Parametric Regret in Uncertain Markov Decision Processes

Huan Xu, and Shie Mannor, *Senior Member, IEEE*

Abstract— **We consider decision making in a Markovian setup where the reward parameters are not known in advance. Our performance criterion is the gap between the performance of the best strategy that is chosen after the true parameter realization is revealed and the performance of the strategy that is chosen before the parameter realization is revealed. We call this gap the parametric regret. We consider two related problems: minimax regret and mean-variance tradeoff of the regret. The minimax regret strategy minimizes the worst-case regret under the most adversarial possible realization. We show that the problem of computing the minimax regret strategy is NP-hard and propose algorithms to efficiently solve it under favorable conditions. The mean-variance tradeoff formulation requires a probabilistic model of the uncertain parameters and looks for a strategy that minimizes a convex combination of the mean and the variance of the regret. We prove that computing such a strategy can be done numerically in an efficient way.**

I. Introduction

Sequential decision making in stochastic dynamic environments is often modeled using Markov Decision Processes (MDP, cf [1], [2]). In the *standard setup*, each strategy is evaluated according to its *performance*, i.e., the expected accumulated reward. The optimal strategy is the one that achieves maximal performance.

In many real applications, the decision maker evaluates strategies in a comparative way. That is, given a strategy, the decision maker is interested in how its performance competes with other strategies rather than the *quantity* of the performance itself. For example, the objective in financial applications such as portfolio optimizations is often to "beat the market", i.e., to perform favorably than a strategy that holds index stocks. The same percentage of growth can be regarded as "incredible success" or ”disastrous failure” purely depending on how others perform in this same market. A natural measurement of strategies in such setup, which we termed *competitive setup* hereafter, is the so-called *parametric regret*: the gap between the performance of a strategy and that of the optimal one. [1]

When the parameters of a MDP are known, minimizing the regret is equivalent to maximizing the performance of a strategy, and hence the competitive setup coincides with the standard setup. However, the formulation of a problem is often subject to *parameter uncertainty* – the deviation of the modeling parameters from the unknown true ones (cf [4]–[7]). In this case, both performance and regret of a strategy are functions of parameter realizations, where in general there is no strategy that is optimal for all parameter realizations.

In the standard setup, there are two formulations to find the "optimal" strategy for MDPs with uncertain parameters. The first formulation [4]–[7] takes a minimax approach, i.e., the true parameters can be any element of a known set, and strategies are evaluated based on the performance under the (respectively) worst possible parameter realization. The second one takes a Bayesian approach (e.g. [8]): The true parameters are regarded as random variables. Thus, given a strategy, its performance is a random variable whose probability distribution can be obtained. And the optimal strategy is the one that maximizes certain risk measure such as percentile loss [8] or mean-variance tradeoff [9].

In this paper we adapt the aforementioned formulations into the competitive setup and discuss parametric regret minimizing in uncertain Markov decision processes. In particular, our contributions include the following:

- In Section II we follow the minimax approach and propose the Minimal Maximum Regret (MMR) decision criterion.
- We show in Section III that finding the MMR strategy is NP-hard in general.
- We investigate the algorithmic aspect of MMR strategy in Section IV. In particular, we propose in Section IV-A an algorithm based on mixed integer program that solves the MMR strategy, and discuss in the rest of Section IV two special cases where the MMR strategy can be found in polynomial time.
- We take a Bayesian approach and propose the Optimal Mean-Variance Tradeoff of Regret criterion in Section V. We further show that such formulation can be converted into a quadratic program on a polytope, and hence solved efficiently.

We need to point out that in this paper we concentrate on the case that the system dynamics is known and only reward parameters are subject to uncertainty, partly due to the prohibitive computational cost. Indeed, as shown in Section III, even in this seemingly simple case finding the MMR strategy is NP-hard. In addition, the known-dynamics case can either model or approximate many practical problems. For instance, a shortest-path problem with uncertain link lengths is an uncertain MDP with known dynamics (e.g., [1]). Another example is using state aggregation to solve large scale MDPs [10]. In such case, states are grouped to a

H. Xu is a Ph.D. student at Department of Electrical and Computer Engineering, McGill University, Canada. *Email: xuhuan@cim.mcgill.ca*

S. Mannor is with Department of Electrical and Computer Engineering, McGill University, Canada and Department of Electrical Engineering, Technion, Israel. *Email: shie.mannor@mcgill.ca; shie@ee.technion.ac.il*

[1] We will use "regret" in the following for simplicity of the expression. However, it should be noted that this is different from the standard notion of regret in online learning - the gap between the average reward of a learning algorithm and the optimal strategy [3].

Preprint submitted to 48th IEEE Conference on Decision and Control.
Received March 6, 2009.

small number of hyper-states and a reduced MDP built on these hyper-states is investigated. Typically, the transition law between hyper-states are known, but the expected reward visiting each hyper-state is uncertain due to the transitions inside each hyper-state.

A. Preliminaries and Notations

Throughout the paper, boldface letters are used for column vectors, where its elements are represented using the same but non-boldfaced letter. For example, the first element of a vector $\mathbf{v}$ is denoted as v_1. Given a function $f(\mathbf{x})$ not necessarily differentiable, we use $\nabla f(\mathbf{x})|_{\mathbf{x}_0}$ to represent the set of subgradients at point $\mathbf{x}_0$.

An uncertain Markov Decision Process (*uMDP*) is a 6-tuple $< T, \gamma, S, A, \mathbf{p}, \mathcal{R} >$ where:

- T is the (possibly infinite) decision horizon;
- $\gamma \in (0, 1]$ is the discount factor. We allow $\gamma = 1$ only when T is finite.
- S is the state set and A is the action set. Both sets are finite.
- $\mathbf{p}$ is the transition probability i.e., $p(s'|s, a)$ is the probability to reach state s' from a state s when action a is taken.
- $\mathcal{R}$ is the admissible set of reward parameter. To be more specific, the reward vector $\mathbf{r}$ is unknown to the decision maker (this is why it is called "uncertain MDP"). To make such decision problem meaningful, some a priori information of $\mathbf{r}$ is known: it is an element of $\mathcal{R}$. In the literature of *robust optimization*, $\mathcal{R}$ is often called the *uncertainty set* (cf [4], [11], [12]).

