Electronic companion to "An Approximate Dynamic Programming Approach to Repeated Games with Vector Losses."

EC.1. Proof of all results.

Proof of Lemma 1. Consider the minimization problem:

$$\min_{\mathbf{x}\in\mathcal{S}} f(\mathbf{x}) = \sum_{k=1}^{K} x_k.$$

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

Proof of Proposition 1. In order to prove the result, we need the following set of results about the Hausdorff distance:

LEMMA EC.1. *a) h is a metric on the space of closed subsets of $\mathbb{R}^K$.*

b) Assume that $(\mathcal{A}_n)_{n\in\mathbb{N}}$ is a sequence of closed subsets of $[0,1]^K$. Then there is a subsequence $(\mathcal{A}_{n_k})_{k\in\mathbb{N}}$ that converges to some closet subset $\mathcal{A}$ of $[0,1]^K$.

c) If the sets $(\mathcal{A}_n)_{n\in\mathbb{N}}$ 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 property of the Hausdorff distance, and the compactness property of the space of closed subsets of a compact set under the Hausdorff metric; see [27, 37].

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

e) Let $\epsilon := h(\mathcal{A},\mathcal{B})$. Pick $\mathbf{x} \in up(\mathcal{A})$. Then $\mathbf{x} = \mathbf{y} + \mathbf{v}$ for some $\mathbf{y} \in \mathcal{A}$ and $v \succeq 0$. There is some $\mathbf{y}' \in \mathcal{B}$ with $\|\mathbf{y} - \mathbf{y}'\|_\infty \leq \epsilon$. Then $\mathbf{x}' = \min\{\mathbf{y}' + \mathbf{v}, \mathbf{1}\} \in up(\mathcal{B})$, where $\mathbf{1}$ is the vector of ones in $\mathbb{R}^K$,

i.e., $(1; k = 1, \cdots, K)$, and the minimization is component-wise. We claim that $\|\mathbf{x}' - \mathbf{x}\|_\infty \leq \epsilon$. If $\mathbf{y}' + \mathbf{v} \in [0,1]^K$, this is clear. Assume $y'_k + v_k > 1$. Then,

$$x'_k = 1 < y'_k + v_k \text{ and } x_k = y_k + v_k \leq 1.$$

Thus,

$$0 \leq x'_k - x_k < y'_k + v_k - y_k - v_k = y'_k - y_k.$$

Hence, $|x'_k - x_k| \leq |y'_k - y_k|$ for any k. Thus, one has $\|\mathbf{x}' - \mathbf{x}\|_\infty \leq \|\mathbf{y}' - \mathbf{y}\|_\infty \leq \epsilon$.

□

Now we can prove the proposition. First, to show that d is a metric on $\mathcal{F}$, we just need to show that if $h(up(\mathcal{V}), up(\mathcal{U})) = 0$ then $\mathcal{V} = \mathcal{U}$. The other properties (e.g., triangle inequality etc.) follow from the corresponding properties for the Hausdorff metric. Note that if $h(up(\mathcal{V}), up(\mathcal{U})) = 0$, then $up(\mathcal{V}) = up(\mathcal{U})$. Suppose that there is some $\mathbf{u} \in \mathcal{U}$ such that $\mathbf{u} \notin \mathcal{V}$. But since $\mathbf{u} \in up(\mathcal{V})$, we have $\mathbf{u} = \mathbf{v} + \mathbf{y}$ for some $\mathbf{v} \in \mathcal{V}$ and some $\mathbf{y} \succeq \mathbf{0}$. But since $h(up(\mathcal{V}), up(\mathcal{U})) = 0$, by the definition of the Hausdorff distance, for each $\epsilon > 0$, there is a point $\mathbf{u}_\epsilon$ in $up(\mathcal{U})$ such that $\mathbf{u}_\epsilon \preceq \mathbf{v} + \epsilon\mathbf{1}$. Consider a sequence $(\epsilon_n)_{n\in\mathbb{N}}$ such that $\epsilon_n \to 0$, and consider the corresponding sequence $(\mathbf{u}_{\epsilon_n})_{n\in\mathbb{N}}$. Now since $up(\mathcal{U})$ is compact, $(\mathbf{u}_{\epsilon_n})_{n\in\mathbb{N}}$ has a convergent subsequence that converges to some $\mathbf{u}^* \in up(\mathcal{U})$ such that $\mathbf{u}^* \preceq \mathbf{v} \preceq \mathbf{v} + \mathbf{y} = \mathbf{u}$, which contradicts the fact that $\mathbf{u} \in \mathcal{U}$. Thus $\mathcal{U} = \mathcal{V}$.

Next, we prove statement (b). From statement (b) and (c) of Lemma EC.1, the subsequence $(up(\mathcal{V}_{n_k}))_{k\in\mathbb{N}}$ converges to some convex set $\mathcal{A}$. But since $h(up(\mathcal{V}_{n_k}), up(\mathcal{A})) \leq h(up(\mathcal{V}_{n_k}), \mathcal{A})$, we have $up(\mathcal{A}) = \mathcal{A}$. And thus the subsequence $(\mathcal{V}_{n_k})_{k\in\mathbb{N}}$ converges to the Pareto frontier of $\mathcal{A}$, which is in $\mathcal{F}$.

Observe that it becomes clear from the above arguments 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. □

Proof of Proposition 2. Suppose that $\max(e(\mathcal{U},\mathcal{V}), e(\mathcal{V},\mathcal{U})) \leq \epsilon$. Consider a point $\mathbf{x} \in up(\mathcal{U})$ such that $\mathbf{x} = \mathbf{y} + \mathbf{v}$ where $\mathbf{y} \in \mathcal{U}$ and $\mathbf{v} \succeq 0$. Suppose that there is no $\mathbf{x}' \in up(\mathcal{V})$ such that $\|\mathbf{x} - \mathbf{x}'\|_\infty \leq \epsilon$, i.e., for any $\mathbf{x}' \in up(\mathcal{V})$, $\|\mathbf{x} - \mathbf{x}'\|_\infty > \epsilon$. This means that $up(\mathcal{V})$ is a subset of the region $\{\mathbf{x}' : x'_k > x_k + \epsilon \text{ for some } k\}$ (this is the region $\mathcal{S}$ shown in the Figure EC.1). But since $\mathbf{y} = \mathbf{x} - \mathbf{v}$, we have $\mathbf{y} \preceq \mathbf{x}$ ($\mathbf{y}$ is in region $\mathcal{S}'$ shown in the Figure EC.1). But then for any $\mathbf{w} \in \mathcal{S}'$, $\|\mathbf{y} - \mathbf{w}\|_\infty > \epsilon$. This contradicts the fact that for $\mathbf{y}$ there is some $\mathbf{y}' \in \mathcal{V}$, such that $\mathbf{y} + \epsilon\mathbf{1} \succeq \mathbf{y}'$. Thus $d(\mathcal{U},\mathcal{V}) \leq \epsilon$. Now suppose that $d(\mathcal{U},\mathcal{V}) \leq \epsilon$. Then for any $\mathbf{x} \in \mathcal{U}$, there is a $\mathbf{x}' \in up(\mathcal{V})$ such

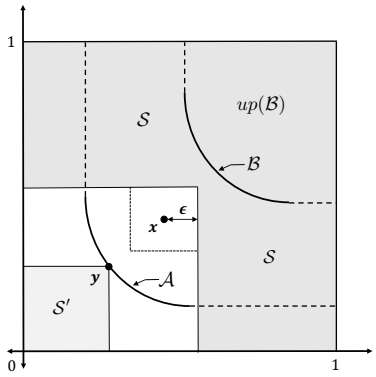

Figure EC.1 Construction in $[0,1]^2$ for the proof of Proposition 2.

that $\|\mathbf{x} - \mathbf{x}'\|_\infty \leq \epsilon$ where $\mathbf{x}' = \mathbf{y} + \mathbf{v}$ for $\mathbf{y} \in \mathcal{V}$ and $\mathbf{v} \succeq 0$. Thus $\mathbf{x} + \epsilon\mathbf{1} \succeq \mathbf{x}' = \mathbf{y} + \mathbf{v}$. The roles of $\mathcal{U}$ and $\mathcal{V}$ can be reversed. Thus $\max(e(\mathcal{U},\mathcal{V}), e(\mathcal{V},\mathcal{U})) \leq \epsilon$. Observe that this proof uses the fact that the sup and inf in the definitions of h and e can be replaced by max and min respectively, which is valid for the space $\mathcal{F}$ as discussed in footnotes 6 and 7. □

Proof of Lemma 2: In order to prove this lemma, we need a few intermediate results. We define the following notion of convexity of Pareto frontiers.

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

We then show the following equivalence.

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

1. *$\mathcal{V}$ is in $\mathcal{F}$.*
2. *$\mathcal{V} \subseteq [0,1]^K$ is p-convex and $up(\mathcal{V})$ is closed.*

Proof. To show that 1 implies 2, we just need to show that $\mathcal{V}$ is p-convex. To see this, suppose that $\mathbf{u}$ and $\mathbf{v}$ are two points in $\mathcal{V}$. Since they also belong to $up(\mathcal{V})$, which is convex, for each $\lambda \in [0,1]$, $\lambda\mathbf{u} + (1-\lambda)\mathbf{v} \in up(\mathcal{V})$ and thus there is some $\mathbf{r} \in \mathcal{V}$ such that $\mathbf{r} \preceq \lambda\mathbf{u} + (1-\lambda)\mathbf{v}$. Thus $\mathcal{V}$ is p-convex.

To show that 2 implies 1, we just need to show that $up(\mathcal{V})$ is convex if $\mathcal{V}$ is p-convex. To see this, suppose that $\mathbf{u}+\mathbf{x}$ and $\mathbf{v}+\mathbf{y}$ are two points in $(\mathcal{V})$ where $\mathbf{u}$, $\mathbf{v} \in \mathcal{V}$ and $\mathbf{x}$, $\mathbf{y} \succeq 0$. By p-convexity of $\mathcal{V}$, for each $\lambda \in [0,1]$, there is a $\mathbf{r} \in \mathcal{V}$ such that $\mathbf{r} \preceq \lambda\mathbf{u} + (1-\lambda)\mathbf{v}$ and thus $\mathbf{r} \preceq \lambda(\mathbf{u}+\mathbf{x}) + (1-\lambda)(\mathbf{v}+\mathbf{y})$. Thus $up(\mathcal{V})$ is convex.

□

We can now prove Lemma 2. Recall that,

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

First, note that $\Lambda(\Psi(\mathcal{V})) = \Lambda(\Psi(up(\mathcal{V})))$. Now one can see that $\Psi(up(\mathcal{V}))$ is the image of a continuous function from the product space $up(\mathcal{V})^{m\times n} \times \Delta(A)$ to a point in $\mathbb{R}^K$, which is a Hausdorff space. Since $up(\mathcal{V})$ is closed and bounded, it is compact. Also the simplex $\Delta(A)$ is compact. Thus the product space $up(\mathcal{V})^{m\times n} \times \Delta(A)$ is compact. Hence by the closed map lemma, f is a closed map and hence $\Psi(up(\mathcal{V}))$ is closed. Hence $up(\Lambda(\Psi(\mathcal{V})))$ is closed.

