Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

An Approximate Dynamic Programming Approach to Repeated Games with Vector Losses

Vijay Kamble, Patrick Loiseau, Jean Walrand

To cite this article:
Vijay Kamble, Patrick Loiseau, Jean Walrand (2022) An Approximate Dynamic Programming Approach to Repeated Games with Vector Losses. Operations Research

Published online in Articles in Advance 29 Aug 2022

. https://doi.org/10.1287/opre.2022.2334

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–16
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

An Approximate Dynamic Programming Approach to Repeated Games with Vector Losses

Vijay Kamble,[a,*] **Patrick Loiseau,**[b,c] **Jean Walrand**[d]

[a] Department of Information and Decision Sciences, University of Illinois, Chicago, Illinois 60607; [b] University Grenoble Alpes, INRIA, Centre National de la Recherche Scientifique, Institut Polytechnique de Grenoble, Laboratoire d'Informatique de Grenoble, 38058 Grenoble, France; [c] Max-Planck Institute for Software Systems, D-66123 Saarbrücken, Germany; [d] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720
*Corresponding author
Contact: kamble@uic.edu, https://orcid.org/0000-0002-9261-1612 (VK); patrick.loiseau@inria.fr (PL); walrand@berkeley.edu (JW)

Received: September 29, 2018
Revised: October 26, 2020; March 3, 2022
Accepted: June 3, 2022
Published Online in Articles in Advance: August 29, 2022

Area of Review: Optimization

https://doi.org/10.1287/opre.2022.2334

Abstract. We describe an approximate dynamic programming (ADP) approach to compute approximations of the optimal strategies and of the minimal losses that can be guaranteed in discounted repeated games with vector-valued losses. Among other applications, such vector-valued games prominently arise in the analysis of worst-case regret in repeated decision making in unknown environments, also known as the adversarial online learning framework. At the core of our approach is a characterization of the lower Pareto frontier of the set of expected losses that a player can guarantee in these games as the unique fixed point of a set-valued dynamic programming operator. When applied to the problem of worst-case regret minimization with discounted losses, our approach yields algorithms that achieve markedly improved performance bounds compared with off-the-shelf online learning algorithms like Hedge. These results thus suggest the significant potential of ADP-based approaches in adversarial online learning.

Funding: This work has been partially supported by the Multidisciplinary Institute in Artificial Intelligence (MIAI) at Grenoble Alpes (ANR-19-P3IA-0003), by the French National Research Agency (ANR) [Grant ANR-20-CE23-0007], by the U.S. Airforce Office of Scientific Research (AFOSR) [Grant MURI FA9550-10-1-0573], by the France-Berkeley Fund, and by the Alexander von Humboldt Foundation.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2022.2334.

Keywords: vector repeated games • online learning • approximate dynamic programming

1. Introduction

In several decision-making scenarios in uncertain and potentially adversarial environments, a decision maker cares about multiple objectives at the same time. For example, in defense operations, an agent might be interested in simultaneously defending multiple targets against an enemy. In repeated decision making in an unknown environment, an agent may want to perform as well in hindsight as every policy in a given class of policies. In asymmetric information games where a player lacks some information that other players have, a natural goal for the player is to choose a strategy that gives appropriate worst-case guarantees simultaneously across the different underlying possibilities (e.g., countries lacking knowledge about the arsenal and defense technologies possessed by other countries). One can model many such scenarios as a vector-valued sequential game between the agent and an adversary.

In this paper, we analyze a simple class of such sequential games: two-player repeated games with vector-valued losses. These are two-player games in which a single-stage, simultaneous-move game with actions that result in vector-valued losses to one player is repeated many times. The player that incurs these losses wants to minimize them while the adversary wants to maximize them; however, because the losses are multidimensional, the relative importance of the different components matters in the choice of a good strategy for the minimizing player.

Focusing on the case where the losses are discounted over time, we describe an approximate dynamic programming (ADP) approach to calculate the best bounds on the losses that a payer can guarantee *simultaneously* in all the components. Formally, our approach approximates the lower Pareto frontier of the set of all points $\mathbf{b} \in \mathbb{R}^K$ (where K is the dimension of the loss vector), such that the loss-minimizing player can guarantee that the expected losses in the game are contained in *the lower corner set* $\{\mathbf{x} \in \mathbb{R}^K : \mathbf{x} \leq \mathbf{b}\}$.[1] We characterize this optimal Pareto frontier as the unique fixed point of a set-

valued dynamic programming (DP) operator. Using this characterization, we propose two computational schemes to derive good policies. The first scheme computes simple finite-state controllers that approximately achieve different points on the optimal frontier but becomes prohibitive for larger values of K. The second scheme addresses such higher-dimensional settings at the cost of loss in optimality. We discuss two applications of our results.

1.1. Application to Worst-Case Regret Minimization

The main application that motivates our study is the problem of worst-case regret minimization in repeated decision-making in unknown environments. The motivating problem is described as follows. A decision maker, Alice, faces a fixed decision-making problem over a finite set of actions every day (e.g., which stocks to invest in, or which investment planning experts' advice to follow). However, the losses corresponding to each action on each day are a priori unknown and revealed only after the decision is taken. One possible way to design a decision-making algorithm in such settings is to assume that an adversary chooses the losses corresponding to these actions to maximize Alice's loss. This results in a zero-sum game between Alice and the adversary, in which Alice wants to choose a randomized strategy that minimizes her worst-case expected loss over the actions chosen by the adversary. Such a design approach, however, results in overly pessimistic decision-making algorithms. For example, if Alice has two actions, 1 and 2, and the adversary is assumed to be constrained to give a loss of one to one action and zero to the other,[2] then the minimax optimal strategy for Alice is to pick either of the actions with equal probability each day because any bias toward a particular action will be exploited by the adversary. However, although such a static strategy is optimized against a fully adversarial environment, it fails to exploit the possibility that the environment may not be fully adversarial; indeed, few decision-making environments are. Perhaps action 1 may turn out to be the one that consistently leads to lower losses, which a smarter algorithm could have realized over time and exploited that knowledge by cautiously increasing the probability of choosing that action. The worst-case optimal algorithm fails to exploit such information, leading to Alice regretting her choices in hindsight.

Addressing this issue, the worst-case regret minimization framework proposes a way to design algorithms that adaptively learn to play well against an unknown adversary. The way it does so is by changing the objective of the decision maker and the adversary: instead of minimizing her worst-case loss, Alice designs an algorithm that minimizes her worst-case "regret," which is the incremental loss relative to the loss of the best fixed action that could have been chosen in hindsight against the sequence of losses chosen by the adversary. The adversary, in turn, is assumed to maximize this regret rather than maximizing Alice's loss. Thus, if the environment turns out to be "nicer," for example, one action consistently leads to a lower loss, then a regret minimizing algorithm would learn that over time and choose that action with a higher probability to keep the regret low.

Because of its practical applicability as an effective decision-making paradigm in unknown environments, the theory of worst-case regret minimization, also known as *adversarial online learning*, has been extensively developed in the literature (Cesa-Bianchi and Lugosi 2003, Hazan 2016). Algorithms are known that achieve an $O(1/\sqrt{T})$ average regret with high probability over a fixed horizon T in the worst-case over sequences of losses that can be chosen by the adversary. In the infinite-horizon case with losses discounted over time by a factor $\beta \in (0, 1)$, similar algorithms achieve an expected discounted average regret of $O(\sqrt{1-\beta})$ as $\beta \to 1$ in the worst-case. Such algorithms are called *no-regret* algorithms, because, remarkably, the worst-case average regret vanishes as $T \to \infty$ or as $\beta \to 1$. Although these algorithms have good asymptotic performance in these regimes, the performance of such algorithms can be far from optimal for a fixed T or β. Our present work addresses this shortcoming: we show that our approach to designing near-optimal strategies for discounted repeated games with vector losses can be used to construct near-optimal algorithms for worst-case regret minimization in such fixed-parameter settings, assuming that the losses chosen by the adversary lie in a finite set.

As an illustration, we use our approach to compute near-optimal algorithms for the well-known problem of prediction using expert advice with binary, discounted losses, and two experts (Bubeck and Cesa-Bianchi 2012). We show that these algorithms achieve lower regret guarantees than those achieved by existing benchmark policies in adversarial online learning, including the well-known Hedge algorithm (also known as the *exponentially weighted average forecaster*; Bubeck and Cesa-Bianchi 2012). To the best of our knowledge, this is the first class of provably near-optimal algorithms for this setting. We also numerically evaluate our scheme for designing good policies for regret minimization in higher-dimensional settings, where we continue to find that these policies achieve lower regret guarantees compared with Hedge. Our results thus suggest the significant potential in using ADP approaches to designing effective adversarial online learning algorithms.

1.2. Application to Repeated Games with Incomplete Information

Finally, our results also have applications to other problems where the theory of repeated games with vector-valued losses is applicable. In Section EC.10 in the online

appendix, we discuss the application of our results to the well-known model of discounted repeated games with incomplete information on one side from Aumann et al. (1995). In particular, our results yield the first known characterization of the optimal policy for the uninformed player in this model.

1.3. Organization of the Paper

The paper is organized as follows. We first discuss related literature in Section 2. In Section 3, we formally introduce the model of repeated games with vector-values losses and define the objective of characterizing the set of losses that can be guaranteed by a player. We also discuss the connection of the problem to adversarial online learning. In Section 4, we introduce the set-valued dynamic programming approach for characterizing the optimal guarantees and the corresponding optimal strategies. In Section 5, we present an approximation procedure to design near-optimal policies based on our characterization. Section 6 discusses a procedure to design good policies in higher-dimensional settings. Section 7 presents a numerical evaluation of the algorithms resulting from our approach in the context of regret minimization. We discuss the limitations of our approach in Section 8 and conclude the paper. The proofs of all our results are presented in Section EC.1 in the online appendix.

2. Related Literature

Our results contribute to several streams of related literature as we discuss here.

2.1. Repeated Games with Vector Losses

Blackwell (1956) pioneered the study of two-player vector-valued repeated games under the long-run average loss criterion. He described necessary and sufficient conditions for any convex set of loss vectors to be *approachable* by a player, which means that there exists a strategy for the player that ensures that the long-run average loss approaches this set almost surely regardless of the adversary's actions. He also defined an adaptive randomized strategy that ensures this. Furthermore, he proved the following remarkable *minimax* theorem: any convex set is either approachable by the player or is *excludable* by the adversary, which means that the adversary has a strategy that guarantees that the long-run average loss remains *outside* this set almost surely. These initial results established the now well-known "approachability" framework to analyze these games.

Approachability theory has developed significantly over the years (see Perchet 2014 or Laraki and Sorin 2015 for a survey). Necessary and sufficient conditions for approachability of general sets have been established in Spinat (2002). Vieille (1992) considers a weaker notion called "weak approachability," and it is shown that every set is either weakly approachable or weakly excludable. There have also been several extensions of this framework beyond the setting of repeated games with finite action spaces, for example, to stochastic games (Milman 2006), to repeated games with payoffs in infinite-dimensional spaces (Lehrer 2003), to repeated games with partial monitoring (Perchet 2011a, b; Perchet and Quincampoix 2014).