We assume that the initial state distribution is known to be $\boldsymbol{\alpha}$. All history-dependent randomized strategies are admissible, and we denote that set by Π^{HR}. We use Π^S and Π^D to denote the set of stationary Markovian random strategies and stationary Markovian deterministic strategies, respectively. For $\pi \in \Pi^S$, we use $\pi(a|s)$ to represent the probability of choosing $a \in A$ at state s following π. Given a strategy $\pi \in \Pi^{HR}$ and a parameter realization $\mathbf{r} \in \mathcal{R}$, its expected performance (i.e., accumulated discounted reward) is denoted by $P(\pi, \mathbf{r})$, that is

$$P(\pi, \mathbf{r}) \triangleq \mathbb{E}_{\pi}\{\sum_{i=1}^{T} \gamma^{i-1} r(s_i, a_i)\}. \qquad (1)$$

We focus on the case when the uncertainty set $\mathcal{R}$ is a polytope. Polytopes are probably the most "natural" formulation of uncertainty set that can model many widely applicable cases. For example, the interval case, i.e., each reward parameter $r(s, a)$ belongs to an interval, is a polytope. We also assume that $\mathcal{R}$ is bounded, to avoid technical problems such as infinitely large regret.

II. MINIMAX REGRET IN MDPS

In this section we propose the MiniMax Regret criterion, i.e., minimizing the parametric regret under the most adversarial parameter realization.

Definition 2.1: Given a uMDP $< T, \gamma, S, A, \mathbf{p}, \mathcal{R} >$ and $\mathbf{r}_0 \in \mathcal{R}$, the *parametric regret* of a strategy π w.r.t. $\mathbf{r}_0$ is defined as

$$\hat{R}(\pi, \mathbf{r}_0) \triangleq \max_{\pi' \in \Pi^{HR}} \{P(\pi', \mathbf{r}_0) - P(\pi, \mathbf{r}_0)\}.$$

In words, regret is the performance gap between a strategy and the optimal strategy. It is thus a natural performance measure in a competitive environment. Observe that for a fixed $\mathbf{r}_0$, the regret is equivalent to the expected reward up to adding a constant.

Definition 2.2: Given a uMDP $< T, \gamma, S, A, \mathbf{p}, \mathcal{R} >$, the *Maximum Regret* of a strategy π is defined as

$$R(\pi) \triangleq \max_{\mathbf{r} \in \mathcal{R}} \hat{R}(\pi, \mathbf{r}) = \max_{\mathbf{r} \in \mathcal{R}, \pi' \in \Pi^{HR}} \{P(\pi', \mathbf{r}) - P(\pi, \mathbf{r})\}. \qquad (2)$$

The maximum regret is the regret of a strategy under the most adversarial parameter realization. It can also be regarded as the performance gap w.r.t. an "all-mighty" opponent strategy that can observe the parameter realization and select the respective optimal solution.

Definition 2.3: Given a uMDP $< T, \gamma, S, A, \mathbf{p}, \mathcal{R} >$, the *MiniMax Regret* (MMR) strategy is

$$\pi^* \triangleq \arg \min_{\pi \in \Pi^{HR}} R(\pi). \qquad (3)$$

The minimax regret strategy is not the same as the robust MDP (i.e., minimax performance) strategy in general, as shown in the following example: Consider the MDP as shown in Figure 1, where $\mathcal{R} = [0,\, 3] \times [1,\, 2]$. Observe that the minimax performance strategy is selecting $a2$, whose maximum regret equals 2. On the other hand, the minimax regret strategy is selecting either action with probability 50%, whose maximum regret is 1.

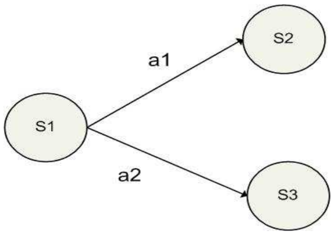

Fig. 1. An example where the MMR policy is different than the robust policy.

A. Existence of stationary optimal solution

Although the definition of MMR considers history dependent strategies, in this subsection we show that without loss of generality we can concentrate on Π^S because there exists a stationary MMR strategy. We need the following lemma first.

Lemma 2.4: Given $\pi_0 \in \Pi^{HR}$, there exists $\hat{\pi} \in \Pi^S$ such that $R(\hat{\pi}) = R(\pi_0)$.

Proof: It is well known that (e.g., [1]) given $\pi_0 \in \Pi^{HR}$, there exists $\hat{\pi} \in \Pi^S$ such that $\forall s \in S, a \in A$

$$\mathbb{E}_{\pi_0} \sum_i \gamma^{i-1} \mathbf{1}(s_i = s, a_i = a) \equiv \mathbb{E}_{\hat{\pi}} \sum_i \gamma^{i-1} \mathbf{1}(s_i = s, a_i = a).$$

Note that the following holds for any $\pi \in \Pi^{HR}$,

$$P(\pi, \mathbf{r}) = \sum_{s\in S}\sum_{a\in A}\{r(s,a)\mathbb{E}_\pi \sum_{i} \gamma^{i-1}\mathbf{1}(s_i = s, a_i = a)\}.$$

Hence,

$$P(\pi', \mathbf{r}) - P(\pi_0, \mathbf{r}) = P(\pi', \mathbf{r}) - P(\hat{\pi}, \mathbf{r}),\ \forall \mathbf{r} \in \mathcal{R}, \pi' \in \Pi^{HR}.$$

Taking maximization over π' and $\mathbf{r}$ establishes the lemma. ■

We now present the main theorem of this subsection: the existence of a stationary MMR strategy.

Theorem 2.5: There exists $\pi^* \in \Pi^S$ such that $R(\pi^*) \leq R(\pi),\ \forall \pi \in \Pi^{HR}$.

Proof: From Lemma 2.4, it suffices to prove that $R(\pi^*) \leq R(\pi),\ \forall \pi \in \Pi^S$. We define a metric $d(\pi_1, \pi_2) \triangleq \max_{s\in S, a\in A} |\pi_1(a|s) - \pi_2(a|s)|$ on Π^S and note that since S and A are finite, the set Π^S is compact. Let sequence $\{\pi_n\} \subseteq \Pi^S$ be such that $R(\pi_n) \to \inf_{\pi\in\Pi^{HR}} R(\pi)$. Due to compactness of Π^S, taking a convergent subsequence $\{\pi_{m_n}\}$ and let $\pi^* \in \Pi^S$ be its limiting point. Let

$$(\hat{\pi}', \hat{\mathbf{r}}) = \arg\max_{(\pi', \mathbf{r})}\{P(\pi', \mathbf{r}) - P(\pi^*, \mathbf{r})\}.$$

By definition of maximum regret we have

$$R(\pi_{m_n}) \geq P(\hat{\pi}', \hat{\mathbf{r}}) - P(\pi_{m_n}, \hat{\mathbf{r}}),\ \forall n.$$

Take limits on both sides and note that $P(\hat{\pi}', \hat{\mathbf{r}}) - P(\pi, \hat{\mathbf{r}})$ is a continuous function of π w.r.t. the aforementioned metric, we have

$$\inf_{\pi\in\Pi^{HR}} R(\pi) \geq R(\pi^*),$$

which establishes the theorem. ■

III. Computational complexity

This section investigates the computational complexity of MMR strategy. We show that the MMR strategy is in general intractable. In fact, even evaluating the maximum regret for a given strategy can be NP-hard, as shown in the next theorem.