Next, recall that any point $\mathbf{u}$ in $\Lambda(\Psi(\mathcal{V}))$ is of the form:

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

for some $\boldsymbol{\alpha} \in \Delta(A)$ and $\mathbf{R}(a,b) \in \mathcal{V}$. But since $\mathcal{V}$ is p-convex from Lemma EC.2, for each $b \in B$, there exists some $\mathbf{Q}(b) \in \mathcal{V}$ such that $\mathbf{Q}(b) \preceq \sum_{a=1}^{m} \alpha_a \mathbf{R}(a,b)$. Hence statement 2 follows.

Now let

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

and

$$\mathbf{v} = \left(\max_{b\in B} \left\{ \sum_{a\in A} \eta_a r_k(a,b) + \beta R_k(b) \right\}; k = 1, \cdots, K \right)$$

be two points in $\Lambda(\Psi(\mathcal{V}))$, where $\boldsymbol{\alpha}$, $\boldsymbol{\eta} \in \Delta(A)$ and $\mathbf{Q}(b)$, $\mathbf{R}(b) \in \mathcal{V}$ for all $b \in B$. For a fixed $\lambda \in [0,1]$, let $\mathbf{z} = \boldsymbol{\alpha}\lambda + \boldsymbol{\eta}(1-\lambda)$. Then

$$\begin{aligned}
&\lambda\mathbf{u} + (1-\lambda)\mathbf{v} \\
&= \left(\lambda \max_{b\in B} \left\{ \sum_{a\in A} \alpha_a r_k(a,b) + \beta Q_k(b) \right\} + (1-\lambda) \max_{b\in B} \left\{ \sum_{a\in A} \eta_a r_k(a,b) + \beta R_k(b) \right\}; k = 1, \cdots, K \right) \\
&\succeq \left(\max_{b\in B} \left\{ \sum_{a\in A} z_a r_k(a,b) + \beta[\lambda Q_k(b) + (1-\lambda) R_k(b)] \right\}; k = 1, \cdots, K \right) \\
&\succeq \left(\max_{b\in B} \left\{ \sum_{a\in A} z_a r_k(a,b) + \beta L_k(b) \right\}; k = 1, \cdots, K \right).
\end{aligned}$$

The first inequality holds since max is a convex function and the second follows since $\mathcal{V}$ is p-convex, and hence $\mathbf{L}(b) \in \mathcal{V}$ that satisfy the given relation exist. Thus $\Lambda(\Psi(\mathcal{V}))$ is p-convex. Combined with the fact that $up(\Lambda(\Psi(\mathcal{V})))$ is closed, this implies that $\Lambda(\Psi(\mathcal{V})) \in \mathcal{F}$ using Lemma EC.2. □

Proof of Lemma 3. Suppose $e(\mathcal{U}, \mathcal{V}) = \epsilon$. Let

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

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

$$\begin{aligned}
\max_{b\in B} \left\{ \sum_{a\in A} \alpha_a r_k(a,b) + \beta Q_k(b) \right\} &= \max_{b\in B} \left\{ \sum_{a\in A} \alpha_a r_k(a,b) + \beta R_k(b) + \beta(Q_k(b) - R_k(b)) \right\} \\
&\leq \max_{b\in B} \left\{ \sum_{a\in A} \alpha_a r_k(a,b) + \beta R_k(b) + \beta\epsilon \right\} \\
&= \max_{b\in B} \left\{ \sum_{a\in A} \alpha_a r_k(a,b) + \beta R_k(b) \right\} + \beta\epsilon.
\end{aligned}$$

Thus

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

But since

$$\left(\max_{b\in B}\left\{\sum_{a\in A}\alpha_a r_k(a,b)+\beta Q_k(b)\right\};k=1,\cdots,K\right)\in\Psi(\mathcal{U}),$$

and since $\Phi(\mathcal{U})=\Lambda(\Psi(\mathcal{U}))$, there exists some $\mathbf{L}\in\Phi(\mathcal{U})$ such that

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

Thus

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

Thus

$$e(\Phi(\mathcal{U}),\Phi(\mathcal{V}))\leq\beta\epsilon=\beta e(\mathcal{U},\mathcal{V}). \tag{EC.1}$$

□

Proof of Theorem 2. In $\mathbb{G}^\infty$, fix $T\geq 1$ and consider a truncated game where Alice can guarantee the cumulative losses in $\beta^{T+1}\mathcal{V}^*$ after time $T+1$. Then the minimal losses that she can guarantee after time T is the set:

$$\Lambda\left(\left\{\left(\max_{b\in B}\beta^T\sum_{a\in A}\alpha_a r_k(a,b)+\beta^{T+1}Q_k(b);k=1,\cdots,K\right)\mid\boldsymbol{\alpha}\in\Delta(A),\ \mathbf{Q}(b)\in\mathcal{V}^*\,\forall\, b\in B\right\}\right).$$

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

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

$$\left\{\mathbf{u}\in[0,1]^K:u_k\in[x_k-\beta^{T+1},x_k+\beta^{T+1}]\text{ for all }k,\ x\in\mathcal{V}^*\right\}.$$

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

Proof of Theorem 3. Assume that Alice can guarantee every pair $\beta^{T+1}\mathbf{u}$ of cumulative losses with $\mathbf{u} \in \mathcal{V}^*$ after time $T+1$ by choosing some continuation strategy in Π_A. Let $\mathbf{x} = \mathbf{F}(\mathbf{p}, \mathcal{V}^*)$. We claim that after time T, Alice can guarantee a loss of no more than $\beta^T \mathbf{x}$ on each component by first choosing $a_T = a$ with probability $\alpha_a(\mathbf{p})$ and then if Bob chooses $b \in B$, choosing a continuation strategy that guarantees her $\mathbf{F}(\mathbf{p}', \mathcal{V}^*)$, where $p' = \mathbf{q}(b, \mathbf{p})$. Indeed by following this strategy, her expected loss on component k after time T is then

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

Thus, this strategy for Alice guarantees that her loss after time T is no more than $\beta^T \mathcal{V}^*$. Hence by induction, following the indicated strategy (in the statement of the theorem) for the first T steps and then using the continuation strategy from time $T+1$ onwards, guarantees that her loss is not more than $\mathbf{F}(\mathbf{p}_1, \mathcal{V}^*)$ after time 0. Now, even if Alice plays arbitrarily after time $T+1$ after following the indicated strategy for the first T steps, she still guarantees that her loss is (componentwise) no more than $\mathbf{F}(\mathbf{p}_1, \mathcal{V}^*) + \beta^{T+1}(1; k = 1, \cdots, K)^T$. Since this is true for arbitrarily large values of T, playing the given strategy indefinitely guarantees that her loss is no more than $\mathbf{F}(\mathbf{p}_1, \mathcal{V}^*)$. □

Proof of Proposition 3. Any point $\mathbf{e}$ in $\Gamma_N(\mathcal{V})$ is of the form $\sum_{k=1}^M \lambda_k \mathbf{v}_k$ where $\mathbf{v}_k \in up(\mathcal{V})$ and $\sum_{k=1}^M \lambda_k = 1$, and $M \leq K$. But then by the definition of an upset, we have $\mathbf{v}'_k \in \mathcal{V}$ for each k such that $\mathbf{v}'_k \preceq \mathbf{v}_k$ and hence $\sum_{k=1}^M \lambda_k \mathbf{v}'_k \preceq \sum_{k=1}^M \lambda_k \mathbf{v}_k$. By the p-convexity of $\mathcal{V}$, there is some $\mathbf{r} \in \mathcal{V}$, such that $\mathbf{r} \preceq \sum_{k=1}^M \lambda_k \mathbf{v}'_k$, and hence $\mathbf{r} \preceq \mathbf{e}$. Thus $e(\Gamma_N(\mathcal{V}), \mathcal{V}) = 0$.

Next, we will show that for any $\mathbf{u} \in \mathcal{V}$, there exists $\mathbf{e} \in \Gamma_N(\mathcal{V})$ such that $\mathbf{e} \preceq \mathbf{u} + (1/N)\mathbf{1}$. For the rest of the proof, all the distances refer to distances in the $\mathcal{L}^\infty$ norm. Consider a line $\mathbf{x} = t\mathbf{1} + \mathbf{p}$, and suppose that the shortest distance between $\mathbf{u}$ and any point on this line is $a > 0$,

i.e., $\min\{||\mathbf{u}-\mathbf{x}||_\infty : \mathbf{x}=t\mathbf{1}+\mathbf{p}\}=a$. Let $\mathbf{x}^*=t^*\mathbf{1}+\mathbf{p}$ be the point on the line that is closest to $\mathbf{u}$. If $a^+ \triangleq \max\{(x_k^*-u_k)^+ : k=1,\cdots,K\}$ and $a^- \triangleq \max\{-(x_k^*-u_k)^- : k=1,\cdots,K\}$, then $a=\max\{a^+,a^-\}$. Consider any point $\mathbf{v}$ that is the smallest point of intersection of $\mathbf{x}=t\mathbf{1}+\mathbf{p}$ and the set $up(\mathcal{V})$. Then this point must lie in the set $\{t\mathbf{1}+\mathbf{p} : t\in[t^*-a^+,t^*+a^-]\}$, because a) if $\mathbf{v}=t'\mathbf{1}+\mathbf{p}$ for some $t'<t^*-a^+$, then it means that $\mathbf{u}$ dominates $\mathbf{v}$ which contradicts the fact that $\mathbf{v}\in up(\mathcal{V})$, and b) if $\mathbf{v}=t'\mathbf{1}+\mathbf{p}$ for some $t'>t^*+a^-$ then $\mathbf{v}$ will *strictly* dominate $\mathbf{u}$ on each dimension, but then the point $\mathbf{v}'=(t^*+a^-)\mathbf{1}+\mathbf{p}$ is strictly smaller than $\mathbf{v}$ and lies in $up(\mathcal{V})$ and on the line $\mathbf{v}=t'\mathbf{1}+\mathbf{p}$, which contradicts the definition of $\mathbf{v}$. Thus $||\mathbf{u}-\mathbf{v}||_\infty \le a^+ + a^- \le 2a$. Thus we have shown that if the shortest distance between $\mathbf{u}$ and some line $\mathbf{x}=t\mathbf{1}+\mathbf{p}$ is a, then the distance between $\mathbf{u}$ and the smallest point of intersection of $\mathbf{x}=t\mathbf{1}+\mathbf{p}$ and the set $up(\mathcal{V})$ is no more than $2a$.