Despite these advances in our understanding of vector repeated games, the characterization and computation of loss vectors that can be guaranteed in expectation under the discounted loss criterion and the strategies that achieve these guarantees has remained a significant gap. Discounting of losses over time has natural interpretations in practice; for example, it may capture a low-risk rate of return on investment. Hence, the closing of this gap in our work has significant practical ramifications.

2.2. Set-Valued Dynamic Programming

A well-known use of set-valued dynamic programs in the context of dynamic games is from Abreu et al. (1986, 1990). They characterize the set of pure strategy subgame-perfect equilibrium payoffs in nonzero sum repeated games with imperfect monitoring as the fixed point of a set-valued DP operator. Although we have a similar fixed-point characterization of the optimal Pareto frontier, there are important differences in the machinery involved in iterative computation of this fixed point. Our ADP approach and related error bounds critically rely on the new metric that we define on the space of Pareto frontiers of convex and compact sets and on the fact that our DP operator is a contraction in this metric. On the other hand, their iterative computation scheme relies on the monotonicity of the DP operator[3]; because they do not define a metric space, they do not obtain error bounds. Moreover, theirs is an exact scheme akin to *value iteration* in DP (Bertsekas 2012) and implementing it in practice would require defining appropriate set approximations that are finitely parameterized. Our carefully defined polytopic set approximation scheme squarely addresses this issue in our setting.

For the use of dynamic programming in zero-sum dynamic games, one can refer to the classic paper by Shapley (1953) on stochastic games. For a general theory of dynamic programming in control problems and Markov decision processes (MDPs), see Bertsekas (2005, 2012) and Puterman (2014).

2.3. Regret Minimization in Repeated Games

The first study of regret minimization in repeated games dates back to the pioneering work of Hannan (1957), who introduced the notion of regret optimality in repeated games and proposed the earliest known no-regret algorithm. Since then, numerous other such algorithms have been proposed, particularly for the problem of prediction using expert advice (Vovk 1990, Littlestone and Warmuth 1994, Cesa-Bianchi et al.

1997, Freund and Schapire 1999), one particularly well-known class being the multiplicative weights update class of algorithms. Regret minimization with discounted losses has been considered before in Perchet (2014), Cesa-Bianchi and Lugosi (2003), and Chernov and Zhdanov (2010). Other settings with limited feedback have been considered, most notably the multiarmed bandit setting (Auer et al. 2002, Bubeck and Cesa-Bianchi 2012). Stronger notions of regret such as internal regret, have also been studied (Foster and Vohra 1997, Stoltz and Lugosi 2005, Cesa-Bianchi and Lugosi 2006, Blum and Mansour 2007). Cesa-Bianchi and Lugosi (2003) offers a fairly comprehensive survey of the literature.

The results on *exact* regret minimization are few. In an early work, Cover (1966) gave the optimal algorithm for the problem of prediction using expert advice over any finite horizon T, for the case of two experts and where the losses are {0, 1}. Recently, Gravin et al. (2016) extended the result to the case of three experts for both the finite horizon and geometrically distributed random horizon problems. Bayraktar et al. (2020) further extended these results to characterize the optimal scaling of the regret for the case of four experts in the geometric horizon model in the regime where the stopping probability approaches zero. In general, the direct dynamic programming approach proposed by Gravin et al. (2016) for this problem has a complexity that is exponential in the number of experts. Characterizing the optimal scaling of regret and defining a polynomial-time scheme to compute an asymptotically optimal algorithm for the general $K > 4$ experts problem in this setting remains an interesting open problem.

Although a geometric time horizon model appears to be an interpretation of the infinite horizon model with discounted losses that we consider, the two problem formulations define regret differently and thus lead to different optimal regrets and algorithms. Moreover, our formulation presents new computational challenges that are absent in the formulation of Gravin et al. (2016), and a significantly different approach is necessary to obtain similar guarantees. We discuss this distinction in detail in Section EC.6 in the online appendix.

Abernethy et al. (2008) considers a related problem, where a gambler places bets from a finite budget repeatedly on a fixed menu of events, the outcomes of which are adversarially chosen from {0, 1} (you win or you lose), and characterizes the minimax optimal strategies for the gambler and the adversary. Luo and Schapire (2014) considers a similar repeated decision-making problem where an adversary is restricted to pick loss vectors (i.e., a loss for each action of the decision maker in a stage) from a set of basis vectors and characterizes the minimax optimal strategy for the decision maker under both a fixed and an unknown horizon. Most of the approaches in these works are specific to their settings and exploit the assumptions on the structure of the loss vectors. However, if the loss vectors are arbitrary, these approaches are difficult to generalize, and it is generally recognized that characterizing the optimal regret and algorithm is difficult (Luo and Schapire 2014). The idea of characterizing the Pareto frontier of all achievable regrets with respect to different actions has been explored in Koolen (2013), again in the specific context of prediction with expert advice with two experts and {0, 1} losses, and for the finite time horizon problem without discounting. We, on the other hand, consider regret minimization in arbitrary repeated games, and moreover, with losses that are discounted over time. Characterizing and computing the optimal policies in this setting necessitates the development of the new machinery of a contractive set-valued dynamic programming operator and its approximation.

Finally, all the previous examples deal with games with finite action spaces, which is the setting that we are concerned with. However, there are many works that consider exact minimax optimality in repeated games with general action sets, with specific types of loss functions (see Koolen et al. 2014, 2015 and Bartlett et al. 2015 and references therein).

3. Model

For the remainder of the paper, $\mathbf{1}$ and $\mathbf{0}$ denote the vector of ones and zeros, respectively, in $\mathbb{R}^K$.

Consider a two-player vector-valued game $\mathbb{G}$ defined by an action set $A = \{1, \cdots, l\}$ for player 1, who is the decision maker and whom we will call Alice, and the action set $B = \{1, \cdots, m\}$ 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 $\mathbf{r}(a, b) \in \mathbb{R}^K$.

The game $\mathbb{G}$ is played repeatedly in stages $t = 1, 2, 3, \cdots, T$. Let $\mathbb{G}^T$ denote this T-stage repeated game. 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 $\mathbf{r}(a_t, b_t)$. Fix a discount factor $\beta \in [0, 1)$. Then the vector of total discounted losses is defined as

$$\sum_{t=1}^{T} \beta^{t-1} \mathbf{r}(a_t, b_t) = \left(\sum_{t=1}^{T} \beta^{t-1} r_k(a_t, b_t); k = 1, \cdots, K \right). \quad (1)$$

An adaptive randomized strategy π_A for Alice specifies for each stage t, a mapping from the set of observations until stage t, that is, $H_t = (a_1, b_1, \cdots, a_{t-1}, b_{t-1})$, to a probability distribution on the action set A, denoted by $\Delta(A)$. Let Π_A be the set of all such strategies of Alice. Similarly, let Π_B be the set of all adaptive randomized strategies for Bob. For a pair of strategies π_A and π_B, the expected discounted loss on

component k in the repeated game is given by

$$R_k^T(\pi_A, \pi_B) = \mathrm{E}_{\pi_A,\pi_B}\left[\sum_{t=1}^{T} \beta^{t-1} r_k(a_t, b_t)\right], \tag{2}$$

where the expectation is over the randomness in the strategies π_A and π_B. Alice would like to minimize her loss in every component k. However, reducing the loss in one dimension typically implies increasing the loss in another dimension. For instance, consider the situation of protecting two different targets against attacks: devoting more resources to protect one target makes the other more vulnerable. Accordingly, it is important to characterize the set of best possible tradeoffs between the different dimensions of the loss.

Consider a fixed strategy $\pi_A \in \Pi_A$. If Alice plays this strategy, then irrespective of the strategy chosen by Bob, Alice guarantees that the long term expected vector loss is no larger than

$$\left(\max_{\pi_B^k \in \Pi_B} R_k^T(\pi_A, \pi_B^k); k = 1, \cdots, K\right)$$

along each dimension. Let the set of all such *simultaneous upper bounds* that correspond to *all* the strategies $\pi_A \in \Pi_A$ be defined as

$$\mathcal{W}^T \triangleq \left\{\left(\max_{\pi_B^k \in \Pi_B} R_k^T(\pi_A, \pi_B^k); k = 1, \cdots, K\right) : \pi_A \in \Pi_A\right\}. \tag{3}$$

Then characterizing the best possible tradeoffs across the different dimensions amounts to finding the *minimal* points in the set $\mathcal{W}$, that is, its *lower Pareto frontier*, which is the set

$$\mathcal{V}^T \triangleq \Lambda(\mathcal{W}^T) \triangleq \{\mathbf{x} \in \mathcal{W}^T : \forall\, \mathbf{x}' \in \mathcal{W}^T \setminus \{\mathbf{x}\}, \exists k \text{ s.t. } x_k < x_k'\}, \tag{4}$$

because all other points are strictly suboptimal. Our goal in this paper is to characterize and approximate the set $\mathcal{V}^\infty$ that can be achieved in the infinite horizon game $\mathbb{G}^\infty$ and compute approximately optimal strategies for Alice in Π_A that approximately guarantee different points in it.

3.1. Application to Adversarial Online Learning

As an application of our model and objective, we now describe how they lead to the solution to the problem of regret minimization in repeated games in the framework of adversarial online learning. This conclusion follows from the transformation of the regret minimization problem into a vector-valued repeated game that we describe here.

We first define the problem of regret minimization in repeated games. Let G be a two-player game with l actions $A = \{1, \ldots, l\}$ for Alice (the decision maker), who is assumed to be the minimizer, and m actions $B = \{1, \ldots, m\}$ for Bob (the adversary), in keeping with the previous notation. 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 game G is played repeatedly for T stages $t = 1, 2, \cdots, T$. 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 $\sum_{t=1}^{T} \beta^{t-1} L(a_t, b_t)$, where $\beta \in (0, 1)$. We define the total discounted regret of Alice as

$$\sum_{t=1}^{T} \beta^{t-1} L(a_t, b_t) - \min_{a \in A} \sum_{t=1}^{T} \beta^{t-1} L(a, b_t), \tag{5}$$

which is the difference between her actual discounted loss and the loss corresponding to the single best action that could have been chosen against the sequence of actions chosen by Bob in hindsight.