Theorem 3.1: Let $\mathcal{R}$ be a polytope defined by a set of n linear inequalities. Then evaluating the maximum regret of a strategy is NP-complete with respect to $|S|$, $|A|$ and n.

Proof: We first show that evaluating the maximum regret can not be computationally more difficult than NP. This is due to the fact that evaluating the regret of a given strategy $\hat{\pi}$ can be written as the following optimization problem on $(\mathbf{x}', \mathbf{r})$:

$$\begin{aligned}
\text{max:}\ & \sum_{a\in A}\sum_{s\in S}\{r(s,a)x'(s,a) - r(s,a)\hat{x}(s,a)\} \\
\text{s.t. :}\ & \sum_{a\in A} x'(s',a) - \sum_{s\in S}\sum_{a\in A}\gamma p(s'|s,a)x'(s,a) = \alpha(s'),\ \forall s', \\
& x'(s,a) \geq 0, \quad \forall s, \forall a, \\
& \mathbf{r} \in \mathcal{R}.
\end{aligned} \tag{4}$$

where $\hat{x}(s,a)$ is given by the $\sum_{i=1}^{T}\gamma^{i-1}\mathbb{E}(\mathbf{1}_{s_i=s, a_i=a})$ under $\hat{\pi}$. Note that Equation (4) is a (non-convex) quadratic program which is known to be equivalent to NP. Hence, evaluating the maximum regret can not be computationally more difficult than NP.

Next we prove that evaluating the maximum regret is NP-hard by showing that the *integer feasibility problem*, which is known to be NP hard (e.g., [13]), can be reduced to evaluating maximum regret for a given strategy.

The integer feasibility problem is to tell for $H \in \mathbb{R}^{m\times n}$ and $\mathbf{t} \in \mathbb{R}^m$, whether there exist a vector $\mathbf{x} \in \{0,1\}^n$ such that $H\mathbf{x} \leq \mathbf{t}$. Now consider the following MDP:

Let $\mathbf{r}_a$ denote the vector form of r_{ai} and let $\mathcal{R}$ be defined by the following linear equalities/inequlities:

$$\begin{aligned}
r_{ai} &= -1 - r_{bi},\ i = 1, \cdots n \\
-1 \leq r_{ai} &\leq 0,\ i = 1, \cdots n \\
r_0 &= -1, \\
-H\mathbf{r}_a &\leq \mathbf{t}.
\end{aligned}$$

We claim that the integer feasibility problem is equivalent to whether the maximum regret of action b_0 is 1. Suppose the maximum regret is 1. Note that all rewards are negative and the performance of b_0 does not depend on the reward realization. Hence there exists $(\pi, \mathbf{r})$ such that $P(\pi, \mathbf{r}) = 0$, which means that the expected reward from s_i must be zero for all $i = 1, \cdots n$. Therefore, either r_{ai} or r_{bi} must equal to zero, i.e., $-r_{ai} \in \{0, 1\}$. Thus, let $x_i = -r_{ai}$, the integer feasibility problem has an affirmative answer. Now suppose that the integer feasibility problem has an affirmative answer, i.e., there exists $\mathbf{x}$ satisfying the integer feasibility, let $r_{ai} = -x_i$. Hence either r_{ai} or r_{bi} equals to zero, and the maximum expected reward equals to zero, i.e., the maximum regret of b_0 is 1. Therefore, we reduce the integer feasibility problem to evaluation of the maximum regret, and hence the latter is NP hard.

Combining the two steps, we conclude that evaluating the maximum regret is NP complete. ■

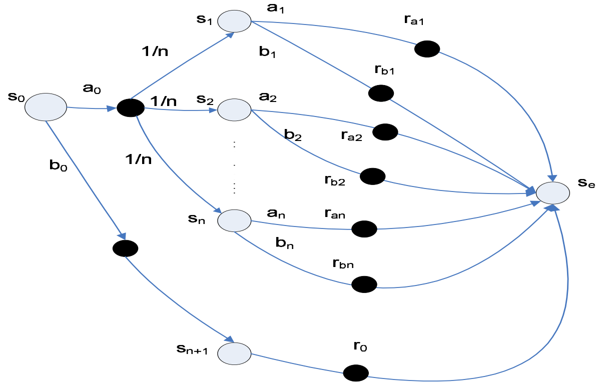

Fig. 2. NP-hard regret evaluation.

IV. Algorithms for finding the MMR solution

Although the MMR solution is generally intractable, we propose in this section several ways to find the MMR strategy. In Subsection IV-A we propose a subgradient method to find MMR, where the subgradient in each step is evaluate by a Mixed Integer Program (MIP). Due to the NP-hardness of MIP, such an algorithm is inherently non-polynomial. We further consider two special cases where polynomial algorithms are possible. (1) In Section IV-B we show that when the number of vertices of $\mathcal{R}$ is small, i.e., $\mathcal{R}$ is the convex hull of a small number of parameter realizations, we can find MMR in polynomial time by solving a linear program. (2) In Section IV-C we show that the MMR has a special property: it is a randomization of "efficient" (defined subsequently) strategies. Furthermore, the weighting coefficients of this randomization can be obtained by LP. Thus we are able to solve MMR in an efficient way if the set of "efficient" strategy, which can be found using action elimination methods, contains a small number of elements.

A. Subgradient approach

In this subsection, we propose a subgradient method to find the MMR solution. The subgradient for each step is indeed the reward parameter that achieves the maximum regret. We further provide an "oracle" based on mixed integer programming that computes this subgradient. This method is non-polynomial, due to the inherent NP-hardness of the problem as shown in Section III.

We first show that minimizing the maximum regret is indeed a convex program (w.r.t. an equivalent form the the decision variable π). Thus, the global optimum (i.e., the MMR strategy) can be found with a subgradient descent/projection method.