Now we will show that the for any $\mathbf{u}\in\mathcal{V}$, there is always a line $\mathbf{x}=t\mathbf{1}+\mathbf{p}$ such that the shortest distance between $\mathbf{u}$ and the line is no more than $1/(2N)$. Let $u_{\min}=\min_k\{u_k\}$. Then $\mathbf{u}=u_{\min}\mathbf{1}+(\mathbf{u}-u_{\min}\mathbf{1})$. Now the vector $(\mathbf{u}-u_{\min}\mathbf{1})$ has value 0 on one dimension, and on every other dimension it has value in $[0,1]$ (since $\mathbf{u}\in[0,1]^K$), and so it can be approximated by some $p_k\in\{0,1/N,\cdots,(N-1)/N,1\}$ where the approximation error on any dimension is at most $1/(2N)$. Thus there is a point $\mathbf{e}'=u_{\min}\mathbf{1}+\mathbf{p}$ where $\mathbf{p}\in\mathcal{P}_N$ such that $\|\mathbf{u}-\mathbf{e}'\|\le 1/(2N)$. Thus there us always a line $\mathbf{x}=t\mathbf{1}+\mathbf{p}$ such that the shortest distance between $\mathbf{u}$ and the line is no more than $1/2N$.

Together, we finally have that for any $\mathbf{u}\in\mathcal{V}$ there is some point $\mathbf{e}''$, which is the smallest point of intersection of some line $\mathbf{x}=t\mathbf{1}+\mathbf{p}$ and the set $up(\mathcal{V})$, such that $\|\mathbf{u}-\mathbf{e}''\|_\infty \le 2\times(1/2N)=1/N$, and thus $\mathbf{e}''\preceq\mathbf{u}+(1/N)\mathbf{1}$. Since there is always some point $\mathbf{e}\in\Gamma_N(\mathcal{V})$ such that $\mathbf{e}\preceq\mathbf{e}''$ (recall the definition (13) of $\Gamma_N(\mathcal{V})$ as the Pareto frontier of the set $\mathbf{ch}\big(\{\mathbf{F}(\mathbf{p},\mathcal{V}):\mathbf{p}\in\mathcal{P}_N\}\big)$), we have $\mathbf{e}\preceq\mathbf{u}+(1/N)\mathbf{1}$. Thus $e(\mathcal{V},\Gamma_N(\mathcal{V}))\le 1/N$. □

Proof of Proposition 4. We have $\mathcal{G}_n=\Gamma_N\circ\Phi(\mathcal{G}_{n-1}))$. Consider another sequence of Pareto frontiers

$$\Big(\mathcal{A}_n=\Phi^n(\mathcal{G}_0)\Big)_{n\in\mathbb{N}}. \tag{EC.2}$$

Then we have

$$\begin{aligned} d(\mathcal{A}_n, \mathcal{G}_n) &= d(\Phi(\mathcal{A}_{n-1}), \Gamma_N(\Phi(\mathcal{G}_{n-1}))) \\ &\overset{(a)}{\leq} d(\Phi(\mathcal{A}_{n-1}), \Phi(\mathcal{G}_{n-1})) + d(\Phi(\mathcal{G}_{n-1}), \Gamma_N(\Phi(\mathcal{G}_{n-1}))) \\ &\overset{(b)}{\leq} \beta d(\mathcal{A}_{n-1}, \mathcal{G}_{n-1}) + \frac{1}{N}, \end{aligned} \tag{EC.3}$$

where inequality (a) is the triangle inequality and (b) follows from (EC.1) and Lemma 3. Coupled with the fact that $d(\mathcal{A}_0, \mathcal{G}_0) = 0$, we have that

$$\begin{aligned} d(\mathcal{A}_n, \mathcal{G}_n) &\leq \frac{1}{N}\left(1 + \beta + \beta^2 + \cdots \beta^{n-1}\right) \\ &= \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right). \end{aligned} \tag{EC.4}$$

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 $\mathcal{A}_0 = \mathcal{G}_0 = \{\mathbf{0}\}$, and since the stage losses $r_k(a,b) \in [0, 1-\beta]$, we have that $d(\mathcal{A}_1, \mathcal{A}_0) \leq 1-\beta$. From the contraction property, this implies that $d(\mathcal{A}_{n+1}, \mathcal{A}_n) \leq \beta^n(1-\beta)$. Thus $d(\mathcal{V}^*, \mathcal{A}_n) \leq \frac{\beta^n(1-\beta)}{1-\beta} = \beta^n$, and thus by triangle inequality we have

$$d(\mathcal{V}^*, \mathcal{G}_n) \leq \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right) + \beta^n. \tag{EC.5}$$

Finally, to show that $e(\mathcal{G}_n, \mathcal{V}^*) \leq \beta^n$, observe that

$$\begin{aligned} e(\mathcal{G}_n, \mathcal{A}_n) &= e(\Gamma_N(\Phi(\mathcal{G}_{n-1})), \Phi(\mathcal{A}_{n-1})) \\ &\overset{(a)}{\leq} e(\Gamma_N(\Phi(\mathcal{G}_{n-1})), \Phi(\mathcal{G}_{n-1})) + e(\Phi(\mathcal{G}_{n-1}), \Phi(\mathcal{A}_{n-1})) \\ &\overset{(b)}{\leq} 0 + \beta e(\mathcal{G}_{n-1}, \mathcal{A}_{n-1}). \end{aligned} \tag{EC.6}$$

Since $\mathcal{A}_0 = \mathcal{G}_0 = \{\mathbf{0}\}$, this implies that $e(\mathcal{G}_n, \mathcal{A}_n) = 0$ for all n. Here, (a) holds since if for three frontiers $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{L}$, $\mathcal{U}$ ϵ_1-dominates $\mathcal{V}$ and $\mathcal{V}$ ϵ_2-dominates $\mathcal{Z}$, then $\mathcal{U}$ $(\epsilon_1 + \epsilon_2)$-dominates $\mathcal{Z}$. (b) follows from the contraction property of Φ under e. Further, $e(\mathcal{A}_n, \mathcal{V}^*) \leq d(\mathcal{A}_n, \mathcal{V}^*) \leq \beta^n$ from above. Thus we have $e(\mathcal{G}_n, \mathcal{V}^*) \leq e(\mathcal{G}_n, \mathcal{A}_n) + e(\mathcal{A}_n, \mathcal{V}^*) \leq \beta^n$.

□

Proof of Proposition 5. In order to prove this result, we need a few intermediate definitions and results. First, we need to characterize the losses guaranteed by any $H(K,N)-$mode stationary strategy. Such a strategy π defines the following operator on any function $\mathbf{F}:\mathcal{P}_N\rightarrow\mathbb{R}^K$ ($\mathcal{P}_N$ is defined in (11)):

$$\Delta_N^{\pi}(\mathbf{F})(\mathbf{p})=\left(\max_{b\in B}\Big\{\sum_{a\in A}\alpha_a(\mathbf{p})r_k(a,b)+\sum_{k'=1}^{K}z_{k'}(b,\mathbf{p})\beta F_k(\mathbf{q}_{k'}(b,\mathbf{p}))\Big\};k=1,\cdots,K\right), \tag{EC.7}$$

where $\mathbf{q}_{k'}(b,\mathbf{p})\in\mathcal{P}_N$ for all k'. Now for a function $\mathbf{F}:\mathcal{P}_N\rightarrow\mathbb{R}^K$, define the following norm:

$$\|\mathbf{F}\|=\max_{\mathbf{p}\in\mathcal{P}_N}\|\mathbf{F}(\mathbf{p})\|_{\infty}.$$

It is easy to show that Δ_N^{π} is a contraction in the norm. We omit the proof for the sake of brevity.

Lemma EC.3.

$$\|\Delta_N^{\pi}(\mathbf{F})-\Delta_N^{\pi}(\mathbf{G})\|\leq\beta\|\mathbf{F}-\mathbf{G}\|. \tag{EC.8}$$

We can then show the following result.

Lemma EC.4. *Consider a $H(K,N)$-mode strategy π. Then there is a unique function*

$$\mathbf{F}^{\pi}:\mathcal{P}_N\rightarrow\mathbb{R}^K$$

such that $\Delta_N^{\pi}(\mathbf{F}^{\pi})=\mathbf{F}^{\pi}$. Further, The strategy π initiated at mode $\mathbf{p}$ where $\mathbf{p}\in\mathcal{P}_N$ guarantees the vector of losses $\mathbf{F}^{\pi}(\mathbf{p})$.

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. The arguments are not repeated here for the sake of brevity. Now let

$$\mathcal{V}^{\pi_n}\triangleq\Lambda\Big(ch(\{\mathbf{F}^{\pi_n}(\mathbf{p}):\mathbf{p}\in\mathcal{P}_N)\Big),$$

where $\mathbf{F}^{\pi_n}$ is the fixed point of the operator $\Delta_N^{\pi_n}$.

Define a sequence of functions $\mathbf{F}^n : \mathcal{P}_N \to \mathbb{R}^K$ where $\mathbf{F}^n(\mathbf{p}) = \mathbf{F}(\mathbf{p}, \Phi(\mathcal{G}^{n-1})) = \mathbf{F}(\mathbf{p}, \mathcal{G}^n)$. We then have that

$$\begin{aligned} d(\mathcal{V}^{\pi_n}, \mathcal{V}^*) &\leq d(\mathcal{V}^{\pi_n}, \mathcal{G}_n) + d(\mathcal{G}_n, \mathcal{V}^*) \\ &\leq d(\mathcal{V}^{\pi_n}, \mathcal{G}_n) + \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right) + \beta^n. \end{aligned} \tag{EC.9}$$

From the definition of d, it is clear that $d(\mathcal{V}^{\pi_n}, \mathcal{G}_n) \leq \|\mathbf{F}^{\pi_n} - \mathbf{F}^n\|$. Next we have

$$\begin{aligned} \|\mathbf{F}^{\pi_n} - \mathbf{F}^n\| &\leq \|\mathbf{F}^{\pi_n} - \Delta_N^{\pi_n}(\mathbf{F}^n)\| + \|\Delta_N^{\pi_n}(\mathbf{F}^n) - \mathbf{F}^n\| \\ &\overset{(a)}{=} \|\Delta_N^{\pi_n}(\mathbf{F}^{\pi_n}) - \Delta_N^{\pi_n}(\mathbf{F}^n)\| + \|\mathbf{F}^{n+1} - \mathbf{F}^n\| \\ &\overset{(b)}{\leq} \beta\|\mathbf{F}^{\pi_n} - \mathbf{F}^n\| + \|\mathbf{F}^{n+1} - \mathbf{F}^n\|. \end{aligned} \tag{EC.10}$$

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

$$d(\mathcal{V}^{\pi_n}, \mathcal{G}_n) \leq \|\mathbf{F}^{\pi_n} - \mathbf{F}^n\| \leq \frac{\|\mathbf{F}^{n+1} - \mathbf{F}^n\|}{1-\beta}. \tag{EC.11}$$

And finally we have:

$$d(\mathcal{V}^{\pi_n}, \mathcal{V}^*) \leq \frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right) + \beta^n + \frac{\|\mathbf{F}^{n+1} - \mathbf{F}^n\|}{1-\beta}. \tag{EC.12}$$

To finish up, we need the following result:

Lemma EC.5.