Alice chooses an adaptive randomized strategy $\pi_A \in \Pi_A$. Bob is assumed to choose a deterministic oblivious strategy; that is, his choice is simply a sequence of actions $\mathbf{b} = (b_1, b_2, b_3, \cdots, b_T) \in B^T$ chosen before the start of the game.[4] Alice's problem of minimizing her worst-case expected discounted regret is defined as

$$\min_{\pi_A \in \Pi_A} \max_{\mathbf{b} \in B^T} \mathrm{E}_{\pi_A}\left[\sum_{t=1}^{T} \beta^{t-1} L(a_t, b_t) - \min_{a \in A} \sum_{t=1}^{T} \beta^{t-1} L(a, b_t)\right]$$

$$= \min_{\pi_A \in \Pi_A} \max_{\mathbf{b} \in B^T} \max_{a \in A} \mathrm{E}_{\pi_A}\left[\sum_{t=1}^{T} \beta^{t-1} (L(a_t, b_t) - L(a, b_t))\right] \tag{6a}$$

$$\overset{(a)}{=} \min_{\pi_A \in \Pi_A} \max_{\pi_B \in \Pi_B} \max_{a \in A} \mathrm{E}_{\pi_A,\pi_B}\left[\sum_{t=1}^{T} \beta^{t-1} (L(a_t, b_t) - L(a, b_t))\right]. \tag{6b}$$

Equality (a) says that in Problem (6a), there is no loss to Alice if Bob is allowed to choose any adaptive randomized strategy $\pi_B \in \Pi_B$ instead of restricting him to choosing a deterministic oblivious strategy. This is because one can show that Alice's optimal strategy in Problem (6a) need not depend on her own past actions; see Lemma 2 and Theorem 3. Hence, strategies in Π^B are not more powerful than those in B^T against the optimal strategy of Alice. Problem (6b) is called the problem of minimizing the *pseudo-regret* in the online learning literature (see chapter 3 in Bubeck and Cesa-Bianchi 2012). As we saw, it is equivalent to minimizing the expected regret when the adversary is restricted to B^T.

Now the connection to our original vector-valued repeated game is straightforward. 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 $\mathbf{r}(a, b)$ with $K = l$ components

(recall that $|A| = l$), where $r_k(a,b) = L(a,b) - L(k,b)$; $r_k(a,b)$ is the single-stage additional loss that Alice bears by choosing action a instead of action k, when Bob chooses b: referred to as the "single-stage regret" with respect to action k.

Defining the set $\mathcal{W}^T$ as in (3) for the repeated game $\mathbb{G}^T$, it is clear that the minimax optimal regret can be written as

$$\min_{\pi_A \in \Pi_A} \max_{\pi_B \in \Pi_B} \max_{a \in A} \mathrm{E}_{\pi_A}\left[\sum_{t=1}^{T} \beta^{t-1}(L(a_t,b_t) - L(a,b_t))\right] = \min_{\mathbf{x} \in \mathcal{W}^T} \max_k x_k.$$

Hence, to compute the minimax optimal regret, it suffices to compute $\mathcal{W}^T$. In fact, it suffices to compute its lower Pareto frontier $\mathcal{V}^T$, and the strategies that achieve this frontier, because all other points are strictly suboptimal.

4. Set-Valued Dynamic Programming

We first present an informal description of our approach. Let $\mathcal{V}^0 = \{\mathbf{0}\}$, that is, the singleton set containing only the zero vector in $\mathbb{R}^K$. We can show that one can obtain the set $\mathcal{V}^{T+1}$ from the set $\mathcal{V}^T$, by decomposing Alice's strategy in $\mathbb{G}^{T+1}$ into a strategy for the first stage, and a continuation strategy for the remainder of the game from stage 2 onward as a function of the action chosen by both the players in the first stage. The inductive argument results from the fact that the minimal guarantees that she can guarantee from stage 2 onward are exactly the set $\mathcal{V}^T$. Suppose that at the start of $\mathbb{G}^{T+1}$, Alice fixes the following plan for the entire game: she will play a mixed strategy $\boldsymbol{\alpha} \in \Delta(A)$ in stage 1. Then depending on her realized action a and Bob's action b, from stage 2 onward, she will play a continuation strategy that achieves the upper-bound $\mathbf{R}(a,b) \in \mathcal{V}^T$ (she will choose one such point $\mathbf{R}(a,b)$ for every $a \in A$ and $b \in B$). It is strictly suboptimal for Alice to choose any points outside $\mathcal{V}^T$ from stage 2 onward. Now this plan for the entire game $\mathbb{G}^{T+1}$ gives Alice the following simultaneous upper bounds on the expected losses on the K dimensions:

$$\left(\max_{b \in B} \sum_{a \in A} \alpha_a [r_k(a,b) + \beta R_k(a,b)]; k = 1, \cdots, K\right).$$

By varying the choice of $\boldsymbol{\alpha}$ and the map $\mathbf{R}(a,b)$, we can obtain the set of all the simultaneous upper bounds that Alice can achieve in the $(T+1)$-stage game. The lower Pareto frontier of this set is exactly $\mathcal{V}^{T+1}$. Thus, there is an operator Φ, such that

$$\mathcal{V}^{T+1} = \Phi(\mathcal{V}^T)$$

for any $T \geq 0$. In what follows, we will show that this operator is a contraction in the space of lower Pareto frontiers of compact and convex sets, with an appropriately defined metric. This space is shown to be complete, and thus the sequence $\mathcal{V}^T$ converges in the metric to a set $\mathcal{V}^*$, which is the unique fixed point of this operator Φ. As one would guess, this $\mathcal{V}^*$ is indeed the set $\mathcal{V}^\infty$ of minimal simultaneous upper bounds that Alice can achieve in the infinitely repeated game $\mathbb{G}^\infty$.

The rest of this section formalizes these arguments. We will begin the formal presentation of our results by first defining the space of Pareto frontiers that we will be working with.

4.1. Space of Pareto Frontiers in $[0,1]^K$

We work with the following basic definitions.

Definition 1. Consider the following notions of domination and Pareto frontiers.

(a) Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^K$. We say that $\mathbf{u} \preceq \mathbf{v}$ if $u_k \leq v_k$ for all k. Also, we say that $\mathbf{u} \prec \mathbf{v}$ if $\mathbf{u} \preceq \mathbf{v}$ and $\mathbf{u} \neq \mathbf{v}$. For some $\epsilon \geq 0$ if $\mathbf{u} \preceq \mathbf{v} + \epsilon\mathbf{1}$, we say that $\mathbf{v}$ *ϵ-dominates* $\mathbf{u}$. If $\epsilon = 0$, we simply say that $\mathbf{v}$ dominates $\mathbf{u}$.

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

(c) For two Pareto frontiers $\mathcal{U}$ and $\mathcal{V}$, we say that $\mathcal{V}$ ϵ-dominates $\mathcal{U}$ if for every point $\mathbf{v} \in \mathcal{V}$, there is a point $\mathbf{u} \in \mathcal{U}$ that it ϵ-dominates. If $\epsilon = 0$, then we simply say that $\mathcal{V}$ dominates $\mathcal{U}$.

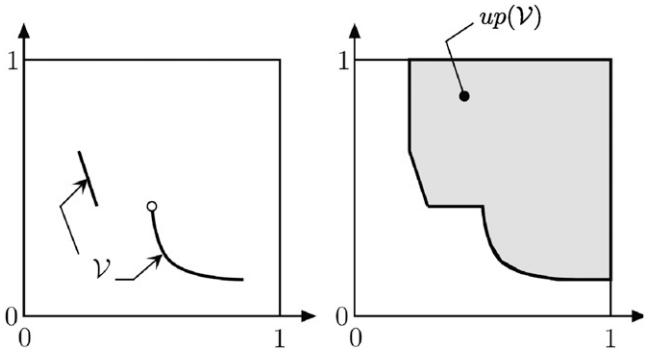
(d) The *lower Pareto frontier* (or simply *Pareto frontier*) of $\mathcal{S} \subset [0,1]^K$, denoted by $\Lambda(\mathcal{S})$, is the set of elements of $\mathcal{S}$ that do not dominate any another element of $\mathcal{S}$.

The Pareto frontier of a set may be empty, as is certainly the case when the set is open. However, one can show that the Pareto frontier of a nonempty compact set is always nonempty.

Lemma 1. *Suppose that $\mathcal{S}$ is a nonempty compact subset of $\mathbb{R}^K$. Then $\Lambda(\mathcal{S})$ is nonempty.*

Because compactness is equivalent to a set being closed and bounded in Euclidean spaces, any closed subset of $[0,1]^K$ has a nonempty Pareto frontier. We next define the *upset* of a set, illustrated in Figure 1.

Figure 1. Pareto Frontier $\mathcal{V}$ and Its Upset $up(\mathcal{V})$ in $[0,1]^2$



Definition 2. Let $\mathcal{A}$ be a subset of $\mathcal{B} \subseteq \mathbb{R}^K$. The *upset* of $\mathcal{A}$ in $\mathcal{B}$ is defined as $up(\mathcal{A}) = \{\mathbf{x} \in \mathcal{B} \,|\, x_k \geq y_k \text{ for all } k, \text{ for some } \mathbf{y} \in \mathcal{A}\}$, that is, $up(\mathcal{A})$ is the set of all points in $\mathcal{B}$ that dominate some point in $\mathcal{A}$. Equivalently, $up(\mathcal{A}) = \{\mathbf{x} \in \mathcal{B} \,|\, \mathbf{x} = \mathbf{y} + \mathbf{v}, \text{ for some } \mathbf{y} \in \mathcal{A} \text{ and } \mathbf{v} \succeq \mathbf{0}\}$.

For a subset of $[0,1]^K$, we will refer to its upset in $[0,1]^K$ as simply its upset. It is immediate that the upset of a closed and convex subset of $[0,1]^K$ is closed and convex. We define the following space of Pareto frontiers.

Definition 3. The term $\mathcal{F}$ is the space of Pareto frontiers in $[0,1]^K$ whose upset is closed and convex.

It is easy to show that $\mathcal{F}$ can be equivalently defined as the space of lower Pareto frontiers of closed and convex subsets of $[0,1]^K$.[5] We will now define a metric on this space. We first recall the definition of *Hausdorff* distance induced by the $\mathcal{L}^\infty$ norm.

Definition 4. Let $\mathcal{A}$ and $\mathcal{B}$ be two subsets of $\mathbb{R}^K$. The Hausdorff distance $h(\mathcal{A}, \mathcal{B})$ between the two sets is defined as

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

The Hausdorff distance defines a metric on the space of nonempty closed subsets of $[0,1]^K$, and furthermore, this space is compact and hence complete in this metric (Henrikson 1999). On the other hand, it only defines a pseudometric on space of all nonempty subsets of $[0,1]^K$. Now, a possible straightforward metric on the space $\mathcal{F}$ could be the one defined by the Hausdorff distance. However, as we discuss in Section EC.2 in the online appendix, if $K > 2$, then a Pareto frontier in $\mathcal{F}$ may not be closed, and hence the Hausdorff distance at best defines a pseudometric on $\mathcal{F}$. Moreover, even this pseudometric is not appropriate for our purposes as demonstrated by the following example.