Recall the well-known equivalence between a strategy of MDP and its expected state-action frequency (cf [1]). We thus change the decision variable $\pi \in \Pi^{HR}$ to its state-action frequency vector $\mathbf{x}$, i.e., the vector form of $x(s,a) = \mathbb{E}_\pi \sum_{i=1}^{\infty} \gamma^{i-1}\mathbf{1}(s_i = s, a_i = a)$, and recast finding MMR strategy as the following minimization problem on $\mathbf{x}$.

$$\min_{\mathbf{x}\in\mathcal{X}} G(\mathbf{x}). \tag{5}$$

Here, $\mathcal{X}$ is the state-action polytope:

$$\begin{aligned} \mathcal{X}: \sum_{a\in A_{s'}} x(s',a) - \sum_{s\in S}\sum_{a\in A_s} \gamma p(s'|s,a)x(s,a) = \alpha(s'),\ \forall s';\\ x(s,a) \geq 0, \quad \forall s, \forall a; \end{aligned}$$

and $G(\cdot): \mathcal{X} \to \mathbb{R}$ is defined by

$$G(\mathbf{x}) \triangleq \max_{\mathbf{r}\in\mathcal{R},\mathbf{x}'\in\mathcal{X}} (\mathbf{r}^\top\mathbf{x}' - \mathbf{r}^\top\mathbf{x}).$$

Theorem 4.1: 1) Problem (5) is a convex program;
2) Given $\mathbf{x}_0 \in \mathcal{X}$,

$$-\arg\max_{\mathbf{r}\in\mathcal{R}} \big\{ \max_{\mathbf{x}'\in\mathcal{X}} (\mathbf{r}^\top\mathbf{x}' - \mathbf{r}^\top\mathbf{x}_0) \big\} \in \nabla G(\mathbf{x})|_{\mathbf{x}_0}.$$

Proof: Observe that $\mathcal{X}$ is convex. To see that the objective function (i.e., the part inside the curled bracket) is convex, we note that for a fixed pair of $(\mathbf{r}, \mathbf{x}')$, function $(\mathbf{r}^\top\mathbf{x}' - \mathbf{r}^\top\mathbf{x})$ is affine. Therefore the objective function is the maximum over a class of affine functions and hence convex. The second claim follows from the Envelope Theorem (e.g., [14]). ∎

Therefore, we propose here a subgradient descent/project algorithm.

Algorithm 4.2:

```
1) Initialize. $n := 1$; choose $\mathbf{r}_0 \in \mathcal{R}$, $\mathbf{x}^* := \arg\max_{\mathbf{x}\in\mathcal{X}} \mathbf{r}_0^\top \mathbf{x}$.
2) Oracle. Solve $\mathbf{r}^* := \arg\max_{\mathbf{r}\in\mathcal{R}} \big\{ \max_{\mathbf{x}'\in\mathcal{X}} (\mathbf{r}^\top\mathbf{x}' - \mathbf{r}^\top\mathbf{x}^*) \big\}$.
3) Descent. $\hat{\mathbf{x}} := \mathbf{x}^* + \frac{\mathbf{r}^*}{n}$.
4) Projection. Solve $\mathbf{x}^* := \arg\max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \hat{\mathbf{x}}\|$.
5) $n := n + 1$. Go to Step 2.
```

Note that the Projection step is a convex quadratic program over a polytope, which can be solved in polynomial time. In contrast, Step 2 is NP-hard as shown in Section III. We thus propose a MIP formulation that finds $\arg\max_{\mathbf{r}\in\mathcal{R}} \big\{ \max_{\mathbf{x}'\in\mathcal{X}} (\mathbf{r}^\top\mathbf{x}' - \mathbf{r}^\top\mathbf{x}^*) \big\}$.

The formulation is based on a "large M" method. Define

$$\begin{aligned} r_{\max} &\triangleq \sup_{\mathbf{r}\in\mathcal{R}} \max_{s\in S, a\in A} r(s,a),\\ M &\triangleq r_{\max}\big(\sum_{i=1}^{T} \gamma^{i-1}\big). \end{aligned}$$

Note that $r_{\max}$ is finite since $\mathcal{R}$ is bounded, and $\sum_{i=1}^{T}\gamma^{i-1}$ is finite because $\gamma = 1$ only when T is finite. Observe that M is larger than or equal to the reward-to-go for any $s \in S$, $\pi \in \Pi^{HR}$ and $\mathbf{r} \in \mathcal{R}$.

Theorem 4.3: Given initial state distribution $\boldsymbol{\alpha}$ and $\mathbf{x}^*$, let $\mathbf{r}^*$ be the optimal solution of the following maximization problem on $(\mathbf{z}, \mathbf{v}, \mathbf{q}, \mathbf{r})$,

$$\begin{aligned} \text{max:}\ & \sum_s \alpha(s)v(s) - \sum_{s\in S}\sum_{a\in A} r(s,a)x^*(s,a)\\ \text{S.T.:}\ & \sum_{a\in A} z_{s,a} = 1, \quad \forall s \in S,\\ & \left.\begin{array}{l} q(s,a) = r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a)v(s'),\\ v(s) \geq q(s,a),\\ v(s) \leq M(1 - z_{s,a}) + q(s,a),\\ z_{s,a} \in \{0,1\}, \end{array}\right\} \forall s,a\\ & \mathbf{r} \in \mathcal{R}. \end{aligned} \tag{6}$$

We have $\mathbf{r}^* = \arg\max_{\mathbf{r}\in\mathcal{R}} \big\{ \max_{\mathbf{x}'\in\mathcal{X}} (\mathbf{r}^\top\mathbf{x}' - \mathbf{r}^\top\mathbf{x}^*) \big\}$.

Proof: We establish the following lemma first.

Lemma 4.4: Fix $\mathbf{r}$, the following set of constraints

$$\begin{aligned} v(s) &= \max_{a\in A} q(s,a);\\ q(s,a) &= r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a)v(s'). \end{aligned} \tag{7}$$

is equivalent to

$$\begin{aligned} & \sum_{a\in A} z_{s,a} = 1, \quad \forall s \in S,\\ & \left.\begin{array}{l} q(s,a) = r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a)v(s'),\\ v(s) \geq q(s,a),\\ v(s) \leq M(1 - z_{s,a}) + q(s,a),\\ z_{s,a} \in \{0,1\}, \end{array}\right\} \forall s,a. \end{aligned} \tag{8}$$

Proof: First note that since M is larger than or equal to the reward to go of any s, π and $\mathbf{r}$, any $\mathbf{v}, \mathbf{q}$ that satisfy (7) also satisfy (8). (Let $z_{s,a^*} = 1$ when a^* maximizes $q(s, \cdot)$. If multiple a^* exist, arbitrarily pick one.)