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

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

Finally, by the triangle inequality we have

$$\begin{aligned} d(\mathcal{G}_{n+1},\mathcal{G}_n) &\le d(\mathcal{A}_{n+1},\mathcal{A}_n)+d(\mathcal{G}_{n+1},\mathcal{A}_{n+1})+d(\mathcal{G}_n,\mathcal{A}_n) \\ &\le (1-\beta)\beta^n+\frac{1}{N}\left(\frac{1-\beta^{n+1}}{1-\beta}\right)+\frac{1}{N}\left(\frac{1-\beta^n}{1-\beta}\right). \end{aligned} \tag{EC.13}$$

Combining with (EC.12) we have the result. □

EC.2. Some remarks on Pareto frontiers of closed and convex sets.

The Pareto frontier of a closed set is not necessarily closed. Figure 1 is one example – the upset is closed, but its Pareto frontier is open. But we can show that the Pareto frontier of a closed and convex set in $\mathbb{R}^2$ is closed.[9]

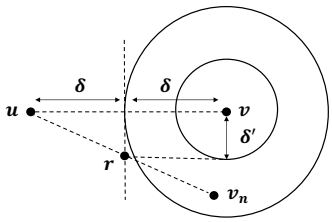

Figure EC.2 Construction in the proof of Proposition *EC.1*.

PROPOSITION EC.1. *Let $\mathcal{V}$ be the lower Pareto frontier of a closed and convex set $\mathcal{S}$ in $\mathbb{R}^2$. Then $\mathcal{V}$ is closed.*

Proof. Suppose that $\{\mathbf{v}_n\}$ is a sequence of points in $\mathcal{V}$ that converge to some point $\mathbf{v}$. Then since $\mathcal{S}$ is closed, $\mathbf{v}\in\mathcal{S}$. We will show that $\mathbf{v}\in\mathcal{V}$. Suppose not. Then there is some $\mathbf{u}\in\mathcal{V}$ such that $\mathbf{u}\preceq\mathbf{v}$. Suppose first that $u_1<v_1$ and $u_2<v_2$. Then let $\epsilon=(\min(v_1-u_1,v_2-u_2)/2$ and consider the $\mathcal{L}^2$ ball of radius ϵ around $\mathbf{v}$, i.e.

$$B_{\mathbf{v}}(\epsilon)=\{\mathbf{y}\in\mathbb{R}^2:\|\mathbf{y}-\mathbf{v}\|_2\le\epsilon\}.$$

[9] Of course, the Pareto frontier of a closed set may be empty, e.g., $\{(x,y)\in\mathbb{R}^2:x=y\}$, in which case it is trivially closed.

Then for any point $\mathbf{y}$ in $B_{\mathbf{v}}(\epsilon)$, we have that $\mathbf{u} \preceq \mathbf{y}$. But since $\{\mathbf{v}^n\}$ converges to $\mathbf{v}$, there exists some point in the sequence that is in $B_{\mathbf{v}}(\epsilon)$, and $\mathbf{u}$ is dominated by this point, which is a contradiction. Hence either $u_1 = v_1$ or $u_2 = v_2$. Suppose w.l.o.g. that $u_1 < v_1$ and $u_2 = v_2$. See Figure EC.2. Let $\delta = (v_1 - u_1)/2$ and consider the ball of radius δ centered at $\mathbf{v}$, i.e. $B_{\mathbf{v}}(\delta)$. Let $\mathbf{v}^n$ be a point in the sequence such that $\mathbf{v}^n \in B_{\mathbf{v}}(\delta)$. Now $v_{n,1} > u_1$ and hence it must be that $v_{n,2} < u_2$. Now for some $\lambda \in (0,1)$, consider a point $\mathbf{r} = \lambda \mathbf{u} + (1-\lambda)\mathbf{v}_n$ such that $r_1 = u_1 + \delta$. It is possible to pick such a point since a) $v_1 = u_1 + 2\delta$ and b) $|v_{n,1} - v_1| \leq \delta$, which together imply that $v_{n,1} \geq u_1 + \delta$ (please see the figure). Now $\mathbf{r} \in S$ since S is convex. Next $r_1 = v_1 - \delta < v_1$, and also $r_2 < u_2 = v_2$ since $\lambda > 0$ and $v_{n,2} < u_2$. Let $\delta' = v_2 - r_2$. Then consider the ball $B_{\mathbf{v}}(\delta')$ centered at $\mathbf{v}$. Clearly $\mathbf{r} \preceq \mathbf{y}$ for any $\mathbf{y} \in B_{\mathbf{v}}(\delta')$. But since $\{\mathbf{v}_n\}$ converges to $\mathbf{v}$, there exists some point in the sequence that is in $B_{\mathbf{v}}(\delta')$, and $\mathbf{r}$ is dominated by this point, which is again a contradiction. Thus $\mathbf{v} \in \mathcal{V}$. □

Interestingly, this result doesn't hold for closed and convex sets in $\mathbb{R}^K$ for $K > 2$. A counterexample can be found in Kruskal [31]. A variant of this counterexample is depicted in Figure EC.3 for completeness. The closed and convex set $\mathcal{S}$ is a solid 3-dimensional cone with apex $(0,0,1)$ and base being the semicircular disc defined by the set $\{(x,y,z) \in \mathbb{R}^3 : x = 5,\ (y-1)^2 + (z-1)^2 \leq 1,\ z \leq 1\}$. The sequence $(\mathbf{v}_n)_{n \in \mathbb{N}}$ lies on the Pareto frontier of this set as shown in the figure, but it converges to the point $(5,0,1)$, which dominates the point $(0,0,1)$.

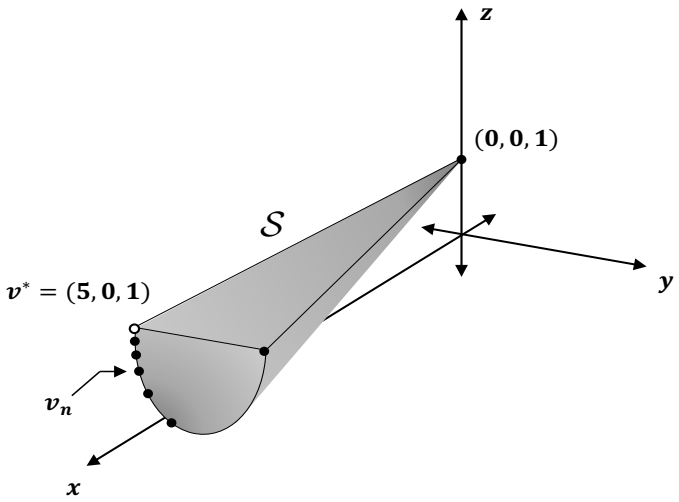

Figure EC.3 An example of a compact and convex set in $\mathbb{R}^3$ whose Pareto frontier is not closed. The sequence $(\mathbf{v}_n)$ on the Pareto frontier converges to the point $(5, 0, 1)$, which dominates $(0, 0, 1)$.

EC.3. Solving Linear Program (14).

When $\mathcal{V}$ is the lower Pareto frontier of a convex polytope, (14) is a linear program. In this program, $\mathbf{x}$ is a dependent vector and can be eliminated. The only question remains is that of addressing the constraint $\mathbf{Q}(b) \in \mathcal{V}$. The vertices of $\mathcal{V}$ are a subset of $\{\mathbf{F}(\mathbf{p}, \Phi(\mathcal{V})) : \mathbf{p} \in \mathcal{P}_N\}$. Thus $\mathbf{Q}(b)$ can be chosen as a convex combination of points in $\{\mathbf{F}(\mathbf{p}, \Phi(\mathcal{V})) : \mathbf{p} \in \mathcal{P}_N\}$. This introduces $H(K, N)$ variables for each $b \in B$ along with the constraint that these variables sum to 1, thus contributing $mH(K, N)$ variables and m constraints. Along with the variables α and t, this makes $mH(K, N) + l + 1$ variables in total. And along with the K domination constraints for each $b \in B$ and the constraint that $\sum_a \alpha_a = 1$, this makes $Km + m + 1$ constraints in total (ignoring non-negativity constraints on all variables except t).

Of the $H(K, N)$ variables associated with each $b \in B$ that determine the point $\mathbf{Q}(b)$, we know that at most K will be non-zero. This sparsity constraint can potentially be utilized to speed up the computation, although we didn't attempt to do so in our computations.

EC.4. Evaluation of finite-mode stationary policies

In this section, we show how the upper bounds on the losses guaranteed by a finite-mode stationary policy can be efficiently computed as the solution to a linear program. Consider an M-mode stationary policy π, with a set of modes $\mathcal{M} = \{1, \cdots, M\}$, where each mode $i \in \mathcal{M}$ is associated with a probability distribution $\boldsymbol{\alpha}_i \in \Delta(A)$ over immediate actions and a transition rule $(\mathbf{z}_i(b) \in \Delta(\mathcal{M}); b \in B)$. Let $\mathbf{v}_i$ be the vector of smallest upper bounds on the total discounted losses guaranteed by this policy starting from mode i. Then $\{\mathbf{v}_i; i \in \mathcal{M}\}$ can be computed as the solution of the following linear optimization problem.

$$\min_{(\mathbf{v}_i)_{i\in\mathcal{M}}} \sum_{i\in\mathcal{M}} \mathbf{1}^T \mathbf{v}_i \tag{EC.14a}$$

$$\text{s.t. } \mathbf{v}_i \succeq \sum_{a\in A} \alpha_{i,a} \mathbf{r}(a,b) + \beta \sum_{j\in\mathcal{M}} z_{i,j}(b) \mathbf{v}_j, \text{ for all } b \in B \text{ and } i \in \mathcal{M}. \tag{EC.14b}$$

It is clear that if $\mathbf{v}_i$ is the vector of smallest upper bounds on the losses corresponding to mode i, then the set $(\mathbf{v}_i)$ satisfies (EC.14b), which captures the Bellman one-step optimality conditions. In other words, $(\mathbf{v}_i)$ is feasible in the above program. On the other hand, due to the same inequalities in (EC.14b), any feasible set $(\mathbf{v}_i)$ can be approximately guaranteed by the given stationary policy in a long but finite discounted repeated game, where the approximation error goes to 0 as the length of the game approaches infinity. Hence, the upper bounds $(\mathbf{v}_i)$ can be guaranteed in the infinitely repeated game (this argument is analogous to the one used in the proof of Theorem 2). We thus conclude that the solution to the linear program above yields the smallest lower bounds on the losses corresponding to the different modes. Note that in the objective function, the weights for the different components of $\mathbf{v}_i$ for the different modes i can be chosen to be any positive numbers. Finally, note that this linear program decouples across dimensions and hence, can be solved for each dimension k to obtain $\{v_{k,i}; i \in \mathcal{M}\}$.

EC.5. Benchmark algorithms

EC.5.1. Hedge

In this algorithm, if $L_t(i)$ is the cumulative loss of expert i till time t, then the probability of choosing expert i at time $t+1$ is

$$p_i(t+1) \propto \exp(-\eta L_t(i)),$$

where η is a parameter. In the undiscounted problem, choosing $\eta = \sqrt{8 \log K/T}$ when the time horizon T is known attains an upper bound of $\sqrt{T \log K/2}$ on the expected cumulative regret (Thm. 2.2, [18]). In a certain sense, this is shown to be asymptotically optimal in K and T for general loss function taking values in $[0,1]$ (Thm. 3.7, [18]). In our implementation, we use discounted cumulative losses in this algorithm, and choose $\eta = \sqrt{8 \log K(1-\beta^2)}$. This resulting algorithm achieves an upper bound of $\sqrt{\log K/2(1-\beta^2)} = \sqrt{\log K/(2(1-\beta)(1+\beta))}$ on the expected discounted regret in the infinitely repeated game (see proof of Thm. 2.2, and Thm. 2.8 in [18]).