Example. Consider a sequence of Pareto frontiers $(\mathcal{V}_n)_{n \in \mathbb{N}}$, where $\mathcal{V}_n$ is the union of the line segment joining (0, 1) and $(1/n, 1/n)$, and the segment joining $(1/n, 1/n)$ and (1, 0), as depicted in Figure 2. Then we would like this sequence of frontiers to converge to the Pareto frontier defined by the singleton set $\{(0,0)\}$, but the Hausdorff distance between $\mathcal{V}_n$ and $\{(0,0)\}$ does not vanish as $n \to \infty$. Under the Hausdorff metric, the sequence $(\mathcal{V}_n)_{n \in \mathbb{N}}$ converges to the union of the line segment joining (0, 1) and (0, 0), and the segment joining (0, 0) and (1, 0), which is not in $\mathcal{F}$.

It is thus clear that we need to define a different metric on $\mathcal{F}$. We now proceed to define one with the desired properties. We define the distance between two Pareto frontiers in $\mathcal{F}$ as the Hausdorff distance between their upsets.[6]

Definition 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 $d(\mathcal{U}, \mathcal{V}) \triangleq h(up(\mathcal{U}), up(\mathcal{V}))$.

We can then show that d is a metric on $\mathcal{F}$, and $\mathcal{F}$ is compact in the metric d. The latter essentially follows from the compactness of the space of closed subsets of $[0,1]^K$ in the Hausdorff metric. It also immediately follows that $\mathcal{F}$ is complete.

Proposition 1. The following statements hold.

(a) *The variable d is a metric on $\mathcal{F}$.*

(b) *Let $(\mathcal{V}_n)_{n \in \mathbb{N}}$ be a sequence in $\mathcal{F}$. Then there is a subsequence $(\mathcal{V}_{n_k})_{k \in \mathbb{N}}$ and a $\mathcal{V} \in \mathcal{F}$ such that $d(\mathcal{V}_{n_k}, \mathcal{V}) \to 0$.*

In the proof, it becomes clear that d induces these properties not just on $\mathcal{F}$, but also on the more general space of Pareto frontiers in $[0,1]^K$ whose upset is closed (though not necessarily convex). Finally, we end this section by presenting another way of defining the same metric d on $\mathcal{F}$.

Definition 6. For two Pareto frontiers $\mathcal{V}$ and $\mathcal{U}$ in $\mathcal{F}$, define

$$e(\mathcal{U}, \mathcal{V}) \triangleq \inf\{\epsilon \geq 0 : \forall \mathbf{u} \in \mathcal{U}, \exists \mathbf{v} \in \mathcal{V} \text{ s.t. } \mathbf{v} \preceq \mathbf{u} + \epsilon \mathbf{1}\}. \quad (7)$$

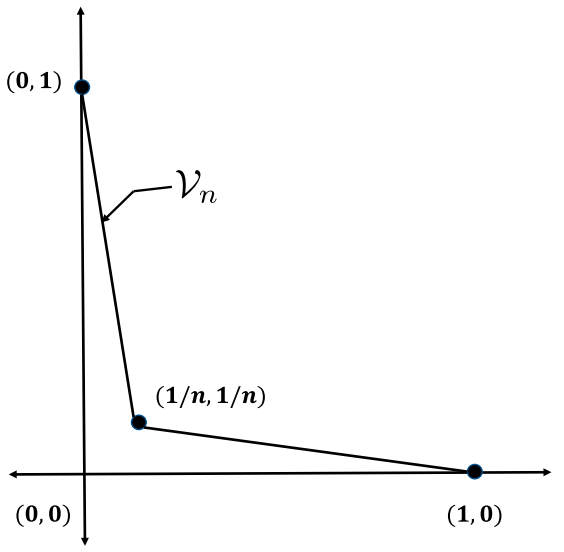
In other words, $e(\mathcal{U}, \mathcal{V})$ is the smallest $\epsilon \geq 0$ such that $\mathcal{U}$ ϵ-dominates $\mathcal{V}$ (note that e is not a symmetric distance).[7] We can then show the following.

Proposition 2. *For any two Pareto frontiers $\mathcal{V}$ and $\mathcal{U}$ in $\mathcal{F}$,*

$$d(\mathcal{U}, \mathcal{V}) = \max(e(\mathcal{U}, \mathcal{V}), e(\mathcal{V}, \mathcal{U})).$$

This means that the distance d between two frontiers $\mathcal{V}$ and $\mathcal{U}$ is less than or equal to ϵ if both $\mathcal{U}$ and $\mathcal{V}$ ϵ-dominate each other. This way of defining d is

Figure 2. A Sequence of Pareto Frontiers Converging to the Point (0, 0)



attractive because it does not require defining upsets of the Pareto frontiers as we do in Definition 5.

4.2. 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 rewards of the game take values in $[0, 1]$ irrespective of the time horizon. Now, for any set $\mathcal{S} \subseteq [0,1]^K$, define the following operator Ψ that maps $\mathcal{S}$ to a subset of $\mathbb{R}^K$:

$$\Psi(\mathcal{S}) = \left\{ \left(\max_{b\in B} \sum_{a\in A} \alpha_a [r_k(a,b) + \beta R_k(a,b)]; k = 1, \cdots, K \right) : \boldsymbol{\alpha} \in \Delta(A),\ \mathbf{R}(a,b) \in \mathcal{S}\ \ \forall a \in A, b \in B \right\}. \quad (8)$$

This operator can be interpreted as follows. Assuming that $\mathcal{S}$ is the set of vectors of simultaneous upper bounds on expected losses that Alice can ensure in $\mathbb{G}^T$, $\Psi(\mathcal{S})$ is the set of vectors of simultaneous upper bounds on expected losses that she can ensure in $\mathbb{G}^{T+1}$. If $\mathcal{S}$ is convex then $\Psi(\mathcal{S})$ is not necessarily convex as we demonstrate in the following example.

Example. Consider the game depicted in Figure 3. Suppose that $\mathcal{S} = \{(0,0)\}$, which is convex. Then for any discount factor β and any $(\alpha, 1-\alpha)$, where $\alpha \in [0, 1]$, one obtains the guarantee:

$$u(\alpha) = (\max(2(1-\alpha), 2\alpha + 4(1-\alpha)), \max(2\alpha, 2(1-\alpha))).$$

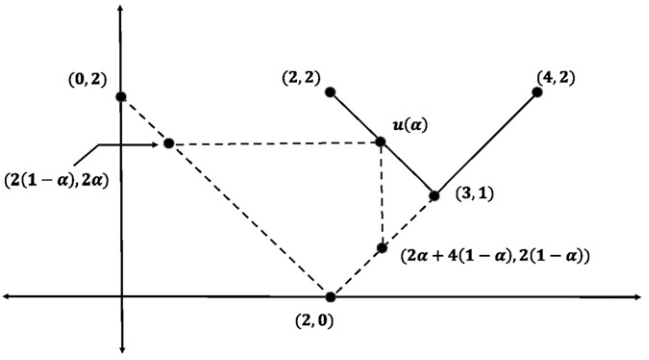
This is depicted in Figure 4. Thus, by varying α, we find that the set $\Psi(\mathcal{S})$ is the union of the line segment joining points (2, 2) and (3, 1) and the segment joining the points (3,1) and (4,2). Clearly, this set is not convex.

Despite the fact that the operator Ψ does not preserve convexity, we can nevertheless show that if $\mathcal{V}$ is a Pareto frontier in $\mathcal{F}$ (which means that it is the Pareto frontier of a convex and closed set), then the Pareto frontier of $\Psi(\mathcal{V})$ is also in $\mathcal{F}$ (observe that in the previous example, the Pareto frontier of $\Psi(\{(0,0)\})$ is the line segment joining points (2, 2) and (3, 1)). Furthermore, we can also show

Figure 3. Game with Vector Losses

	1	2
1	(0, 2)	(2, 0)
2	(2, 0)	(4, 2)

Figure 4. Construction of the Set $\Psi(\mathcal{S})$ Where $\mathcal{S} = \{(0,0)\}$, for the Game Shown in Figure 3



that if $\mathcal{V} \in \mathcal{F}$ is the set of vectors of simultaneous upper bounds on expected losses that Alice can ensure in $\mathbb{G}^T$, then in any optimal plan for Alice in $\mathbb{G}^{T+1}$, the continuation strategy from stage 2 onward need not depend on her own action in stage 1.

Lemma 2. *Let $\mathcal{V} \in \mathcal{F}$. Then:*

1. *We have that $\Lambda(\Psi(\mathcal{V})) \in \mathcal{F}$.*
2. *Any point $\mathbf{u}$ in $\Lambda(\Psi(\mathcal{V}))$ is of the form:*

$$\mathbf{u} = \left(\max_{b\in B} \left[\sum_{a\in A} \alpha_a r_k(a,b) + \beta Q_k(b) \right]; k = 1, \cdots, K \right),$$

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

We next define the following dynamic programming operator Φ on $\mathcal{F}$.

Definition 7 (Dynamic Programming Operator). For $\mathcal{V} \in \mathcal{F}$, we define $\Phi(\mathcal{V}) = \Lambda(\Psi(\mathcal{V}))$.

From Lemma 2, 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.

Lemma 3. *We have that $e(\Phi(\mathcal{U}), \Phi(\mathcal{V})) \le \beta e(\mathcal{U}, \mathcal{V})$, and hence $d(\Phi(\mathcal{U}), \Phi(\mathcal{V})) \le \beta d(\mathcal{U}, \mathcal{V})$.*

Finally, the completeness of $\mathcal{F}$ and the fact that Φ is a contraction in d directly implies the following result as a consequence of the Banach fixed point theorem (Munkres 2000).

Theorem 1. *For any $\mathcal{V} \in \mathcal{F}$, the sequence $(\mathcal{A}_n = \Phi^n(\mathcal{V}))_{n\in\mathbb{N}}$ converges in the metric d to the Pareto frontier $\mathcal{V}^* \in \mathcal{F}$, which is the unique fixed point of the operator Φ, that is, the unique solution of $\Phi(\mathcal{V}) = \mathcal{V}$.*

We can then show that $\mathcal{V}^*$ is indeed the optimal set $\mathcal{V}^\infty$ that we are looking for.

Theorem 2. *We have that $\mathcal{V}^\infty = \mathcal{V}^*$.*

4.3. Optimal Strategies: Existence and Structure

For a Pareto frontier $\mathcal{V} \in \mathcal{F}$, one can define a one-to-one function from some compact parameter set $\mathcal{P}$ to $\mathcal{V}$.