Now consider any $\mathbf{q}, \mathbf{v}, \mathbf{z}$ satisfying (8). Fix a s. We have $v(s) \le q(s, a^*)$ for some $a^* \in A$. This is because $z(s, a) \in \{0,\ 1\}$ and $\sum_a z(s, a) = 1$ implies the existence of a^* such that $z(s, a^*) = 1$. Thus,

$$v(s) \le M(1 - z_{s,a^*}) + q(s, a^*) = q(s, a^*).$$

Combining this with $v(s) \ge q(s, a)$ for all $a \in A$ implies that $v(s) = \max_{a \in A} q(s, a)$. Therefore, $(\mathbf{q}, \mathbf{v})$ satisfies Equation (7). ■

We now prove the theorem. Note that for a fixed $\mathbf{r}$, (7) uniquely determines the reward-to-go $\mathbf{v}$ [1]. Thus, the unique solution of (8) is the reward-to-go and hence $\sum_s \alpha(s)v(s)$ is the expected performance under $\mathbf{r}$. This implies that $\mathbf{r}^* = \arg\max_{\mathbf{r}\in\mathcal{R}} \big\{ \max_{\mathbf{x}'\in\mathcal{X}} (\mathbf{r}^\top \mathbf{x}' - \mathbf{r}^\top \mathbf{x}^*) \big\}$. ■

B. Vertices approach

We consider a special type of uMDP: the uncertainty set $\mathcal{R}$ has a small number of vertices. That is, there exists $\mathbf{r}_1, \cdots, \mathbf{r}_t$ such that

$$\mathcal{R} = \mathrm{conv}\{\mathbf{r}_1, \cdots, \mathbf{r}_t\} \triangleq \Big\{ \sum_{i=1}^{t} c_i \mathbf{r}_i | \sum_{i=1}^{t} c_i = 1;\ c_i \ge 0,\ \forall i \Big\}.$$

Theorem 4.5: Given uMDP $< T, \gamma, S, A, \mathbf{p}, \mathcal{R} >$, suppose $\mathcal{R} = \mathrm{conv}\{\mathbf{r}_1, \cdots, \mathbf{r}_t\}$ and the initial state-distribution is $\boldsymbol{\alpha}$. Let $\hat{x}_i(s, a) \triangleq \mathbb{E}_{\pi'_i} \sum_{i=1}^{T} \gamma^{i-1} \mathbf{1}(s_i = s, a_i = a)$ where $\pi'_i = \arg\max_{\pi'\in\Pi^D} P(\pi', \mathbf{r_i})$; and h^*, $\mathbf{x}^*$ be the optimal strategy of the following LP,

$$\begin{aligned}
\text{Min: } & h\\
\text{S. T.: } & h \ge \sum_{s\in S} \sum_{a\in A} \big[r_i(s,a)\hat{x}_i(s,a) - r_i(s,a)x(s,a) \big],\ \forall i,\\
& \sum_{a\in A} x(s',a) - \sum_{s\in S}\sum_{a\in A} \gamma p(s'|s,a)x(s,a) = \alpha(s'),\ \forall s',\\
& x(s,a) \ge 0,\ \forall s, \forall a.
\end{aligned}$$

Then the MMR strategy π^* is such that $\pi^*(a|s) = x(s,a)/\sum_{a'\in A} x(s,a')$ for all s, a. Here, the denominator is guaranteed to be nonzero.

Proof: We establish the following lemma first.

Lemma 4.6: For any $\pi \in \Pi^{HR}$ the following holds,

$$R(\pi) = \max_{i=1,\cdots,t} \Big\{ P(\pi'_i, \mathbf{r_i}) - P(\pi, \mathbf{r_i}) \Big\}.$$

Proof: Fix a strategy $\pi \in \Pi^{HR}$. Define the following function ranging over $\mathcal{R}$:

$$R^\pi(\mathbf{r}) \triangleq \max_{\pi'\in\Pi^{HR}} \{ P(\pi', \mathbf{r}) - P(\pi, \mathbf{r}) \}.$$

It is easy to see that $R^\pi(\cdot)$ is convex because $P(\pi', \mathbf{r}) - P(\pi, \mathbf{r})$ is a linear function of $\mathbf{r}$ for any π', and hence $R^\pi(\cdot)$ is convex as it is the maximum of a class of linear functions.

By convexity of $R^\pi(\cdot)$ and definition of π'_i we have

$$\begin{aligned}
R(\pi) &= \max_{\mathbf{r}\in\mathcal{R}} \Big\{ \max_{\pi'\in\Pi^R} \big[P(\pi', \mathbf{r_i}) - P(\pi, \mathbf{r_i}) \big] \Big\} = \max_{\mathbf{r}\in\mathcal{R}} R^\pi(\mathbf{r})\\
&= \max_{i=1,\cdots,t} R^\pi(\mathbf{r_i}) = \max_{i=1,\cdots,t} \Big\{ P(\pi'_i, \mathbf{r_i}) - P(\pi, \mathbf{r_i}) \Big\},
\end{aligned}$$

which establishes the lemma. ■

Now we prove the theorem. By Lemma 4.6, we have

$$R(\pi) = \min_h \Big\{ h | h \ge P(\pi'_i, \mathbf{r}_i) - P(\pi, \mathbf{r}_i),\ i = 1, \cdots, t \Big\}.$$

Taking minimization over $\pi \in \Pi^S$ on both sides, the theorem follows immediately by writing MDP as its dual LP form, see [1] for the details. ■

C. Efficient-strategy approach

Definition 4.7: A strategy $\pi \in \Pi^D$ is called *efficient* if there is no $\pi' \in \Pi^{HR}$ such that $P(\pi, \mathbf{r}) < P(\pi', \mathbf{r})$ holds for all $\mathbf{r} \in \mathcal{R}$.