EC.5.2. GPS

The GPS algorithm for $K = 2$ experts is defined as follows [24]. Let $\xi = (1 - \sqrt{1-\beta^2})/\beta$.Let d be the difference in the cumulative (undiscounted) losses of the leading expert (the one with the lower cumulative loss) and the lagging expert (the one with the higher loss). Then at every stage, the algorithm chooses the leading expert with probability $1 - (1/2)\xi^d$ and the lagging expert with probability $(1/2)\xi^d$.

EC.6. Comparison to the model of expert selection considered in GPS [24].

In the expert selection game considered in [24] (GPS), at each stage, the game ends with a probability $1-\beta$ and continues with probability β, amounting to a geometric distribution on the decision horizon with parameter β. This is essentially a reinterpretation of our model of discounted losses, where the discount factor is interpreted as the probability of continuation at each stage. But the difference between their formulation and our formulation is in the definition of regret.

Assuming this interpretation of a game with a random decision horizon, in defining our regret, the loss of the decision-maker is compared to the lowest expected loss across the experts, where the expectation is over the time horizon. Formally, let S denote the random time horizon. Then our regret minimization objective is the following.

$$\textbf{Our problem: } \min_{\pi_A \in \Pi_A} \max_{\mathbf{b} \in B^\infty} \mathrm{E}_{\pi_A, S}\left[\sum_{t=1}^{S} L(a_t, b_t)\right] - \underbrace{\min_{a \in A} \mathrm{E}_S\left[\sum_{t=1}^{S} L(a, b_t)\right]}_{\text{Our regret benchmark}} \tag{EC.15}$$

$$= \min_{\pi_A \in \Pi_A} \max_{\mathbf{b} \in B^\infty} \mathrm{E}_{\pi_A}\left[\sum_{t=1}^{\infty} \beta^{t-1} L(a_t, b_t)\right] - \min_{a \in A} \sum_{t=1}^{\infty} \beta^{t-1} L(a, b_t). \tag{EC.16}$$

The interpretation is that the decision-maker measures regret relative to the best, fixed action she could have chosen if she had known the adversary's sequence of actions, but not the horizon. In contrast, in the formulation of GPS, the loss of the decision-maker is compared to the expectation of the lowest total loss across the experts over the time horizon.

$$\textbf{GPS problem: } \min_{\pi_A \in \Pi_A} \max_{\mathbf{b} \in B^\infty} \mathrm{E}_{\pi_A, S}\left[\sum_{t=1}^{S} L(a_t, b_t)\right] - \underbrace{\mathrm{E}_S\left[\min_{a \in A} \sum_{t=1}^{S} L(a, b_t)\right]}_{\text{Regret benchmark of GPS}} \tag{EC.17}$$

$$= \min_{\pi_A \in \Pi_A} \max_{\mathbf{b} \in B^\infty} \mathrm{E}_{\pi_A}\left[\sum_{t=1}^{\infty} \beta^{t-1} L(a_t, b_t)\right] - (1-\beta) \sum_{s=1}^{\infty} \beta^{s-1} \min_{a \in A} \sum_{t=1}^{s} L(a, b_t). \tag{EC.18}$$

The interpretation is that the decision-maker measures regret relative to the best, fixed action she could have chosen if she had known the adversary's sequence of actions and also the decision horizon.

Notice that the interchange of the minimum and the expectation operators compared to our regret benchmark results in a lower, or in other words, a more ambitious regret benchmark under the GPS formulation. Consequently, the optimal regret in the GPS formulation is at least as large as the optimal regret in our formulation. For example, for $K = 2$ and $\beta = 0.9$, the optimal expected total regret in their formulation is ≈ 1.147,[10] while in our formulation, the optimal regret in this case is at most ≈ 0.9338 (see Figure 9).

[10] The optimal regret for [24]'s formulation is $\frac{1}{2\sqrt{1-\beta^2}}$ (or an average discounted regret of $\frac{1-\beta}{2\sqrt{1-\beta^2}} = (1/2) \times \sqrt{(1-\beta)/(1+\beta)}$). The expression in [24] is off by a factor of β compared to this expression since they discount the first period by β (i.e., the game could end before the first stage begins), whereas we discount it by 1.

Moreover, if a policy π_A for the decision-maker guarantees some maximal value of a in the GPS problem, then it will also guarantee a in our problem. This is because, for such a policy π_A, we have,

$$\begin{aligned}
&\max_{\mathbf{b}\in B^\infty} \mathrm{E}_{\pi_A,S}\left[\sum_{t=1}^{S} L(a_t,b_t)\right] - \min_{a\in A} \mathrm{E}_S\left[\sum_{t=1}^{S} L(a,b_t)\right] \\
&\leq \max_{\mathbf{b}\in B^\infty} \mathrm{E}_{\pi_A,S}\left[\sum_{t=1}^{S} L(a_t,b_t)\right] - \mathrm{E}_S\left[\min_{a\in A}\sum_{t=1}^{S} L(a,b_t)\right] \\
&\leq a.
\end{aligned}$$

Thus, the optimal algorithm for GPS formulation yields the same guarantee for our definition of regret as the optimal regret in their definition, i.e., for instance, for $\beta = 0.9$, the optimal GPS algorithm will guarantee an expected regret of at most ≈ 1.147 according to our definition. This, however, leaves the question open of whether the optimal algorithm of GPS may achieve the optimal regret under our formulation. We show that this is not the case in general: in Section EC.7 in the Online Appendix, in the 2 Experts setting and $\beta = 0.8$, we design an adversary that induces the optimal algorithm for the GPS formulation to exceed the upper bound on the achievable regret under our formulation, thereby demonstrating its sub-optimality for our problem.

We emphasize that for decision-making over a long time horizon with discounted losses, where the discount factor captures the time value of money (e.g., in a portfolio optimization context) rather than representing a distributional parameter for a random decision horizon, our notion of regret is the more natural one to consider; indeed, in this scenario, it is difficult to attribute any meaning to the regret benchmark of GPS.

The dynamic programming technique of GPS cannot be efficiently adapted to our formulation. The distinction between the two objectives has crucial implications on the computational approaches to solving these problems. The GPS formulation is amenable to a dynamic programming approach where one keeps track of only the undiscounted total losses for each expert in the state information. This is feasible since, if the game ends at time S, then the total losses of the experts

until S are sufficient to determine the regret benchmark under the GPS formulation. Formally, observe that the GPS problem can be decomposed as:

$$\min_{\pi_A\in\Pi_A}\max_{\mathbf{b}\in B^\infty}\mathrm{E}_{\pi_A,S}\left[\sum_{t=1}^{S}L(a_t,b_t)-\min_{a\in A}\sum_{t=1}^{S}L(a,b_t)\right] \tag{EC.19}$$

$$\begin{aligned}&=\min_{\pi_A\in\Pi_A}\max_{\mathbf{b}\in B^\infty}P(S=1)\mathrm{E}_{\pi_A}\left[L(a_1,b_1)-\min_{a\in A}L(a,b_1)\right]\\&\quad+P(S>1)\mathrm{E}_{\pi_A,S}\left[L(a_1,b_1)+\sum_{t=2}^{S}L(a_t,b_t)-\min_{a\in A}\left(L(a,b_1)+\sum_{t=2}^{S}L(a,b_t)\right)\mid S>1\right]\end{aligned} \tag{EC.20}$$

$$\begin{aligned}&\overset{(a)}{=}\min_{\mathbf{p}\in\Delta(A),\pi'_A\in\Pi_A}\max_{b_1\in B}(1-\beta)\mathrm{E}_{\mathbf{p}}\left[L(a_1,b_1)-\min_{a\in A}L(a,b_1)\right]+\beta\mathrm{E}_{\mathbf{p}}[L(a_1,b_1)]\\&\quad+\max_{(b_2,b_3,\cdots)\in B^\infty}\beta\mathrm{E}_{\pi'_A,S}\left[\sum_{t=2}^{S}L(a_t,b_t)-\min_{a\in A}\left(L(a,b_1)+\sum_{t=2}^{S}L(a,b_t)\right)\mid S>1\right]\end{aligned} \tag{EC.21}$$

$$\begin{aligned}&\overset{(b)}{=}\min_{\mathbf{p}\in\Delta(A)}\max_{b_1\in B}\mathrm{E}_{\mathbf{p}}[L(a_1,b_1)]-(1-\beta)\min_{a\in A}L(a,b_1)\\&\quad+\beta\underbrace{\min_{\pi'_A\in\Pi_A}\max_{(b_2,b_3,\cdots)\in B^\infty}\mathrm{E}_{\pi'_A,S}\left[\sum_{t=2}^{S}L(a_t,b_t)-\min_{a\in A}\left(L(a,b_1)+\sum_{t=2}^{S}L(a,b_t)\right)\mid S>1\right]}_{\text{Continuation term}}.\end{aligned} \tag{EC.22}$$

Here, π'_A is the continuation strategy of the decision-maker if the game doesn't end at time $t=1$. (a) results from the fact that $P(S=1)=1-\beta$, and (b) results from the fact that the continuation strategy π'_A of the decision-maker is allowed to depend on b_1. Notice that in the continuation term, the distribution of the residual horizon conditioned on $S>1$ is once again geometric with parameter β. Hence, this term is the same as the original objective, except that the starting cumulative losses of different experts are $(L(a,b_1);a\in A)$. Suppose we denote the starting cumulative losses for the different experts as $\mathbf{c}=(c_a;\,a\in A)$. Then we can denote the value of the decision maker's objective as a function of these starting quantities as:

$$\mathrm{Val}(\mathbf{c})=\min_{\pi_A\in\Pi_A}\max_{\mathbf{b}\in B^\infty}\mathrm{E}_{\pi_A,S}\left[\sum_{t=1}^{S}L(a_t,b_t)-\min_{a\in A}\left(c_a+\sum_{t=1}^{S}L(a,b_t)\right)\right]. \tag{EC.23}$$

Then, similar to (EC.22), we can write the dynamic programming equations for this value function as:

$$\mathrm{Val}(\mathbf{c})=\min_{\mathbf{p}\in\Delta(A)}\max_{b_1\in B}\mathrm{E}_{\mathbf{p}}[L(a_1,b_1)]-(1-\beta)\min_{a\in A}(c_a+L(a,b_1))+\beta\mathrm{Val}\left((c_a+L(a,b_1);\,a\in A)\right). \tag{EC.24}$$

When the losses are in $\{0,1\}$, as in the Experts setting considered in [24], we have that $\mathbf{c} \in \mathbb{N}^K$. While the resulting state-space is infinite, we can truncate the time horizon at some T with negligible loss. In particular, truncating at T would lead to a loss of β^T on the discounted average regret ($(1-\beta)$ times the regret). So to obtain an ϵ-optimal policy, one can truncate the horizon at $T = \log \epsilon / \log(\beta) \approx \log \epsilon/(\beta - 1)$, resulting in a state space of size $\Theta(T^K) = \Theta((\frac{\log(1/\epsilon)}{(1-\beta)})^K)$. This size, though exponential in K, is significantly smaller than the worst-case state-space of a naïve dynamic programming approach, which can grow exponentially in the truncated time-horizon T.