Such a function parameterizes the Pareto frontier. We present one such parameterization that will be used later in our approximation procedure. Define the set

$$\mathcal{P} \triangleq \cup_{k=1}^{K} \{(p_1, \cdots, p_{k-1}, 0, p_k, \cdots, p_{K-1});\ p_r \in [0,1] \text{ for all } r = 1, \cdots, K-1\}.$$

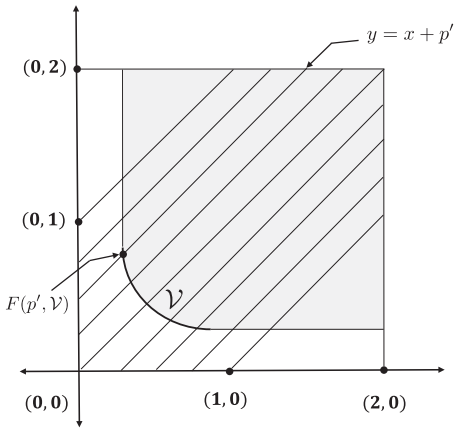
The term $\mathcal{P}$ is thus the union of K, $K-1$ dimensional faces of the hypercube $[0,1]^K$, where each face is obtained by pinning the value along one dimension to zero. For instance for $K=2$, we have $\mathcal{P} = [0,1] \times \{0\} \cup \{0\} \times [0,1]$, that is, the union of the line segment joining (0, 0) and (0, 1) and the segment joining (0,0) and (1, 0). Now consider the function $\mathbf{F} : \mathcal{P} \times \mathcal{F} \to \mathbb{R}^K$, where we define

$$\mathbf{F}(\mathbf{p}, \mathcal{V}) = \arg\min_{\mathbf{x}} t \tag{9}$$
$$\text{s.t. } \mathbf{x} = t\mathbf{1} + \mathbf{p},\ t \in \mathbb{R},$$
$$\mathbf{x} \succeq \mathbf{u},\ \mathbf{u} \in \mathcal{V}.$$

$\mathbf{F}(\mathbf{p}, \mathcal{V})$ is essentially the component-wise smallest point of intersection of the line $\mathbf{x} = t\mathbf{1} + \mathbf{p}$ (for a fixed $\mathbf{p}$) with the upset of $\mathcal{V}$ in $[0,2]^K$. In $\mathbb{R}^2$, this is simply the family of lines $y = x + p'$ where $p' = p_2 - p_1$, for $p' \in [-1,1]$ (Figure 5).

Then for a given $\mathcal{V}$, the function $\mathbf{F}(.,\mathcal{V}) : \mathcal{P} \to up(\mathcal{V})$ defines a map such that for every point $\mathbf{u}$ on $\mathcal{V}$, there is a unique $\mathbf{p} \in \mathcal{P}$ that maps *only* to that point. This is the $\mathbf{p}$ such that the line $\mathbf{x} = t\mathbf{1} + \mathbf{p}$ intersects $\mathcal{V}$ at $\mathbf{u}$ (if the line intersects $\mathcal{V}$ at two or more points, then one of those points is dominated by the other(s), which is a contradiction). For some values of $\mathbf{p}$, the line $\mathbf{x} = t\mathbf{1} + \mathbf{p}$ may not intersect $\mathcal{V}$, but it will definitely intersect the upset of $\mathcal{V}$ in $[0,2]^K$, which is why in (9), we optimize over $\mathbf{x}$ that dominate $\mathbf{u} \in \mathcal{V}$, rather than directly optimizing over $\mathbf{u} \in \mathcal{V}$.

Figure 5. Parameterization of $\mathcal{V}$



We can now express the DP operator in the form of such a parametrization. Assume that $\mathcal{V}^*$ is such that $\mathcal{V}^* = \Phi(\mathcal{V}^*)$. Then for $\mathbf{p} \in \mathcal{P}$, one can choose $\boldsymbol{\alpha}(\mathbf{p}) \in \Delta(A)$ and $\mathbf{q}(b, \mathbf{p}) \in \mathcal{P}$ for each $b \in B$ such that for $k \in \{1, \cdots, K\}$,

$$F_k(\mathbf{p}, \mathcal{V}^*) = \max_{b \in B} \left\{ \sum_{a \in A} \alpha_a(\mathbf{p}) r_k(a,b) + \beta F_k(\mathbf{q}(b, \mathbf{p}), \mathcal{V}^*) \right\}. \tag{10}$$

Then we have the following result.

Theorem 3. *For any $\mathbf{p}_1 \in \mathcal{P}$, the upper bound $\mathbf{F}(\mathbf{p}_1, \mathcal{V}^*) \in \mathcal{V}^*$ on losses is guaranteed by Alice in the infinite horizon game by first choosing action $a_1 \in A$ with probability $\alpha_{a_1}(\mathbf{p}_1)$. Then if Bob chooses an action $b_1 \in B$, the optimal guarantees to choose from the second step onward are then $\beta\mathbf{F}(\mathbf{p}_2, \mathcal{V}^*)$ in $\beta\mathcal{V}^*$, where $\mathbf{p}_2 = \mathbf{q}(b_1, \mathbf{p}_1)$, which can be guaranteed by Alice by choosing action $a_2 \in A$ with probability $\alpha_{a_2}(\mathbf{p}_2)$, and so on.*

This implies that $\mathcal{P}$ can be thought of as a compact *state space* for the strategy. In the remainder of the paper, however, we will refer to these states as *modes* to distinguish them from the classical notion of an exogenously defined state in, for example, MDPs. Each mode is associated with an immediate optimal randomized action and a transition rule that depends on the observed action of Bob. To attain a point in $\mathcal{V}^*$, Alice starts with the corresponding mode, plays the associated randomized action, transitions into another mode depending on Bob's observed action as dictated by the rule, then plays the randomized action associated with the new mode, and so on. In particular, the strategy does not depend on the past actions of Alice, and it depends on the past actions of Bob only through this information state, that is, the mode, that Alice keeps track of. It is interesting to note that unlike in stochastic games or MDPs (Puterman 2014), the state transitions are not exogenously defined, but they are endogenously specified by the dynamic programming operator.

5. Approximation

In general, except for simple examples (such an example is presented in Section EC.8 in the online appendix), it is difficult to analytically compute $\mathcal{V}^*$ and the optimal strategies $\{(\boldsymbol{\alpha}(\mathbf{p}), q(b, \mathbf{p})) : \mathbf{p} \in \mathcal{P}\}$ that satisfy (10) by simply using the fixed point relation. Hence, we now propose an approximate dynamic programming procedure to approximate the optimal Pareto frontier and devise approximately optimal strategies. To do so, we first define an appropriate finitely parameterized approximation of any Pareto frontier where one gets an increasingly finer approximation as the size of the parameter space increases.

Consider the following approximation scheme for a Pareto frontier $\mathcal{V} \in \mathcal{F}$. For a fixed positive integer N, define the set

$$\mathcal{P}_N = \cup_{k=1}^{K} \Big\{ (p_1, \cdots, p_{k-1}, 0, p_k, \cdots, p_{K-1}); \; p_r \in \Big\{0, \frac{1}{N}, \frac{2}{N}, \cdots, \frac{N-1}{N}, 1\Big\} \; \forall r = 1, \cdots, K-1 \Big\}. \tag{11}$$

In words, $\mathcal{P}_N$ is obtained by approximating each of the $K, K-1$ dimensional faces in $\mathcal{P}$ by a uniformly distributed grid of $(N+1)^{K-1}$ points. The number of distinct points in this set is

$$H(K,N) \triangleq (N+1)^K - N^K. \tag{12}$$

Next, define the approximation operator to be

$$\Gamma_N(\mathcal{V}) = \Lambda(\mathbf{ch}(\{\mathbf{F}(\mathbf{p}, \mathcal{V}) : \mathbf{p} \in \mathcal{P}_N\})), \tag{13}$$

where $\mathbf{F}(\mathbf{p}, \mathcal{V})$ was defined in (9). Here **ch** denotes the closed convex hull of a set. Thus, $\Gamma_N(\mathcal{V}) \in \mathcal{F}$ is the lower Pareto frontier of a convex polytope, and it has at most $H(K,N)$ vertices, where each vertex is the point of intersection of the line $\mathbf{x} = t\mathbf{1} + \mathbf{p}$ with the upset of $\mathcal{V}$ for some $\mathbf{p} \in \mathcal{P}_N$. The following approximation guarantee is instrumental in driving our results.

Proposition 3. *Consider a $\mathcal{V} \in \mathcal{F}$. Then*

$$e(\Gamma_N(\mathcal{V}), \mathcal{V}) = 0 \text{ and } e(\mathcal{V}, \Gamma_N(\mathcal{V})) \le \frac{1}{N},$$

and hence

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

Next, we can express the compound operator $\Gamma_N \circ \Phi$ via a set of explicit optimization problems as in (9) that only take $\mathcal{V}$ as input:

$$\begin{aligned} \mathbf{F}(\mathbf{p}, \Phi(\mathcal{V})) = \arg\min_{\mathbf{x}} t \qquad & (14) \\ \text{s.t. } \mathbf{x} = t\mathbf{1} + \mathbf{p}, \;\; t \in \mathbb{R}, \\ \mathbf{x} \succeq \sum_{a \in A} \alpha_a \mathbf{r}(a,b) + \beta \mathbf{Q}(b) \;\; \forall b \in B, \\ \boldsymbol{\alpha} \in \Delta(A), \; \mathbf{Q}(b) \in \mathcal{V} \;\; \forall b \in B. \end{aligned}$$

If $\mathcal{V} \in \mathcal{F}$ is the lower Pareto frontier of a convex polytope, then this is a linear program, and furthermore, $\Gamma_N \circ \Phi(\mathcal{V})$ is also the lower Pareto frontier of a convex polytope. We then we have the following result.

Proposition 4. *Let $\mathcal{G}_0 = \{\mathbf{0}\}$ and let $\mathcal{G}_n = (\Gamma_N \circ \Phi)^n(\mathcal{G}_0)$. Then*

$$e(\mathcal{G}_n, \mathcal{V}^*) \le \beta^n \text{ and } e(\mathcal{V}^*, \mathcal{G}_n) \le \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right) + \beta^n. \tag{15}$$

Thus,

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

Hence, for any ϵ, there is a pair (N, n) such that $d(\mathcal{V}^*, \mathcal{G}_n) \le \epsilon$. This result implies an iterative procedure for approximating $\mathcal{V}^*$ by successively applying the compound operator $\Gamma_N \circ \Phi$ to $\mathcal{G}_0$, by solving the linear program in (14) for each $\mathbf{p} \in \mathcal{P}_N$ at each step. Because $\mathcal{G}_n$ is a lower Pareto frontier of a convex polytope, with at most $H(K,N)$ vertices for each n, the size of these linear programs remain the same throughout. More details on solving these programs can be found in Section EC.3 in the online appendix.

The fact that $e(\mathcal{G}_n, \mathcal{V}^*) \le \beta^n$ implies that $\mathcal{G}_n$ β^n-dominates $\mathcal{V}^*$ for all n, and thus the optimal upper bounds in $\mathcal{V}^*$ cannot be larger than in $\mathcal{G}_n + \beta^n\mathbf{1}$. Thus, as n gets larger, the set $\mathcal{G}_n + \beta^n\mathbf{1}$ approaches $\mathcal{V}^*$ "from above," and in the limit, ends up within a $1/(N(1-\beta))$ distance of $\mathcal{V}^*$.