Theorem 4.8: Suppose $\mathcal{R} = \{\mathbf{r} | A\mathbf{r} \le \mathbf{b}\}$ and $\{\pi'_1, \cdots, \pi'_t\} \subset \Pi^D$ is a superset of the set of efficient strategies. Let $\hat{x}_i(s, a) \triangleq \mathbb{E}_{\pi'_i} \sum_{i=1}^{T} \gamma^{i-1} \mathbf{1}(s_i = s, a_i = a)$, whose vector form is denoted by $\hat{\mathbf{x}}_i$. Let $\mathbf{c}^*$ be the optimal solution of the following LP on h, $\mathbf{c}$ and $\mathbf{z}(i)$,

$$\begin{aligned}
\min:\ \ & h\\
\text{S.T.:}\ \ & \sum_{i=1}^{\top} c_i = 1;\\
& \mathbf{c} \ge \mathbf{0};\\
& \left.\begin{aligned} & h \ge \mathbf{b}^\top \mathbf{z}(i);\\ & A^\top \mathbf{z}(i) + \hat{X}\mathbf{c} = \hat{\mathbf{x}}_i;\\ & \mathbf{z}(i) \ge \mathbf{0}; \end{aligned}\right\} i = 1, \cdots, t,
\end{aligned}$$

where $\hat{X} = (\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_t)$, then the MMR strategy π^* is:

$$\pi^*(a|s) = \frac{\sum_{i=1}^{t} c_i \hat{x}_i(s,a)}{\sum_{a'\in A} \sum_{i=1}^{t} c_i \hat{x}_i(s,a')};\ \forall s, a.$$

Here, the denominator is guaranteed to be nonzero.

Proof: We first show that the MMR strategy is a *randomization* over $\pi'_1, \cdots, \pi'_t$, where "randomization" stands for the following: given a pool of deterministic strategies pick one according to an exogenous stochastic source and then follow it forever. It is well known that (cf [1]) for any stationary strategy, there is an equivalent randomization over all deterministic strategies and vice versa. Hence there is a MMR that is a randomization due to Theorem 2.5. Further note that the probability of picking a non-efficient strategy must be zero, or there exists a strategy that performs strictly better for all $\mathbf{r}$ which contradicts the MMR condition. Hence the MMR strategy is a randomization over $\pi'_1, \cdots, \pi'_t$.

Observe that if the probability of picking π'_i is c_i, then the state-action frequency equals $\sum_{i=1}^{t} c_i \hat{\mathbf{x}}_i$. Thus, the MMR strategy is the following optimization problem:

$$\min_{\mathbf{c}: \sum_{j=1}^{t} c_j = 1; \mathbf{c}\ge\mathbf{0}} \Big\{ \max_{i\in\{1,\cdots,t\}, \mathbf{r}\in\mathcal{R}} \big[\mathbf{r}^\top \hat{x}_i - \mathbf{r}^\top \sum_{j=1}^{t} c_j \hat{\mathbf{x}}_j \big] \Big\}.$$

This can be rewritten as

$$\begin{aligned} \min:\ & h\\ \text{S.T.:}\ & \sum_{i=1}^{\top} c_i = 1;\\ & \mathbf{c} \geq \mathbf{0};\\ & h \geq \max_{\mathbf{r}\in\mathcal{R}} (\hat{\mathbf{x}}_i^\top - \mathbf{c}^\top X^\top)\mathbf{r}, \quad i = 1, \cdots, t. \end{aligned} \tag{9}$$

By duality of LP (cf [15], [16]) and $\mathcal{R} = \{\mathbf{r}|A\mathbf{r} \leq \mathbf{b}\}$, $\max_{\mathbf{r}\in\mathcal{R}}(\hat{\mathbf{x}}_i^\top - \mathbf{c}^\top X^\top)\mathbf{r}$ equals to the following LP on $\mathbf{z}(i)$:

$$\begin{aligned} \text{Min:}\ & \mathbf{b}^\top \mathbf{z}(i);\\ \text{S.T.:}\ & A^\top \mathbf{z}(i) + \hat{X}\mathbf{c} = \hat{\mathbf{x}}_i;\\ & \mathbf{z}(i) \geq \mathbf{0}. \end{aligned}$$

Substituting it into (9) establishes the theorem. ■

Observe that if a strategy maximizes the performance $P(\cdot, \mathbf{r}_0)$ for some parameter realization $\mathbf{r}_0 \in \mathcal{R}$, then it is efficient. The following proposition shows that the reverse also holds. The proof is deferred to Appendix I.

Proposition 4.9: Suppose $\mathcal{R}$ is convex and its relative interior is non-empty[2]. If a strategy $\pi \in \Pi^{HR}$ is efficient, then there exists $\mathbf{r}_0 \in \mathcal{R}$ such that $v^\pi(\mathbf{r}_0) = v^*(\mathbf{r}_0)$.

We may thus use *action elimination* [1] [18] [19] to find a "small" superset of efficient strategies: if an action of a state that can be determined to *not* belong to optimal policy for any parameter realization, it can be discarded and disregarded. If only a small number of strategies remains after action elimination[3], then we can solve MMR in a less computational expensive way.

V. Mean variance tradeoff of regret

So far we regarded the true parameters as deterministic but unknown. In this section we take a Bayesian approach: we treat the true parameters as a random vector following distribution μ known a-priori. Thus, given a strategy, its regret is a random variable whose probability distribution can be evaluated. We use the mean-variance tradeoff criterion to compare such random variables. That is, the strategy that minimizes the tradeoff (i.e., the convex combination) of the mean and variance of the regret is considered optimal.

Definition 5.1: Suppose the true reward parameter $\mathbf{r}^t$ follows a distribution μ supported by a compact $\mathcal{R}$. For a strategy $\pi \in \Pi^{HR}$:

1) the *regret mean* is

$$\begin{aligned} E^R(\pi) &\triangleq \mathbb{E}_{\mathbf{r}^t}\Big\{ \max_{\pi'\in\Pi^{HR}} P(\pi', \mathbf{r}^t) - P(\pi, \mathbf{r}^t)\Big\}\\ &= \int \big[\max_{\pi'\in\Pi^{HR}} P(\pi', \mathbf{r}) - P(\pi, \mathbf{r})\big]\mu(d\mathbf{r}); \end{aligned} \tag{10}$$

2) the *regret variance* is

$$\begin{aligned} \mathrm{Var}^R(\pi) \triangleq & \mathbb{E}_{\mathbf{r}^t}\Big[\max_{\pi'\in\Pi^{HR}} P(\pi', \mathbf{r}^t) - P(\pi, \mathbf{r}^t)\Big]^2\\ & - (E^R(\pi))^2. \end{aligned} \tag{11}$$

Definition 5.2: Suppose the true reward parameter $\mathbf{r}^t$ follows a distribution μ supported by a compact $\mathcal{R}$. Fix $\lambda \in [0, 1]$, the *Optimal Mean-Variance Tradeoff of Regret* (OMVTR) strategy is

$$\pi_\lambda \triangleq \arg\min_{\pi\in\Pi^{HR}} \big[\lambda E^R(\pi) + (1-\lambda)\mathrm{Var}^R(\pi)\big].$$

To simplify notations, define function $P^*(\cdot) : \mathcal{R} \to \mathbb{R}$ as

$$P^*(\mathbf{r}) \triangleq \max_{\pi\in\Pi^{HR}} P(\pi, \mathbf{r}),$$

i.e., the optimal reward-to-go given $\mathbf{r}$. Note that $P^*(\mathbf{r})$ is easy to compute, using for example dynamic programming. Observe that OMVTR strategy is trivial when $\lambda = 1$.