However, notice that the decomposition that led to this gain is not feasible for our objective in (EC.16) since the optimal regret benchmark is determined *after* the expectation is taken over the random time horizon. Consequently, the information of the total losses $\sum_{t=1}^{S} L(a, b_t)$ for each expert $a \in A$ until the end of the horizon S is not sufficient to determine the expected losses $\mathrm{E}_S[\sum_{t=1}^{S} L(a, b_t)]$ for the experts. E.g., if we know that the total losses across two experts at the end of the horizon are $(5, 2)$ then that information is not sufficient to deduce the best expert for our regret benchmark. For the GPS formulation, on the other hand, this information is sufficient to deduce that the best expert is Expert 2 and the regret benchmark is 2.

Even if we consider a truncated horizon T such that $P(S > T)$ is negligible, a naïve dynamic programming approach to solve our problem would, at the very least, necessitate storing information about the total *discounted* losses incurred by any of the experts until time $1 \le s \le T$ (i.e., values of $\sum_{t=1}^{s} \beta^{t-1} L(a, b_t)$ for each $a \in A$) to approximate the regret benchmark at time T. The resulting state-space is equivalent to storing the entire history of losses across the K experts (since the losses are weighted by different discount factors across time), leading to a state-space of size $\Theta(2^{KT})$, which is exponentially larger than the $\Theta(T^K)$ state-space that suffices for the GPS formulation under the same truncation of the horizon. To obtain an ϵ-optimal algorithm, we need to truncate the horizon at a time T such that $\beta^T = \epsilon$, which requires that $T = \log(\epsilon)/\log(\beta) \approx \log(1/\epsilon)/(1-\beta)$.

The time taken to compute the optimal policy in each state is expected to be $\Omega(2^K)$,[11] resulting in $\Omega(2^K(1/\epsilon)^{\frac{K}{1-\beta}})$ computations overall.

In contrast, using our approximation approach from Section 5, to obtain an ϵ-optimal policy in the Experts setting where $m = 2^K$, we need to solve $nH(K,N)$ linear programs, each of size $\Theta(2^K H(K,N))$, where $n \approx \frac{\log(1/\epsilon)}{1-\beta}$ and $N \approx \frac{1}{(1-\beta)^2\epsilon}$. Since $H(K,N) = \Theta(N^K)$, this results in a computation time of $\mathrm{O}\left(\frac{\log(1/\epsilon)}{1-\beta}\left(\frac{2}{(1-\beta)^2\epsilon}\right)^{4K}\right)$ since linear programs can be solved in time at most cubic in the input size [48]. This is an exponential improvement in the dependence on the expected time horizon $1/(1-\beta)$ compared to the naïve dynamic programming approach. This gain essentially results from leveraging the fact that the space of possible discounted losses across experts, although infinite, is a subset of a compact space. This space is amenable to an efficient approximation as we demonstrate in Section 5.

EC.7. Hedge and GPS are suboptimal for the expert selection problem with $K = 2$ and $\beta = 0.8$.

Consider the experts problem with $K = 2$ experts and a discount factor $\beta = 0.8$. Consider the following strategy for the adversary.

Adversary A: In this strategy, the probability that expert 1 incurs a loss (and expert 2 doesn't) at time t is $0.9^{1/t}$ if t is odd, and 0.9^t if t is even. This adversary never gives equal losses to both experts.

We compare the performance of Hedge and GPS against this adversary to that of two approximately optimal policies that we design. The first policy is a $203-$mode ($H(2,101)$) stationary policy and the second is a $21-$mode ($H(2,10)$) stationary policy ($n = 28$ in both cases). Table EC.1 shows the theoretical upper bounds on the regret guaranteed by these algorithms.

Figure EC.4 compares the expected average discounted regret incurred by the four algorithms. The expected regret is estimated in each case by averaging over 10000 runs, where each run is a

[11] Given the continuation value functions, the objective in (EC.24) amounts to solving a zero-sum game in K actions for the decision-maker and 2^K actions for the adversary. This can be solved as a linear program with $\Theta(K)$ variables and $\Theta(2^K)$ constraints, which takes time $\Theta(2^{\alpha K})$ for some $\alpha > 1$ [48].

Table EC.1 Upper bounds on the average discounted regret under different policies for $\beta = 0.8$

Policy	Regret upper bound
Hedge	0.1962
GPS	0.1666
203-mode	0.1357
21-mode	0.1374

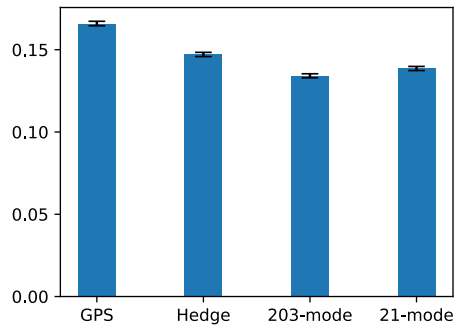

Figure EC.4 Estimates of the expected average discounted regret of different algorithms for $K = 2$ and $\beta = 0.8$ against Adversary A, along with associated error bars.

game with time horizon $T = 100$. The associated error bars are shown in the graph. Note that our strategies significantly outperform Hedge and GPS. Importantly, the regrets of Hedge and GPS significantly exceed the upper bounds on the regret guaranteed by our algorithms. This eliminates the possibility of these algorithms being optimal for our problem with high probability.

EC.8. Exact characterization of $\mathcal{V}^*$ in the expert selection problem with $K = 2$ and $\beta = 0.5$.

In some simple examples, we can exactly determine the set $\mathcal{V}^*$ by "solving" the fixed point relation given by the dynamic programming operator. We demonstrate this by determining the optimal Pareto frontier in the game of combining expert advice (Figure 7) from 2 experts for $\beta = 0.5$. Note that the points $(0, 1/(1-\beta)) = (0, 2)$ and $(1/(1-\beta), 0) = (2, 0)$ lie on $\mathcal{V}^*$ (achieved by choosing Expert 1 always or Expert 2 always, respectively). We can thus represent $\mathcal{V}^*$ by a convex and decreasing function $f(x)$ defined on $x \in [0, 2]$ such that $f(0) = 2$ and $f(2) = 0$, so that $\mathcal{V}^* = \{(x, f(x)) : x \in [0, 2]\}$. $\beta\mathcal{V}^*$ for $\beta = 0.5$ is thus the set $\{(\beta x, \beta f(x)) : x \in [0, 2]\} = \{(x, 0.5f(x/0.5) : x \in [0, 1]\}$, which thus can be represented by the convex, decreasing function $\bar{f}(x) = f(2x)/2$ defined on $x \in [0, 1]$, where $\bar{f}(0) = 1$ and $\bar{f}(1) = 0$.

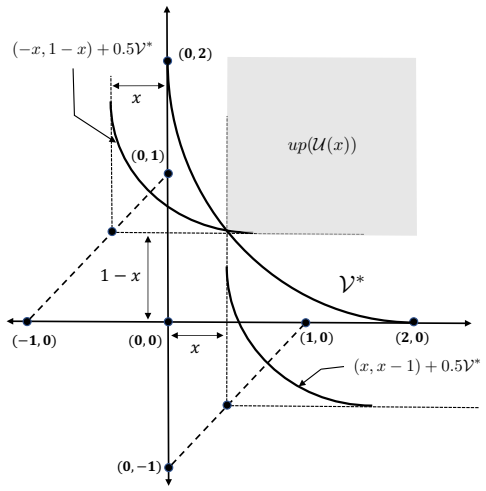

Figure EC.5 Construction of $up(\mathcal{U}(x))$.

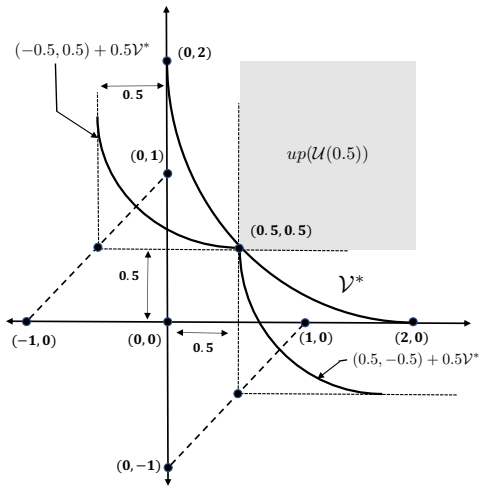

Figure EC.6 Construction of $up(\mathcal{U}(0.5))$.

Now for a fixed randomization over Alice's actions, $(1-x, x)$, by choosing different points in $\mathcal{V}^*$ from the next stage onwards, one obtains the set of guarantees

$$\mathcal{U}(x) = \Big\{ \big(\max(-x + 0.5\, Q_1(1),\, x + 0.5\, Q_1(2)), \\ \max(1 - x + 0.5\, Q_2(1),\, x - 1 + 0.5\, Q_2(2)) \big) : \mathbf{Q}(1),\, \mathbf{Q}(2) \in \mathcal{V}^* \Big\}. \tag{EC.25}$$

If we denote the set $(-x, 1-x) + 0.5\mathcal{V}^*$ (which is obtained by mapping each element $\mathbf{u}$ of $\mathcal{V}^*$ to $(-x, 1-x) + 0.5\mathbf{u}$)), by $\mathcal{U}_1(x)$, and the set $(x, x-1) + 0.5\mathcal{V}^*$ by $\mathcal{U}_2(x)$, it is straightforward to see that

$$up(\mathcal{U}(x)) = up(\mathcal{U}_1(x)) \cap up(\mathcal{U}_2(x)),$$

where $up(.)$ is the upset of the set in $[0,2]^2$. This is depicted in Figure EC.5. The fixed point relation says that

$$\mathcal{V}^* = \Lambda\Big(\cup_{x\in[0,1]}\, up(\mathcal{U}(x)) \Big).$$

From the figure, one can see that $\mathcal{V}^*$ is the curve traced by the lower left corner point of $up(\mathcal{U}(x))$ as x varies between 0 and 1. Since we already know that the two extreme points on $\mathcal{V}^*$ are $(0,2)$ and $(2,0)$, for $x = 0.5$, we know that the lower left corner point of $up(\mathcal{U}(0.5))$ is $(0.5, 0.5)$, and hence is contained in $\mathcal{V}^*$, as shown in Figure EC.6. Since we know that $\mathcal{V}^*$ is symmetric around the line $x = y$,

we know that $f(x) = f^{-1}(x)$, and thus it is sufficient to determine $f(x)$ in the range $x \in [0, 0.5]$. In this range, f satisfies the following fixed point relation (again, see Figure EC.5):

$$\begin{aligned} f(x) &= \bar{f}(2x) + 1 - x \\ &= f(4x)/2 + 1 - x. \end{aligned} \tag{EC.26}$$

Taking the derivative twice on both sides, we obtain:

$$f''(x) = 8f''(4x).$$

This gives us $f''(x) = ax^{-\frac{3}{2}}$ for any $a \in \mathbb{R}$. Integrating, we obtain $f(x) = a\sqrt{x} + x + 2$. Since we want $f(0.5) = 0.5$, we obtain $a = -2\sqrt{2}$. Thus we have $f(x) = -2\sqrt{2x} + x + 2$. Note that $f(2) = 0$, and it turns out that $f(x)$ restricted to the domain $x \in [0, 2]$ is such that $f(x) = f^{-1}(x)$. Thus $f(x)$ is the function we are looking for and $\mathcal{V}^* = \{(x, -2\sqrt{2x} + x + 2) : x \in [0, 2]\}$.

