}$, $H \setminus \cap_{u \in R(w)} A(u) \neq \phi$. But by the one-step optimality condition this means that $H \setminus A \neq \phi$ for any $A \in \mathcal{V}(w)$. Thus the two approaches give the same answer if H is not excludable.

Chapter 5

Conclusion

This thesis extended the dynamic programming paradigm that is ubiquitously used in dynamic optimization problems and games, to the problem of computing minmax optimal strategies in dynamic vector-valued games. Since these vector spaces of payoffs are only partially ordered, the extremal elements of compact subsets of these spaces are not singletons, but rather a set of potentially multiple elements. Hence the outer minimization in the minmax operator to compute the simultaneous guarantees results in a set of points. Each of these points is minimal in the sense that Alice cannot achieve a simultaneous guarantee that is dominated by one of these points. One of the key messages of this thesis is that if one wishes to compute the minimal achievable simultaneous guarantees in dynamic games, and if one wishes to do so by temporally decomposing the problem and using a backward inductive procedure, then one needs to operate on the entire set of achievable guarantees at each stage.

Dynamic programs that operate on a compact state space are commonly used in solving partially observable Markov Decision Processes and Stochastic games, see [27, 21, 19] (also, e.g. such a program is used to solve the optimal maxmin policy for the informed player in Aumann and Maschler's model as described in Chapter 3). There the state space is the space of beliefs. The key difference between these dynamic programming operators and our operator is that in the former case, the state transitions are a result of Bayesian updates of the beliefs, whereas in our case these transition rules are control variables that are optimally chosen.

5.1 Future directions

We propose the following directions.

- Our characterization of the minimal simultaneous guarantees for discounted vector-

valued repeated games opens up the possibility of several dynamic programming based algorithms to approximate the optimal guarantees and compute approximately optimal strategies. We presented one such algorithm based on value iteration combined with finitely parametrized approximations of the set of guarantees. It would be interesting to explore other approaches and characterize the optimal error-complexity tradeoffs

- Aumann and Maschler also considered the model of zero-sum Repeated games with incomplete information on both sides, where both Alice and Bob have some partial information of the underlying repeated game. It would be interesting to extend our dynamic programming approach to compute the minmax policy for Alice (or maxmin for Bob) in this case. We believe that this should not be too difficult. One can imagine a dynamic programming operator defined on the product of two state spaces: one is the space of commonly held Bayesian beliefs about the information known by Alice, and the other is the compact state space coming from Alice trying to adaptively simultaneously optimize for the different possibilities of the information held by Bob. The transitions of the first state will be Bayesian, while that of the second will be optimally chosen by the operator.

- The set-valued dynamic programming approach that we presented in this thesis for discounted vector-valued dynamic games has a lot of potential for extensions. First, note that most of the results in Chapters 2-4 are for the case of discounted payoffs. It would be interesting to see if one can extend these to the case of limiting average payoffs defined in Chapter 1. Note that dynamic programming operators for MDPs (see [27]) and certain classes of Stochastic games (see [20]) with limiting average payoffs are well studied. This could potentially result in a derivation of new no-regret strategies paralleling Blackwell's strategy and others, based on purely dynamic programming based approaches.

- Similarly, it would be interesting to extend our approach to computing minimal guarantees in Reachability games where the players move simultaneously. More generally, a Reachability game is a special type of recursive game [13]. These games are stochastic games where each stage either does not give any immediate payoff, or is terminating, i.e., the game never gets out of that state. Everett [13] showed that with limiting average payoffs, these games have a value (i.e., a minmax theorem holds) and further, the players have stationary optimal strategies. We believe that the vector valued versions of both repeated games and recursive games with limiting average payoffs possess sufficiently nice structure that make them attractive immediate subjects for the extension of our approach.

Bibliography

[1] Jacob Abernethy, Peter L. Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In *Proceedings of COLT*, 2011.

[2] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The non-stochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[3] Robert J. Aumann and Michael Maschler. *Repeated Games with Incomplete Information.* MIT Press, 1995.

[4] David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.*, 6(1):1–8, 1956.

[5] David Blackwell. Controlled random walks. In J De Groot and J.C.H Gerretsen, editors, *Proceedings of the International Congress of Mathematicians 1954*, volume 3, pages 336–338, 1956.

[6] Avrim Blum and Yishay Mansour. From external to internal regret. In *Proceedings of COLT*, pages 621–636, 2005.

[7] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and non-stochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[8] Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, 1997.

[9] Nicolò Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006.

[10] Nicol Cesa-Bianchi and Gbor Lugosi. Potential-based algorithms in on-line prediction and game theory. *Machine Learning*, 51(3):239–261, 2003.

[11] Alexey Chernov and Fedor Zhdanov. Prediction with expert advice under discounted loss. In *Algorithmic Learning Theory*, pages 255–269. Springer, 2010.

[12] Luca De Alfaro, Thomas Henzinger, Orna Kupferman, et al. Concurrent reachability games. In *Foundations of Computer Science, 1998. Proceedings. 39th Annual Symposium on*, pages 564–575. IEEE, 1998.

[13] Hugh Everett. Recursive games. *Contributions to the Theory of Games*, 3(39):47–78, 1957.

[14] Dean P. Foster and Rakesh V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1–2):40–55, 1997.

[15] Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1–2):79–103, 1999.

[16] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[17] James Hannan. Approximation to Bayes risk in repeated plays. In M. Dresher, A. W. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games*, volume 3, pages 97–139. Princeton University Press, 1957.

[18] Sergiu Hart and Andreu Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98(1):26 – 54, 2001.

[19] Milos Hauskrecht. Value-function approximations for partially observable markov decision processes. *Journal of Artificial Intelligence Research*, pages 33–94, 2000.

[20] Alan J Hoffman and Richard M Karp. On nonterminating stochastic games. *Management Science*, 12(5):359–370, 1966.

[21] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.

[22] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[23] Ehud Lehrer. Approachability in infinite dimensional spaces. *International Journal of Game Theory*, 31(2):253–268, 2003.

[24] N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

[25] Shie Mannor, Vianney Perchet, and Gilles Stoltz. Set-valued approachability and online learning with partial monitoring. *Journal of Machine Learning Research*, 15:3247–3295, 2014.

[26] Vianney Perchet. Approachability, regret and calibration: Implications and equivalences. *Journal of Dynamics and Games*, 1(2):181–254, 2014.

[27] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[28] Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, 3rd edition, 1986.

[29] Lloyd S Shapley. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 39(10):1095, 1953.

[30] Sylvain Sorin. *A First Course on Zero Sum Repeated Games*. Springer, 2002.

[31] Gilles Stoltz and Gbor Lugosi. Internal regret in on-line portfolio selection. *Machine Learning*, 59(1–2):125–159, 2005.

[32] Nicolas Vieille. Weak approachability. *Mathematics of Operations Research*, 17(4):pp. 781–791, 1992.

[33] Volodimir G. Vovk. Aggregating strategies. In *Proceedings of COLT*, pages 371–386, 1990.

[34] Shmuel Zamir. Chapter 5 repeated games of incomplete information: Zero-sum. In Robert Aumann and Sergiu Hart, editors, *Handbook of Game Theory with Economic Applications*, volume 1, pages 109–154. Elsevier, 1992.