Theorem 5.3: For $\lambda \in [0, 1)$, let $\mathbf{x}_\lambda$ be the optimal solution to the following convex quadratic program

$$\begin{aligned} \text{min:}\ & (1-\lambda)\mathbf{x}^\top \mathbb{E}(\mathbf{r}\mathbf{r}^\top)\mathbf{x}\\ & + \Big\{[(1-\lambda)\mathbb{E}(P^*(\mathbf{r})) - \lambda]\mathbb{E}(\mathbf{r}) - (1-\lambda)\mathbb{E}[P^*(\mathbf{r})\mathbf{r}]\Big\}^\top \mathbf{x}\\ \text{S.T.:}\ & \sum_{a\in A} x(s', a) - \sum_{s\in S}\sum_{a\in A} \gamma p(s'|s, a)x(s, a) = \alpha(s'),\ \forall s'\\ & x(s, a) \geq 0,\ \forall s, a. \end{aligned} \tag{12}$$

The OMVTR strategy π_λ is such that $\pi_\lambda(a|s) = x_\lambda(s, a)/\sum_{a'\in A} x_\lambda(s, a')$ for all s, a. Here, the denominator is guaranteed to be nonzero.

Proof: We again use the equivalence between Π^{HR} and state-action frequency polytope. Let $\mathbf{x}(\pi)$ be the state-action vector of a a strategy π. Observe that

$$\begin{aligned} E^R(\pi) &= \mathbb{E}(P^*(\mathbf{r}^t)) - \mathbb{E}(\mathbf{r}^t)^\top \mathbf{x}(\pi);\\ \mathrm{Var}^R(\pi) &= \mathbb{E}\Big[\max_{\pi'\in\Pi^{HR}} P(\pi', \mathbf{r}^t) - P(\pi, \mathbf{r}^t) - E^R(\pi)\Big]^2\\ &= \mathbb{E}\Big[P^*(\mathbf{r}^t) - {\mathbf{r}^t}^\top \mathbf{x}(\pi) - \mathbb{E}(P^*(\mathbf{r}^t)) + \mathbb{E}(\mathbf{r}^t)^\top \mathbf{x}(\pi)\Big]^2. \end{aligned}$$

Thus algebra yields

$$\begin{aligned} & \lambda E^R(\pi) + (1-\lambda)\mathrm{Var}^R\\ =& \lambda\mathbb{E}(P^*(\mathbf{r})^t) - \lambda\mathbb{E}(\mathbf{r}^t)^\top \mathbf{x}(\pi) - (1-\lambda)[\mathbb{E}(\mathbf{r}^t)^\top \mathbf{x}(\pi)]^2\\ & + (1-\lambda)\mathbb{E}\big\{P^*(\mathbf{r}^t)^2 - 2P^*(\mathbf{r}^t){\mathbf{r}^t}^\top \mathbf{x}(\pi) + ({\mathbf{r}^t}^\top \mathbf{x}(\pi))^2\big\}\\ & - (1-\lambda)[\mathbb{E}(P^*(\mathbf{r}^t))]^2 + 2(1-\lambda)\mathbb{E}[P^*(\mathbf{r}^t)]\mathbb{E}(\mathbf{r}^t)^\top \mathbf{x}(\pi). \end{aligned}$$

The r.h.s. is equivalent to the minimizing objective in (12) up to adding a constant, which establishes the theorem. ■

We can use Monte Carlo methods to solve Problem (12). Denote the objective function by $O(\mathbf{x})$. Note that it is a convex quadratic program. Further note that all its coefficients are expectations of random variables. Thus we can generate independent samples $\mathbf{r}(1), \cdots, \mathbf{r}(n)$ according to μ, and use the corresponding empirical average to approximate each coefficient. The following theorem establishes an error bound of the solution to the approximated problem $\overline{O}(\mathbf{x})$.

Theorem 5.4: Let π^* and $\overline{\pi}$ be the OMVTR and the solution to the approximated problem using n i.i.d. samples respectively. Denote $\hat{T} \triangleq \sum_{i=1}^{T} \gamma^{i-1}$; $V \triangleq |S| \times |A|$

[2] See page 23 of [17] for for the definition of relative interior. In particular, all polytopes have non-empty relative interior.

[3] Of course this is not guaranteed due to the NP-hardness of MMR.

and $\hat{R} \triangleq \sup_{\mathbf{r}\in\mathcal{R}} \max_{s\in S, a\in A} |r(s,a)|$. Then, the following holds:

$$\begin{aligned}&\Pr\big\{\lambda E^R(\overline{\pi}) + (1-\lambda)\mathrm{Var}^R(\overline{\pi})\\ &\qquad \geq \lambda E^R(\pi^*) + (1-\lambda)\mathrm{Var}^R(\pi^*) + 2\epsilon\big\}\\ \leq &(2V^2+4V+2)\exp\left(\frac{-n\epsilon^2}{2\hat{R}^2(4\hat{T}^2\hat{R}+\hat{T})^2}\right).\end{aligned}$$

Proof: We use overline to represent the empirical average of a quantity from n iid sampling. Note that each element of the $V\times V$ random matrix $\mathbf{r}(i)\mathbf{r}(i)^\top$ belongs to $[-\hat{R}^2, \hat{R}^2]$; $P^*(\mathbf{r}(i)) \in [-\hat{T}\hat{R}, \hat{T}\hat{R}]$; each element of the V dimension random vector $\mathbf{r}(i)$ belongs to $[-\hat{R}, \hat{R}]$; each element of the V dimension random vector $P^*(\mathbf{r}(i))\mathbf{r}(i)$ belongs to $[-\hat{T}\hat{R}^2, \hat{T}\hat{R}^2]$. Let $\epsilon_0 = \epsilon/(4\hat{T}^2\hat{R}+\hat{T})$. By Hoeffding's inequality, the followings hold:

$$\Pr\Big(\big\|\overline{\mathbf{r}\mathbf{r}^\top} - \mathbb{E}(\mathbf{r}\mathbf{r}^\top)\big\|_{\max} \geq R\epsilon_0\Big) \leq 2V^2\exp\left(-\frac{n\epsilon_0^2}{2\hat{R}^2}\right). \tag{13}$$

$$\Pr\Big(\big|\overline{P^*(\mathbf{r})} - \mathbb{E}(P^*(\mathbf{r}))\big| \geq \hat{T}\epsilon_0\Big) \leq 2\exp\left(-\frac{n\epsilon_0^2}{2\hat{R}^2}\right). \tag{14}$$

$$\Pr\Big(\big\|\overline{\mathbf{r}} - \mathbb{E}(\mathbf{r})\big\|_\infty \geq \epsilon_0\Big) \leq 2V\exp\left(-\frac{n\epsilon_0^2}{2\hat{R}^2}\right). \tag{15}$$

$$\Pr\Big(\big\|\overline{P^*(\mathbf{r})\mathbf{r}} - \mathbb{E}(P^*(\mathbf{r})\mathbf{r})\big\|_\infty \geq \hat{T}\hat{R}\epsilon_0\Big) \leq 2V\exp\left(-\frac{n\epsilon_0^2}{2\hat{R}^2}\right). \tag{16}$$