We can compare this exact characterization with the approximate frontier that we computed using our approximation procedure for $\beta = 0.5$ (approximation error less than 0.06). Both these frontiers are plotted in Figure EC.7. As we can observe, the two frontiers are close to identical.

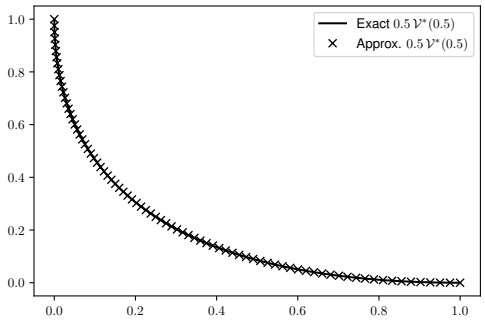

Figure EC.7 Comparison of the approximation of $0.5\mathcal{V}^*(0.5)$ and the exact characterization $0.5\mathcal{V}^*(0.5) = \{(x, -2\sqrt{x} + x + 1) : x \in [0, 1]\}$.

Optimal Policy: To attain the point $(x, f(x))$ for $x \in [0, 0.5]$, the optimal strategy of Alice chooses a randomization $(1 - x, x)$; then if the adversary chooses action 1, the next point she chooses

to attain is $(4x, f(4x))$, whereas if he chooses action 2 then the next point she chooses to attain is $(0, f(0))$. To attain the point $(f(x), x)$ for $x \in [0, 0.5]$, the optimal strategy of Alice chooses a randomization $(x, 1-x)$; then if the adversary chooses action 2, the next point she chooses to attain is $(f(4x), 4x)$, whereas if he chooses action 1, then the next point she chooses to attain is $(f(0), 0)$.

EC.9. Zero-sum repeated games with scalar losses: a review of results

The scalar counterpart of the vector-valued repeated game we study in the paper is a relatively much simpler object of study. Preserving the notation in the paper, suppose that in the single stage game $\mathbb{G}$, a pair of actions $a \in A$ for Alice and $b \in B$ for Bob leads Alice to incur a scalar loss $r(a, b)$. The value of game $\mathbb{G}$ is then defined to be,

$$v^* = \min_{\boldsymbol{\alpha} \in \Delta(A)} \max_{b \in B} \sum_{a \in A} \alpha_a r(a, b) \overset{(a)}{=} \max_{\boldsymbol{\nu} \in \Delta(B)} \min_{a \in A} \sum_{b \in B} \nu_b r(a, b), \tag{EC.27}$$

where the equality (a) follows from the Von Neumann minmax theorem [38]. It is well-known that both the above optimization problems can be solved as linear programs. For example, the minmax optimization problem for Alice can be solved as the following linear program.

$$\min v \tag{EC.28a}$$

$$v \geq \sum_{a \in A} \alpha_a r(a, b), \text{ for all } b \in B, \tag{EC.28b}$$

$$\boldsymbol{\alpha} \in \Delta(A). \tag{EC.28c}$$

Let $\boldsymbol{\alpha}^*$ be an arbitrary minmax optimal strategy for Alice and $\boldsymbol{\nu}^*$ be an arbitrary maxmin optimal strategy for Bob in $\mathbb{G}$ (since optimal strategies may not be unique).

Now consider a repeated game $\mathbb{G}^T$, in which $\mathbb{G}$ is repeated T times, with the cumulative loss of $\mathbb{G}^T$ defined to be

$$\frac{\sum_{t=1}^{T} \beta^{t-1} r(a_t, b_t)}{\sum_{t=1}^{T} \beta^{t-1}},$$

which is the average discounted loss with a discount factor $\beta \in [0, 1]$ (we obtain the simple average loss for $\beta = 1$). It is straightforward to argue that the smallest upper bound on the expected loss that

Alice can guarantee in $\mathbb{G}^T$ is simply v^*, i.e., the value of the game $\mathbb{G}$. Or in other words, the value of the game $\mathbb{G}^T$ is v^* for any T and $\beta \in [0,1]$. To see this, note that by playing $\boldsymbol{\alpha}^*$ in every repetition, Alice can guarantee that the expected stage loss in every repetition is *at most* v^*. Similarly, by playing $\boldsymbol{\nu}^*$ all the time, Bob can guarantee that the expected stage loss in every repetition is *at least* v^*. Thus irrespective of how one averages the daily losses, the optimal guarantee on the average loss that Alice can guarantee is v^*. And the strategy that achieves this guarantee is the one where Alice plays any equilibrium strategy of $\mathbb{G}$ in each repetition. Now consider the game $\mathbb{G}^\infty$, in which $\mathbb{G}$ is repeated infinitely often, with the cumulative loss defined to be

$$(1-\beta)\sum_{t=1}^{T}\beta^{t-1}r(a_t,b_t)$$

for $\beta \in [0,1)$. The previous argument extends to this game as well and we can conclude that the optimal guarantee on the average loss that Alice can guarantee is again v^*.

EC.9.1. Simultaneous vector guarantees vs. guarantees on the combined scalar loss

In the present paper, we concerned ourselves with characterizing the best *simultaneous guarantees* that Alice can guarantee across the vector components of the losses. A natural question is whether the frontier of such simultaneous guarantees can be characterized by solving a set of scalar repeated games, where each game is obtained from a different weighted combination of the vector losses. The example below shows that this is not the case. The key point is that optimal strategies that achieve simultaneous guarantees across the different dimensions must adapt to the evolution of the profile of losses across the different dimensions over time. If one dimension suffers excessive losses, these strategies must shift to focusing on minimizing losses on that dimension. This profile information is lost when one combines the losses into a single scalar. In other words, when we combine the dimensions, the optimal strategy in the resulting scalar repeated game only guarantees an upper bound on the combined loss, as opposed to simultaneously guaranteeing upper bounds on losses across the different dimensions. Hence, one must directly address the multi-dimensional nature of the game as we do in the paper as opposed to attempting a reduction from the scalar case.

Example. Consider the game with vector losses shown in Figure EC.8. This is the game that corresponds to the single-stage vector regrets in the expert selection problem with $K = 2$ experts. If the losses are combined across the two dimensions with weights $(\alpha, 1-\alpha)$, then the resulting game is shown in Figure EC.9. In this game, the unique minmax optimal strategy for Alice is to play action 1 with probability α and action 2 with probability $1-\alpha$. Now suppose that this scalar game is repeated infinitely often, with losses in the n^{th} repetition discounted by $(1-\beta)\beta^{n-1}$. In this repeated game, as we argued above, the unique minmax optimal strategy for Alice is to play $(\alpha, 1-\alpha)$ in every repetition. Against this strategy, the total discounted vector losses corresponding to the two actions (also repeated forever) of the adversary are depicted in Figure EC.10. Considering worst-case choices of the adversary, we can deduce that Alice guarantees a maximum loss of $1-\alpha$ on dimension 1 and α on dimension 2, i.e., this strategy achieves the vector guarantee $(1-\alpha, \alpha)$. Let $\mathcal{V}_s^*$ denote the set of vector guarantees achievable using this scalar reduction by varying α. $\mathcal{V}_s^*$ is shown in Figure EC.11: it is simply the line segment joining $(0,1)$ and $(1,0)$, irrespective of the value of β. However, note that we have shown in Section EC.8 above that we can achieve significantly better guarantees for $\beta = 0.5$; see Figure EC.7. This demonstrates that the optimal guarantees cannot be achieved via a scalar reduction of the vector-valued game.

EC.10. An application to zero-sum repeated games with lack of information on one side

One of the most celebrated and well-studied models of dynamic games is the model of zero-sum repeated games with incomplete information on one side, which was introduced to study nuclear disarmament in the cold-war era by Aumann and Maschler. While the analysis was classified at the time, it was later declassified and published [8]. The high-level goal of this model is to understand how a player (such as the leadership of a country) can learn about her opponent's (some other country's) unknown preferences in a dynamic game (modeling a political scenario such as mutual disarmament) by observing her actions (e.g., proposals it agrees to), and then leverage that information to play better, while the other player rationally responds to this possibility in choosing her actions. The formal model is described as follows.

Alice \ Bob	1	2
1	$(0,1)$	$(0,-1)$
2	$(-1,0)$	$(1,0)$

Figure EC.8 A game with vector losses.

Alice \ Bob	1	2
1	$1-\alpha$	$-(1-\alpha)$
2	$-\alpha$	α

Figure EC.9 A scalar game corresponding to weights $(\alpha, 1-\alpha)$.

	Bob 1	Bob 2
	$(-(1-\alpha), \alpha)$	$(1-\alpha, -\alpha)$

Figure EC.10 Vector losses as a function of Bob's actions given Alice's stationary strategy $(\alpha, 1-\alpha)$.

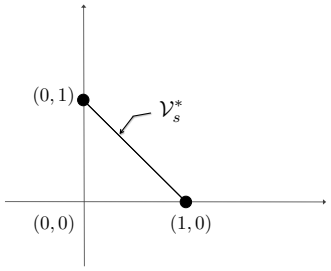

Figure EC.11 The set of vector guarantees on losses achievable by varying α.

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). Let the loss to Alice corresponding to actions a and b of Alice and Bob respectively be denoted by $r_k(a, b)$ in game G_k.

We define the game G^∞ as follows. One of the K games G_k is chosen by nature with probability p_k, such that $\sum_k p_k = 1$. 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 (random) 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 (if she did, the game could potentially be identified immediately), but she observes Bob's actions. An adaptive randomized strategy (also called a behavioral strategy) π_A for Alice specifies for each time t, a mapping from her set of observations till time t, i.e. $H_t^A = (a_1, b_1, \cdots, a_{t-1}, b_{t-1})$, to $\Delta(A)$ (the simplex of probability distributions over A). Similarly, an adaptive randomized strategy

π_B for Bob specifies for each time t a mapping from his set of observations till time t *and the chosen game* G, i.e., $H_t^B = (G, a_1, b_1, \cdots, a_{t-1}, b_{t-1})$, to $\Delta(B)$. We will express the strategy π_B of Bob as $\pi_B = (\pi_B^k; k = 1, \cdots, K)$, where π_B^k is his strategy conditioned on the event $\{G = G_k\}$.