5.1. Extracting an Approximately Optimal Strategy

From $\mathcal{G}_n$, one can also extract an approximately optimal strategy π_n in the infinite horizon game. Suppose $\boldsymbol{\alpha}^*(\mathbf{p})$ and $\mathbf{Q}^*(b, \mathbf{p})$ for $b \in B$ are the optimal values that solve the program (14) to compute $\mathbf{F}(\mathbf{p}, \Phi(\mathcal{G}_n))$ for different $\mathbf{p} \in \mathcal{P}_N$. Then these define an approximately optimal strategy in the following class.

Definition 8. A *$H(K,N)$–mode stationary strategy* π is a mapping from each $\mathbf{p} \in \mathcal{P}_N$ to the following pair of elements.

1. *A probability distribution over actions* $\boldsymbol{\alpha}(\mathbf{p}) \in \Delta(A)$, and
2. *A transition rule* $(\mathbf{q}_1(b, \mathbf{p}), \cdots, \mathbf{q}_K(b, \mathbf{p}), \mathbf{z}(b, \mathbf{p}))$, where for all $b \in B$, $\mathbf{q}_k(b, \mathbf{p}) \in \mathcal{P}_N$ for all $k = 1, \cdots, K$ and $\mathbf{z}(b, \mathbf{p}) \in \Delta^K$.

Here, Δ^K is the unit simplex in $\mathbb{R}^K$. The interpretation is as follows. One starts with some initial *mode*, that is, a value of $\mathbf{p} \in \mathcal{P}_N$. Then at any step, if the current mode is $\mathbf{p}$, then Alice first chooses action $a \in A$ with probability $\alpha_a(\mathbf{p})$. Then if Bob plays action $b \in B$, Alice samples the new mode to be $\mathbf{q}_k(b, \mathbf{p})$ with probability $z_k(b, \mathbf{p})$ for each k, and after having sampled a new mode, plays accordingly thereafter.

Now, $\boldsymbol{\alpha}^*(\mathbf{p})$ defines $\boldsymbol{\alpha}(\mathbf{p})$ in π_n, and $(\mathbf{q}_1(b, \mathbf{p}), \cdots, \mathbf{q}_K(b, \mathbf{p}), \mathbf{z}(b, \mathbf{p}))$ are defined such that they satisfy

$$\mathbf{Q}^*(b, p) = \sum_{k'=1}^{K} z_{k'}(b, \mathbf{p}) \mathbf{F}(\mathbf{q}_{k'}(b, \mathbf{p}), \mathcal{G}_n). \tag{16}$$

These $(\mathbf{q}_1(b, \mathbf{p}), \cdots, \mathbf{q}_K(b, \mathbf{p}), \mathbf{z}(b, \mathbf{p}))$ are directly obtained as the output of the linear program; see Section EC.3 in

the online appendix. The interpretation is as follows. If $\mathcal{V}$ is the lower Pareto frontier of a convex polytope with each vertex lying on the line $\mathbf{x} = t\mathbf{1} + \mathbf{p}$ for some $\mathbf{p} \in \mathcal{P}_N$, $\mathbf{Q}^*(b, \mathbf{p})$ for each $b \in B$ that results from solving (14) will lie on one of the faces of this Pareto frontier. Thus, $\mathbf{Q}^*(b, \mathbf{p})$ can be expressed as a convex combination of (at most K) extreme points of the face as expressed in (16).

Let $\mathcal{V}^{\pi_n}$ be the corresponding Pareto frontier that is attained by the strategy π_n (each point on this frontier is guaranteed by choosing different possible initial randomizations over the $H(K,N)$ modes). In Section EC.4 in the online appendix, we discuss how this "policy evaluation" can be performed by solving a linear program. Simply from the definition of $\mathcal{V}^*$ as the optimal frontier, we know that $\mathcal{V}^{\pi_n}$ dominates $\mathcal{V}^*$, that is, $e(\mathcal{V}^{\pi_n}, \mathcal{V}^*) = 0$. However, we can further show the following.

Proposition 5. The following bound holds.

$$d(\mathcal{V}^{\pi_n}, \mathcal{V}^*) \le \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right) + 2\beta^n + \frac{1}{N}\left(\frac{2-\beta^n-\beta^{n+1}}{(1-\beta)^2}\right). \tag{17}$$

Thus, an approximately optimal strategy can be obtained by choosing an appropriate (N, n).

Remark 1. For a fixed (N, n), to approximate the optimal frontier, the procedure needs to solve $nH(K, N)$ linear programs to give the corresponding error bound in Proposition 4. In our implementation described in Section EC.3 in the online appendix, each linear program is composed of $mH(K,N) + l + 1$ variables and $Km + K + 1$ constraints. One can focus on two terms in the approximation error separately: the first term is the quantization error, which is bounded by $\frac{1}{N(1-\beta)}$, and the second is the iteration error, which is bounded by β^n. The second term is benign because it decays exponentially in n. The first term is dominant and requires $N = \frac{1}{(1-\beta)\epsilon}$ to achieve an error of ϵ. To find an ϵ-optimal strategy, we require $N \approx \frac{1}{(1-\beta)^2\epsilon}$. Thus, for fixed values of β not too close to one, the N required to obtain a good approximation is not too large. The main concern, however, is that $H(K,N)$ grows exponentially in the dimension K, and hence the computation is expected to be prohibitive when K is large. In Section 6, we propose and evaluate a heuristic approach to get around this difficulty at the cost of loss in optimality.

6. Optimal Finite-Mode Policies for Larger K

In this section, we propose a different approach to designing good policies with a small number of modes in settings where K is large when the computation of near-optimal policies discussed in Section 5 becomes prohibitive.

To design a good policy with a small number of modes, the modes must be carefully chosen. The following natural question guides our approach in this section: given an instance of a vector repeated game with discounted losses and a finite budget of modes, how do we design the "best" stationary policy within this budget? The first step is to specify what we mean by "best." Motivated by our goal of regret minimization in repeated decision making, we aim to minimize the losses along the ray $\{\mathbf{x} = t\mathbf{1}; t \in \mathbb{R}\}$; that is, we wish to minimize t such that the losses on all dimensions are guaranteed to be at most t, irrespective of the actions of the adversary.

Let such a policy start from mode 0, and let $\mathbf{v}_0 \in \{\mathbf{x} = t\mathbf{1}; t \in \mathbb{R}\}$ denote vector of losses guaranteed by this policy starting from mode 0, where we wish to minimize $\mathbf{v}_0$. Let there be M additional modes allowed by our budget, denoted by $i = 1, \cdots, M$, making a total of $M + 1$ modes. Let $\mathcal{M}$ denote the set of all modes. Let $\mathbf{v}_i$ be the vector of losses guaranteed by the policy starting from mode i. Associated with each mode $i = 0, \cdots, M$, let $\boldsymbol{\alpha}_i \in \Delta(A)$ denote the probability distribution over immediate actions and let $(\mathbf{z}_i(b) \in \Delta(\mathcal{M}); b \in B)$ denote the randomized transition rule to other modes as a function of the adversary's action. Then the problem of minimizing $\mathbf{v}_0$ can be written as the following optimization problem:

$$\min_{\boldsymbol{\alpha}, \mathbf{z}, \mathbf{v}, t} \; t \tag{18a}$$

$$\text{s.t. } \mathbf{v}_0 = t\mathbf{1}, \quad t \in \mathbb{R}, \tag{18b}$$

$$\mathbf{v}_i \succeq \sum_{a \in A} \alpha_{i,a}\mathbf{r}(a,b) + \beta \sum_{j \in \mathcal{M}} z_{i,j}(b)\mathbf{v}_j, \text{ for all } b \in B \text{ and } i \in \mathcal{M}. \tag{18c}$$

$$\boldsymbol{\alpha}_i \in \Delta(A), \; \mathbf{z}_i(b) \in \Delta(\mathcal{M}); \text{ for all } b \in B \text{ and } i \in \mathcal{M}. \tag{18d}$$

Here, the objective and (18b) express the fact that we are minimizing losses along the said ray. Equation (18c) captures the Bellman one-step optimality conditions, which express the fact that the vector guarantees $(\mathbf{v}_i)$ are feasible under the stationary policy that associates $\boldsymbol{\alpha}$ and $\mathbf{z}$ with the different modes (see also Section EC.4 in the online appendix).

The optimization problem defined previously is a quadratically constrained linear program (QCLP) because of the bilinear constraints in (18c), which is known to be nonconvex. The size of the problem grows as $O(KMm)$, which has a significantly milder dependence on K compared with the approximation approach of Section 5. Although this problem is nonconvex, a wide range of optimization algorithms has been developed over the years that efficiently solve even large-scale instances of such programs to local optimality. For example, in a similar spirit as in our case, such

QCLPs arise in the context of finding optimal finite-state controllers in partially observable Markov decision processes (POMDPs). Numerical solutions to these QCLPs have been shown to yield significantly better policies than those obtained via other heuristic approaches (Amato et al. 2006, 2010). In our case, we analogously find that in our numerical evaluations in the context of regret minimization discussed in Section 7.2, the solutions to our QCLP define finite-mode online learning algorithms that provide guarantees that are significantly better than those provided by Hedge.

7. Numerical Experiments

In this section, we present a numerical evaluation of our approaches discussed in Sections 5 and 6 in the context of adversarial online learning.

7.1. Designing Near-Optimal Strategies for Expert Selection with Binary Losses

First, we illustrate our approximation scheme discussed in Section 5 by applying it to the well-known problem of regret minimization in expert selection with binary losses. We design, to the best of our knowledge, the first-known provably near-optimal algorithms for the case of $K = 2$ experts and discounted losses and show that these algorithms guarantee significantly smaller upper bounds on the regret than existing algorithms in adversarial online learning.

The problem of expert selection with binary losses is described as follows. There are K experts who give Alice recommendations for a decision-making task: say predicting which route will be the quickest to commute to work the next day. On each day, Alice decides to act on the recommendation made by one of the experts. The experts' recommendations may be correct or wrong, and if Alice acts on an incorrect recommendation, she bears a loss of one; otherwise, she does not incur any loss. Each day, any set of experts may be correct, whereas others are wrong. We omit the possibilities that all experts are correct, or all experts are wrong, because it is wasteful for the adversary to choose these options. For $K = 2$, this model can be represented by the matrix shown in Figure 6. The rows correspond to the choice made by Alice and the columns correspond to the different possibilities for the outcomes on each day. The

Figure 6. Possible Loss Scenarios with $K = 2$ Experts

	1	2
Expert 1	1	0
Expert 2	0	1

Figure 7. Single-Stage Regret w.r.t. Experts 1 and 2

	1	2
Expert 1	$(0, 1)$	$(0, -1)$
Expert 2	$(-1, 0)$	$(1, 0)$

matrix of single-stage regrets, in this case, is shown in Figure 7.

In this case, because K is small, the optimal frontier of regrets can be efficiently computed with high accuracy.

Figure 8 shows the computed approximately optimal Pareto frontiers of regret for a range of values of β. (N, n) is chosen in each case so that the error in the approximation of $(1-\beta)\mathcal{V}^*(\beta)$ (i.e., the optimal Pareto frontier of *average* discounted regrets) is at most 0.06 (Proposition 4). The lower and upper bounds on the discounted average optimal regret are plotted in Figure 9; the lower bounds result from our theoretical guarantees in Proposition 4, and the upper bounds result from the evaluation of our policies as shown in Section EC.4 in the online appendix.

In the figure, we also plot the theoretical upper bound on regret guaranteed by two other policies: (a) the well-known exponentially weighted average forecaster, also known as Hedge and (b) the optimal algorithm given by Gravin et al. (2016) (which we will refer to as GPS) for the two-experts problem with a geometrically distributed time horizon.[8] Hedge guarantees an upper bound of $\sqrt{\log K(1-\beta)/(2(1+\beta))}$ on the expected average discounted regret for the K experts problem, which is the best-known bound for this problem (Cesa-Bianchi and Lugosi 2003). Hedge is defined in Section EC.5.1 in the online appendix. For the two-experts problem, GPS

Figure 8. (Color online) Approximations of the Optimal Frontier $(1-\beta)\mathcal{V}^*(\beta)$ for Different β Values and the Associated (N, n)

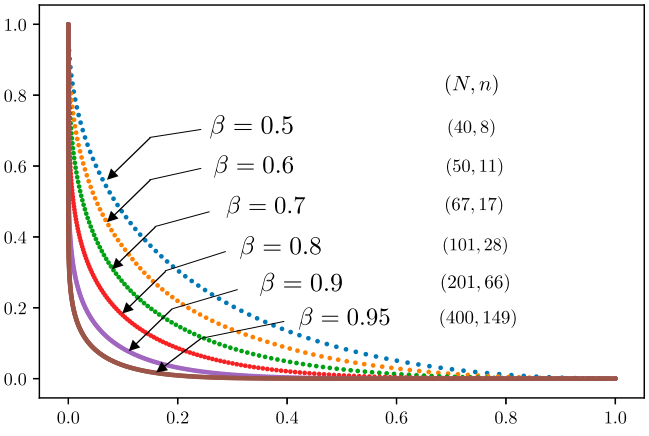
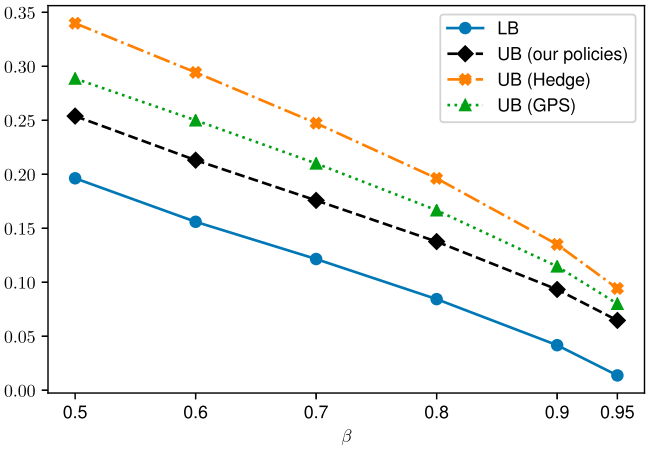
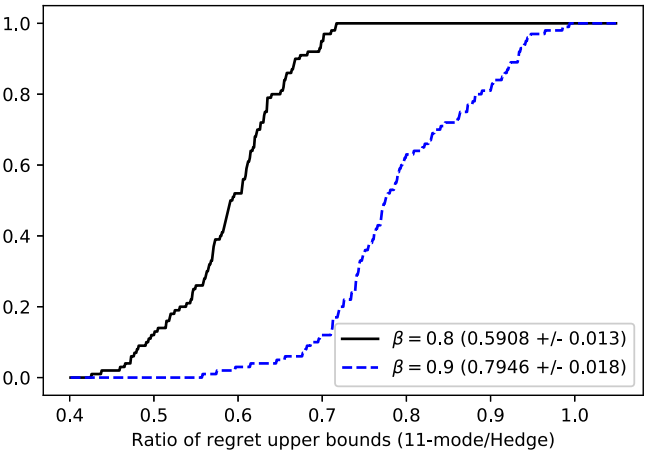


Figure 9. (Color online) Upper Bounds on the Optimal Average Discounted Regret Achieved by the Different Policies Plotted as a Function of the Discount Factor β



Note. Also plotted is the theoretical lower bound on regret.

Figure 10. (Color online) Empirical c.d.f. of the Ratio of the Upper Bounds Guaranteed by Our 11-Mode Policy and by Hedge Across 100 Instances



guarantees an upper bound of $(1/2)\times\sqrt{(1-\beta)/(1+\beta)}$ on the expected average discounted regret (see Endnote 10 in the online appendix). The GPS algorithm is presented in Section EC.5.2 in the online appendix. Both these algorithms achieve significantly higher upper bounds on the regret than those guaranteed by our policies. Additionally, in Section EC.7 in the online appendix, for $\beta = 0.8$, we design an adversary that induces both Hedge and GPS to exceed the upper bound on the regret guaranteed by our policies, thereby demonstrating their suboptimality.

7.2. Numerical Evaluation of Optimal Finite-Mode Policies

In this section, we test the approach discussed in Section 6 for designing optimal finite-mode policies in higher-dimensional settings. We consider a set of randomly generated repeated decision-making instances with $l = 10$ actions for the decision maker and $m = 10$ actions for the adversary. Each instance is generated by drawing losses corresponding to each pair of actions uniformly in the set [0, 1]. We generate 100 such instances. In each instance, the corresponding vector-valued game of single-stage regrets has losses with $K = l = 10$ dimensions for each pair of actions. Each dimension tracks the additional regret relative to playing each of the $l = 10$ actions. We choose $M = 10$, resulting in a budget of $M + 1 = 11$ modes. We consider $\beta \in \{0.8, 0.9\}$. We use the open-source nonlinear optimization software APOPT available via Gekko, a Python package and server for optimization (Beal et al. 2018), to solve our QCLP.

In Figure 10, we plot the empirical cumulative distribution function (c.d.f.) of the ratio of the upper bound guaranteed by our 11-mode stationary policy resulting from the solution of the QCLP and that guaranteed by Hedge, across the 100 instances, for $\beta \in \{0.8, 0.9\}$. The mean ratio and its standard error are presented in the legend. When losses are in $[0, a]$, the optimally tuned Hedge algorithm guarantees an average discounted regret of $a \times \sqrt{\log K(1-\beta)/(2(1+\beta))}$; thus, the regret upper bound guaranteed by Hedge in each of our instances depends on the maximal loss that the decision maker can incur in that instance. We find that in all instances and both the settings, the upper bound on the regret guaranteed by our 11-mode stationary policy is smaller than that guaranteed by Hedge.

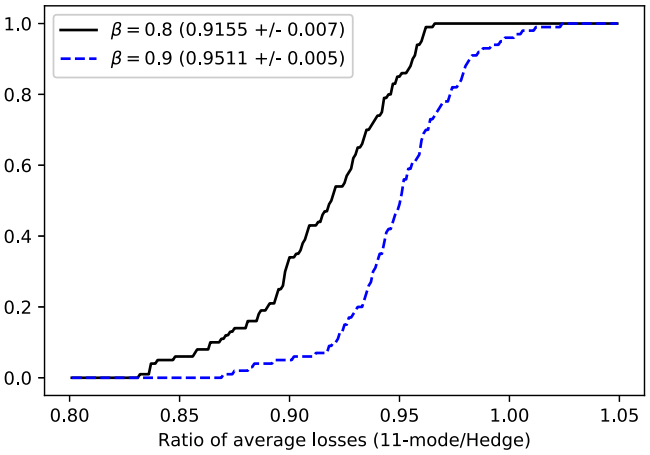
Additionally, for each of the 100 instances, we compare the empirical performance of our 11-mode stationary policy and of Hedge against a set of 1,000 sequences of adversary's actions chosen uniformly at random (for $T = 50$). The empirical c.d.f. of the ratio of average losses incurred by our 11-mode policy and by Hedge across the 100 instances is presented in Figure 11. The mean ratio and its standard error are presented in the legend. We observe that our algorithms yield better empirical performance than Hedge across almost all instances in the two settings.

We thus conclude that, although our finite-mode policy is expected to only crudely leverage detailed information about the instance, it can potentially lead to algorithms with significantly better performance than off-the-shelf algorithms in adversarial online learning.

8. Discussion and Conclusion

We presented a novel approximate dynamic programming approach to approximate the set of minimal guarantees that a player can achieve in a discounted repeated game with vector losses and finite action sets. We showed that this optimal set is the fixed point of a contractive dynamic programming operator, and it is

Figure 11. (Color online) Empirical c.d.f. of the Ratio of the Average Losses Incurred by Our 11-Mode Policy and by Hedge Across 100 Instances Against 1,000 Randomly Generated Adversary Action Sequences



the Pareto frontier of a convex and closed set. We also established the structure of the optimal strategies that achieve the different points on this set. We then proposed an iterative procedure to approximately compute this set and find approximately optimal strategies. The main motivating application of this machinery is to the problem of regret minimization in repeated games within the framework of adversarial online learning. We illustrated our approach by designing provably approximately optimal strategies for prediction using expert advice with binary losses, for $K = 2$ experts. In the process, we demonstrated the suboptimality of well-known off-the-shelf adversarial online learning algorithms. Although our approximation approach can become computationally intensive in higher dimensions, we proposed and tested an approach to design well-performing finite-mode policies in such cases based on a QCLP formulation of the problem of finding the optimal stationary strategy with a finite budget of modes. We showed that such policies can result in better performance guarantees compared with existing adversarial online learning algorithms.

It is important to note that adversarial online learning algorithms like Hedge are able to deal with more general adversaries that inflict arbitrary losses lying in a bounded, continuous set, for example, in [0, 1]. This power comes at the cost of relatively conservative guarantees on the regret. A drawback of our approach is that we cannot explicitly handle continuous action spaces for the adversary; this is indeed an interesting and important direction for future research. The application of our current approach in such cases would require clustering the vectors of losses inflicted by the adversary based on past data. On the one hand, this approximation may lead to performance loss, and one may prefer algorithms like Hedge in such situations. On the other hand, our approach allows one to exploit the structure in the adversary's choice of losses, for example, based on past data, one may conclude that the losses can indeed be effectively clustered, although the choice of the cluster is best modeled as adversarial. In these cases, algorithms like Hedge may be too cautious, and one may wish to incorporate such information in the decision-making process while still seeking robust performance guarantees. Our approach provides a way of doing so by appropriately modifying the set of actions available to the adversary, thus bridging the gap between an overly pessimistic adversarial view of the environment and a fully nonadversarial stochastic view of the environment. As we saw from our numerical evaluations in Section 7, our approach may lead to better performance guarantees than off-the-shelf algorithms like Hedge in these cases. This is also an important distinguishing point compared with the related contribution of Gravin et al. (2016). Apart from the difference in the regret minimization objective considered in that paper relative to ours (as we discuss in detail in Section EC.6 in the online appendix), their focus is on the Experts setting where the loss vectors are in $\{0,1\}^K$. Our model, in contrast, incorporates the flexibility of specifying the loss vectors available to the adversary as an arbitrary finite set.