We now specify the objectives of the two players in G^∞. For a discount factor $\beta \in (0,1)$ and for a choice of strategies π_A and π_B of the two players, the ex-ante expected loss is given by

$$R(\pi_A, \pi_B) = \mathrm{E}_{\pi_A, \pi_B, G}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_G(a_t, b_t)\right] = \sum_{k=1}^{K} p_k \mathrm{E}_{\pi_A, \pi_B^k}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_k(a_t, b_t)\right]. \quad \text{(EC.29)}$$

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

$$\overline{\mathbf{V}} = \min_{\pi_A} \sum_{k=1}^{K} p_k \max_{\pi_B^k} \mathrm{E}_{\pi_A, \pi_B^k}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_k(a_t, b_t)\right]. \quad \text{(EC.30)}$$

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

$$\underline{\mathbf{V}} = \max_{(\pi_B^k; k=1,\cdots,K)} \min_{\pi_A} \left(\sum_{k=1}^{K} p_k \mathrm{E}_{\pi_A, \pi_B^k}\left[\sum_{t=1}^{\infty} \beta^{t-1} r_k(a_t, b_t)\right]\right). \quad \text{(EC.31)}$$

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 $\overline{\mathbf{V}} = \underline{\mathbf{V}}$ [7, 45, 51].

Characterizing the maxmin optimal policy for Bob: To characterize and compute the maxmin policy for Bob, one can use a dynamic programming approach that exploits the structural relationship between the original game and the game after one stage has elapsed [45, 51]. Suppose $V(\mathbf{p})$ is a function that assigns to every prior probability distribution $\mathbf{p}$ over the game G, the maxmin value of the associated infinitely repeated game. Then it is easy to show [45, 51] that the maxmin value as a function of the prior $\mathbf{p}$ is the fixed point of the following contractive dynamic

programming operator defined on the function $V : \Delta^K \to \mathbb{R}$ (here Δ^K is the $K-1$ dimensional unit simplex) :

$$\mathbf{T}(V)(\mathbf{p}) = \max_{(\mathbf{q}_B^k \in \Delta(B); k=1,\cdots,K)} \min_{\mathbf{q}_A \in \Delta(A)} \sum_{k=1}^{K} p_k \mathrm{E}_{\mathbf{q}_B^k, \mathbf{q}_A}[r_k(a,b)] + \sum_{b \in B} \beta(\sum_{k=1}^{K} p_k q_B^k(b)) V\left(\left(\frac{p_k q_B^k(b)}{\sum_{k=1}^{K} p_k q_B^k(b)}; k = 1, \cdots, K \right) \right).$$

This operator can be understood as follows. Notice that in the first stage, the probability distributions over Bob's actions chosen by him for each of the K games as a part of his strategy π_B makes his realized action a (potentially) informative signal of the true game chosen by nature. Since Alice can be assumed to know this strategy in the computation of the maxmin, she can perform a Bayesian update of her belief about the chosen game after having observed Bob's action. Thus once the randomization of Bob in the first stage is fixed, there is a one-stage expected loss that is minimized by Alice, and then every realized action of Bob results in a new game with an associated maxmin value, that is identical in structure to the original game, except that the original prior is replaced with the posterior distribution conditional on that action (note that the state transitions in the space of posteriors are independent of Alice's actions). Bob thus chooses a randomization over his actions for each of the K games that maximize the sum of these two values. Consistency then requires that the function $V(\mathbf{p})$ has to be the fixed point of this resulting operator. It also follows that the optimal policy for Bob is a stationary policy that depends only on the posterior p_t at stage t.

Characterizing the minmax optimal policy for Alice: Now it turns out that a similar approach cannot be used to compute and characterize the minmax optimal policy for Alice, the uninformed player. The problem is that in order to perform the Bayesian update as a part of her policy π_A, Alice needs to know Bob's policy π_B, which means that π_A presupposes the knowledge of π_B, 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 instead depend in some form on Bob's actions and the corresponding losses incurred in the different possible choices of games.

In the case of the limiting average objective with undiscounted losses, Alice's optimal policy is derived using Blackwell approachability theory [8, 14]. However, to the best of our knowledge, the exact computation or even approximation of Alice's optimal policy has been an open problem in the case of discounted losses, essentially because the counterpart of Blackwell approachability theory for the case of discounted losses has been missing in the literature. Structurally, all that is known (see Corollary 3.25 in [45]) about the optimal policy is that Alice's decision at stage t doesn't depend on her own past actions, but can depend on potentially all of Bob's actions till time t. This suggests the possibility that any dynamic programming-based procedure to compute this policy may suffer from the curse of dimensionality, i.e., the state may include the entire history of actions.

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

$$\mathcal{W}=\left\{\left(\max_{\pi_B^k} E\left[\sum_{t=1}^{\infty}\beta^{t-1}r_k(a_t,b_t)\right];k=1,\cdots,K\right):\pi_A\in\Pi_A\right\}. \tag{EC.32}$$

This is the set of upper bounds on losses that Alice can simultaneously achieve on the K components of the vector of the long term discounted losses, by playing all the possible strategies in π_A. In fact, one need not compute the entire set $\mathcal{W}$, but just its lower Pareto frontier $\Lambda(\mathcal{W})$. If we determine this set, then one can simply choose a point $\mathbf{r}(\mathbf{p})\in\Lambda(\mathcal{W})$ such that

$$\mathbf{r}(\mathbf{p})=\arg\min_{\mathbf{r}\in\Lambda(\mathcal{W})}\sum_{k=1}^{K}p_k r_k. \tag{EC.33}$$

The corresponding strategy of Alice that results in the simultaneous guarantee $\mathbf{r}$ is then the optimal policy in the original game. Thus we are interested in characterizing the set $\mathcal{V}^*=\Lambda(\mathcal{W})$. But this is exactly the set we characterized using the set-valued dynamic programming approach.

Using our approach, first, we immediately deduce the known fact that the optimal strategy of Alice doesn't depend on her own past actions (Theorem 3). Second, we also obtain a previously

unknown insight into its structure: Alice's optimal strategy is stationary relative to a compact state space as described in Theorem 3. This state space can be compared to the compact state space of Bob's optimal policy, which is the space of probability distributions on the K games. It is interesting to note that the state transitions for Bob's policy are exogenously determined by the Bayesian updating of the posterior, whereas the state transitions for Alice's policy are endogenously determined by the dynamic programming operator (as we had remarked earlier in Section 4.3). Third, the compactness of the state space opens up the possibility to approximate Alice's optimal strategy, as we have demonstrated in Section 5. To the best of our knowledge, this is the first known approximation procedure for this problem.

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 $\mathbf{p}$, 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 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 $\mathbf{p}$ and to compute the optimal strategy, one needs to simultaneously solve for games starting from all possible priors $\mathbf{p} \in \Delta^K$.

References

[1] Jacob Abernethy, Manfred K Warmuth, and Joel Yellin. Optimal strategies from random walks. In *Proceedings of The 21st Annual Conference on Learning Theory*, 2008.

[2] Dilip Abreu, David Pearce, and Ennio Stacchetti. Optimal cartel equilibria with imperfect monitoring. *Journal of Economic Theory*, 39(1):251 – 269, 1986.

[3] Dilip Abreu, David Pearce, and Ennio Stacchetti. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica*, 58(5):pp. 1041–1063, 1990.

[4] Christopher Amato, Daniel S Bernstein, and Shlomo Zilberstein. Solving pomdps using quadratically constrained linear programs. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 341–343, 2006.

[5] Christopher Amato, Daniel S Bernstein, and Shlomo Zilberstein. Optimizing fixed-size stochastic controllers for pomdps and decentralized pomdps. *Autonomous Agents and Multi-Agent Systems*, 21(3):293–320, 2010.

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

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

[8] Robert J Aumann, Michael Maschler, and Richard E Stearns. *Repeated games with incomplete information*. MIT press, 1995.

[9] Peter L Bartlett, Wouter M Koolen, Alan Malek, Eiji Takimoto, and Manfred K Warmuth. Minimax fixed-design linear regression. In *Proceedings of The 28th Annual Conference on Learning Theory*, 2015.

[10] Erhan Bayraktar, Ibrahim Ekren, and Yili Zhang. On the asymptotic optimality of the comb strategy for prediction with expert advice. *The Annals of Applied Probability*, 30(6):2517–2546, 2020.

[11] Logan DR Beal, Daniel C Hill, R Abraham Martin, and John D Hedengren. Gekko optimization suite. *Processes*, 6(8):106, 2018.

[12] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, 2005.

[13] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, 2012.

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

[15] Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(Jun):1307–1324, 2007.

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

[17] 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.

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

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

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

[21] Thomas M Cover. Behavior of sequential predictors of binary sequences. Technical report, DTIC Document, 1966.

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

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

[24] Nick Gravin, Yuval Peres, and Balasubramanian Sivan. Towards optimal algorithms for prediction with expert advice. In *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2016.

[25] 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.

[26] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

[27] Jeff Henrikson. Completeness and total boundedness of the Hausdorff metric. In *MIT Undergraduate Journal of Mathematics*. Citeseer, 1999.

[28] Wouter M Koolen. The Pareto regret frontier. In *Advances in Neural Information Processing Systems*, 2013.

[29] Wouter M Koolen, Alan Malek, and Peter L Bartlett. Efficient minimax strategies for square loss games. In *Advances in Neural Information Processing Systems*, 2014.

[30] Wouter M Koolen, Alan Malek, Peter L Bartlett, and Yasin Abbasi. Minimax time series prediction. In *Advances in Neural Information Processing Systems*, 2015.

[31] JB Kruskal. Two convex counterexamples: A discontinuous envelope function and a nondifferentiable nearest-point mapping. *Proceedings of the American Mathematical Society*, 23(3):697–703, 1969.

[32] Rida Laraki and Sylvain Sorin. Advances in zero-sum dynamic games. In *Handbook of game theory with economic applications*, volume 4, pages 27–93. Elsevier, 2015.

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

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

[35] Haipeng Luo and Robert Schapire. Towards minimax online learning with unknown time horizon. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[36] Emanuel Milman. Approachable sets of vector payoffs in stochastic games. *Games and Economic Behavior*, 56(1):135–147, 2006.

[37] James R Munkres. Topology: a first course, 1975.

[38] J v Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

[39] Vianney Perchet. Approachability of convex sets in games with partial monitoring. *Journal of Optimization Theory and Applications*, 149(3):665–677, 2011.

[40] Vianney Perchet. Internal regret with partial monitoring: Calibration-based optimal algorithms. *Journal of Machine Learning Research*, 12(Jun):1893–1921, 2011.

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

[42] Vianney Perchet and Marc Quincampoix. On a unified framework for approachability with full or partial monitoring. *Mathematics of Operations Research*, 40(3):596–610, 2014.

[43] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

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

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

[46] Xavier Spinat. A necessary and sufficient condition for approachability. *Mathematics of Operations Research*, 27(1):31–44, 2002.

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

[48] Pravin M Vaidya. Speeding-up linear programming using fast matrix multiplication. In *30th annual symposium on foundations of computer science*, pages 332–337. IEEE Computer Society, 1989.

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

[50] Volodimir G Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.

[51] 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.