Finally, we mention that the extension of this approach to the case of long-run average losses in infinitely repeated games appears to be less straightforward, despite the fact that average cost dynamic programming for standard dynamic optimization problems like MDPs is quite well understood. Such an extension, along with the extension to continuous losses, would fill a significant part of the remaining gap in viewing the approximate dynamic programming paradigm as a methodical approach to designing adversarial online learning algorithms.

Endnotes

[1] An extension to general convex polytopes of the form $\{\mathbf{x} \in \mathbb{R}^K : A\mathbf{x} \le \mathbf{b}\}$ follows by considering games with appropriate linear transformations of the vector losses. The corresponding results in the single-dimensional case, that is, where $A\mathbf{x} \in \mathbb{R}$ (i.e., $K = 1$), are well known, and we review them in Section EC.9.1 in the online appendix.

[2] The constraint on the adversary's space of actions can result from certain side information that Alice has about the environment. For example, it may be known that two experts always give opposite advice.

[3] A set-valued operator $\mathcal{B}$ is monotone if $A \subseteq A'$ implies that $\mathcal{B}(A) \subseteq \mathcal{B}(A')$ (Abreu et al. 1990).

[4] A deterministic, oblivious adversary is a standard assumption in regret-minimization literature (Auer et al. 2002, Cesa-Bianchi and Lugosi 2006, Bubeck and Cesa-Bianchi 2012). This makes sense when "nature" is modeled as an adversary in an application, which would be the case for instance in weather forecasting. Also see chapter 3 in

Bubeck and Cesa-Bianchi (2012) for a discussion of different adversary models and their implications on different definitions of regret.

[5] One direction is clear because a Pareto frontier in $\mathcal{F}$ is the lower Pareto frontier of its upset, which is closed and convex. The other direction follows from the observation that the upset of the lower Pareto frontier of a set is the upset of the set itself and the upset of a closed and convex set is closed and convex.

[6] Because the upsets of the sets in $\mathcal{F}$ are compact, the sup and the inf in the definition of the Hausdorff distance can be replaced by min and max, respectively.

[7] The inf in the definition can be replaced by a min because $e(\mathcal{U},\mathcal{V})$ can be equivalently defined as $\inf\{\epsilon \geq 0 : \forall \mathbf{u} \in up(\mathcal{U}),\ \exists \mathbf{v} \in up(\mathcal{V})\ \text{s.t.}\ \mathbf{v} \preceq \mathbf{u} + \epsilon \mathbf{1}\}$, and $up(\mathcal{U})$ and $up(\mathcal{V})$ are compact sets for any $\mathcal{U},\mathcal{V} \in \mathcal{F}$.

[8] This model and its relation to our model of discounted losses is discussed in Section EC.6 in the online appendix.

References

Abernethy J, Warmuth MK, Yellin J (2008) Optimal strategies from random walks. Servedio R, Zhang T, eds. *Proc. 21st Annual Conf. on Learn. Theory* (Association for Computational Learning, Mountain View, CA), 437–446.

Abreu D, Pearce D, Stacchetti E (1986) Optimal cartel equilibria with imperfect monitoring. *J. Econom. Theory* 39(1):251–269.

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

Amato C, Bernstein DS, Zilberstein S (2006) Solving POMDPs using quadratically constrained linear programs. Weiss G, Stone P, eds. *Proc. 5th Internat. Joint Conf. Autonomous Agents Multiagent Systems* (Association for Computing Machinery, New York), 341–343.

Amato C, Bernstein DS, Zilberstein S (2010) Optimizing fixed-size stochastic controllers for pomdps and decentralized pomdps. *Autonomic Agent Multi Agent Systems* 21(3):293–320.

Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (2002) The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* 32(1): 48–77.

Aumann RJ, Maschler M, Stearns RE (1995) *Repeated Games with Incomplete Information* (MIT Press, Cambridge, MA).

Bartlett PL, Koolen WM, Malek A, Takimoto E, Warmuth MK (2015) Minimax fixed-design linear regression. Grünwald P, Hazan E, eds. *Conf. Learn. Theory* (Association for Computational Learning, Mountain View, CA), 437–446.

Bayraktar E, Ekren I, Zhang Y (2020) On the asymptotic optimality of the comb strategy for prediction with expert advice. *Ann. Appl. Probability* 30(6):2517–2546.

Beal LDR, Hill DC, Martin RA, Hedengren JD (2018) Gekko optimization suite. *Processes (Basel)* 6(8):106.

Bertsekas DP (2005) *Dynamic Programming and Optimal Control*, vol 1 (Athena Scientific).

Bertsekas DP (2012) *Dynamic Programming and Optimal Control*, vol 2 (Athena Scientific).

Blackwell D (1956) An analog of the minimax theorem for vector payoffs. *Pacific J. Math.* 6(1):1–8.

Blum A, Mansour Y (2007) From external to internal regret. *J. Machine Learn. Res.* 8(Jun):1307–1324.

Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations Trends Machine Learn.* 5(1):1–122.

Cesa-Bianchi N, Lugosi G (2003) Potential-based algorithms in on-line prediction and game theory. *Machine Learn.* 51(3):239–261.

Cesa-Bianchi N, Lugosi G (2006) *Prediction, Learning, and Games* (Cambridge University Press, Cambridge, UK).

Cesa-Bianchi N, Freund Y, Haussler D, Helmbold DP, Schapire RE, Warmuth MK (1997) How to use expert advice. *J. ACM* 44(3): 427–485.

Chernov A, Zhdanov F (2010) Prediction with expert advice under discounted loss. *Algorithmic Learning Theory* (Springer, Berlin), 255–269.

Cover TM (1966) Behavior of sequential predictors of binary sequences. Technical report, Stanford University, Stanford, CA.

Foster DP, Vohra RV (1997) Calibrated learning and correlated equilibrium. *Games Econom. Behav.* 21(1–2):40–55.

Freund Y, Schapire RE (1999) Adaptive game playing using multiplicative weights. *Games Econom. Behav.* 29(1–2):79–103.

Gravin N, Peres Y, Sivan B (2016) Toward optimal algorithms for prediction with expert advice. Kraughgamer R, ed. *Proc. 27th Annual ACM-SIAM Sympos. on Discrete Algorithms* (Society for Industrial and Applied Mathematics, Philadelphia), 528–547.

Hannan J (1957) Approximation to Bayes risk in repeated plays. Dresher M, Tucker AW, Wolfe P, eds. *Contributions to the Theory of Games*, vol. 3 (Princeton University Press, Princeton, NJ), 97–139.

Hazan E (2016) Introduction to online convex optimization. *Foundations Trends Optim.* 2(3-4):157–325.

Henrikson J (1999) Completeness and total boundedness of the Hausdorff metric. *MIT Undergraduate J. Math.* 1:69–80.

Koolen WM (2013) The Pareto regret frontier. Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY), 863–871.

Koolen WM, Malek A, Bartlett PL (2014) Efficient minimax strategies for square loss games. Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, eds. *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY), 3230–3238.

Koolen WM, Malek A, Bartlett PL, Abbasi Y (2015) Minimax time series prediction. Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, eds. *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY), 2557–2565.

Laraki R, Sorin S (2015) Advances in zero-sum dynamic games. *Handbook of Game Theory with Economic Applications*, vol. 4 (Elsevier, New York), 27–93.

Lehrer E (2003) Approachability in infinite dimensional spaces. *Internat. J. Game Theory* 31(2):253–268.

Littlestone N, Warmuth MK (1994) The weighted majority algorithm. *Inform. Comput.* 108(2):212–261.

Luo H, Schapire R (2014) Toward minimax online learning with unknown time horizon. Xing Ep, Jebara T, eds. *Internat. Conf. Machine Learn. Proc. 31st Internat. Conf. on Machine Learn.*, 226–234.

Milman E (2006) Approachable sets of vector payoffs in stochastic games. *Games Econom. Behav.* 56(1):135–147.

Munkres JR (2000) *Topology: A First Course* (Prentice Hall, Inc., Hoboken, NJ).

Perchet V (2011a) Approachability of convex sets in games with partial monitoring. *J. Optim. Theory Appl.* 149(3):665–677.

Perchet V (2011b) Internal regret with partial monitoring: Calibration-based optimal algorithms. *J. Machine Learn. Res.* 12(Jun):1893–1921.

Perchet V (2014) Approachability, regret and calibration: Implications and equivalences. *J. Dynamic Games* 1(2):181–254.

Perchet V, Quincampoix M (2014) On a unified framework for approachability with full or partial monitoring. *Math. Oper. Res.* 40(3):596–610.

Puterman ML (2014) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (John Wiley & Sons, Hoboken, NJ).

Shapley LS (1953) Stochastic games. *Proc. National Acad. Sci. USA* 39(10):1095.

Spinat X (2002) A necessary and sufficient condition for approachability. *Math. Oper. Res.* 27(1):31–44.

Stoltz G, Lugosi G (2005) Internal regret in on-line portfolio selection. *Machine Learn*. 59(1–2):125–159.

Vieille N (1992) Weak approachability. *Math. Oper. Res.* 17(4):781–791.

Vovk VG (1990) Aggregating strategies. Fulk M, Case J, eds. *Proc. 3rd Annual Workshop Comput. Learning Theory* (COLT '90) (Morgan Kaufmann Publishers Inc., San Francisco), 371–386.

Vijay Kamble is an assistant professor of information and decision sciences in the College of Business Administration at the University of Illinois Chicago with a courtesy affiliation with the Department of Computer Science. His research is centered on the design and optimization of online platforms and marketplaces, with a primary focus on learning and experimentation on these platforms.

Patrick Loiseau is a research scientist at Inria and a part-time professor of computer science at Ecole Polytechnique (Saclay, France). He is also currently the co-holder of the chair on "Explainable and Responsible AI" in the multidisciplinary institute in artificial intelligence at Grenoble Alpes. His research interests include game theory and statistical learning, with a particular interest in security, privacy, and fairness aspects and in applications to online platforms and algorithms.

Jean Walrand is professor emeritus and a professor of the Graduate School in the Department of EECS at the University of California, Berkeley. His research interests include stochastic processes, queuing theory, communication networks, game theory, machine learning applied to stochastic scheduling, and the economics of the Internet. He is a recipient of the Lanchester Prize, the Stephen O. Rice Prize, the IEEE Kobayashi Award, and the ACM Sigmetrics Achievement Award.