Here $\|\|_{\max}$ stands for the largest absolute value of elements of a matrix. Now note that $\mathbf{x}\in\mathcal{X}$ implies $\|\mathbf{x}\|_\infty \leq \hat{T}$. Algebraic manipulations easily yield that for $\mathbf{x}\in\mathcal{X}$:

$$\begin{aligned}&\overline{O}(\mathbf{x}) - O(\mathbf{x})\\ \leq&(1-\lambda)\hat{T}^2\left\|\overline{\mathbf{r}\mathbf{r}^\top} - \mathbb{E}(\mathbf{r}\mathbf{r}^\top)\right\|_{\max}\\ &+(1-\lambda)\hat{T}\left\|\overline{P^*(\mathbf{r})\mathbf{r}} - \mathbb{E}(P^*(\mathbf{r})\mathbf{r})\right\|_\infty\\ &+(1-\lambda)\hat{T}\hat{R}|\overline{P^*(\mathbf{r})} - \mathbb{E}(P^*(\mathbf{r}))|\\ &+\hat{T}[(1-\lambda)\hat{T}\hat{R}+\lambda]\|\overline{\mathbf{r}} - \mathbb{E}(\mathbf{r})\|_\infty.\end{aligned}$$

Combining this with Inequalities (13) to (16), we have:

$$\begin{aligned}&\Pr\Big\{\max_{\mathbf{x}\in\mathcal{X}}|\overline{O}(\mathbf{x}) - O(\mathbf{x})| \geq \epsilon\Big\}\\ \leq&(2V^2+4V+2)\exp\left(\frac{-n\epsilon^2}{2\hat{R}^2(4\hat{T}^2\hat{R}+\hat{T})^2}\right),\end{aligned}$$

which implies the theorem. ■

VI. CONCLUSION

In this paper we investigated decision making in a Markovian setup where the reward parameters are not known in advance. In contrast to the standard setup where a strategy is evaluated by its accumulated reward-to-go, we focused on the so-called competitive setup where the criterion is the parametric regret, i.e., the gap between the performance of the best strategy that is chosen after the true parameter realization is revealed and the performance of the strategy that is chosen before the parameter realization is revealed.

We considered two related formulations: minimax regret and mean-variance tradeoff of the regret. In the minimax regret formulation, the true parameters are regarded as deterministic but unknown, and the optimal strategy is the one that minimizes the worst-case regret under the most adversarial possible realization. We showed that the problem of computing the minimax regret strategy is NP-hard and proposed algorithms to efficiently solve it under favorable conditions. The mean-variance tradeoff formulation requires a probabilistic model of the uncertain parameters and looks for a strategy that minimizes a convex combination of the mean and the variance of the regret. We proved that computing such a strategy can be done numerically in an efficient way.

MDPs in a competitive setup can model many real applications. However, unlike the standard setup, robust decision making in such a setup has not been thoroughly investigated. This paper aims to address this absence by recasting solution concepts that were successfully implemented for standard setup to the competitive setup and solve them efficiently.

APPENDIX I
PROOF OF PROPOSITION 4.9:

We define the following to simplify the expression:

$$\begin{aligned}&v^\pi(\mathbf{r}) \triangleq P(\pi,\mathbf{r}); \quad \pi\in\Pi^{HR}.\\ &v^*(\mathbf{r}) \triangleq \max_{\pi\in\Pi^{HR}} v^\pi(\mathbf{r}).\end{aligned}$$

Before proving the proposition, we establish the following lemma.

Lemma 1.1: Let $\mathcal{R}$ be convex, then

1) for any $\pi\in\Pi^S$, $v^\pi(\cdot)$ is an affine function;
2) $v^*(\cdot)$ is a convex, piecewise affine function.

Proof: Note that given strategy $\pi\in\Pi^S$, we have

$$v^\pi(\mathbf{r}) = \sum_{s\in S}\sum_{a\in A}\{r(s,a)\mathbb{E}_\pi\sum_i \gamma^{i-1}\mathbf{1}(s_i = s, a_i = a)\}.$$

The right-hand side is affine of $\mathbf{r}$, which implies the first claim.

To prove the second claim, recall that (e.g., [1]) for a fixed $\mathbf{r}$, the optimal strategy is determined and stationary, i.e.,

$$v^*(\mathbf{r}) = \max_{\pi\in\Pi^{HR}} P(\pi,\mathbf{r}) = \max_{\pi\in\Pi^D} P(\pi,\mathbf{r}).$$

Further note that Π^D is a finite set, and $v^\pi(\mathbf{r})$ is affine. Thus $v^*(\cdot)$ is convex and piecewise affine, since it is a pointwise maximum over a finite number of affine functions. ■

We now prove the proposition by contradiction. Assume there exists an efficient strategy π^* which does not maximize the expect reward for any realization. Note $v^{\pi^*}(\cdot)$ is affine. We construct a function $v'(\cdot)$ such that $v'(\mathbf{r}) > v^{\pi^*}(\mathbf{r})$ for all $\mathbf{r}\in\mathcal{R}$, and show that there exists a strategy $\pi'\in\Pi^{HR}$ such that $v^{\pi'}(\mathbf{r}) \geq v'(\mathbf{r})$ for all $\mathbf{r}\in\mathcal{R}$.

Step 1: To construct $v'(\cdot)$, note that by assumption $v^{\pi^*}(\mathbf{r}) < v^*(\mathbf{r})$ for all $\mathbf{r}\in\mathcal{R}$. Hence let $c_0 \triangleq \min_{\mathbf{r}\in\mathcal{R}}\left[v^*(\mathbf{r}) - v^{\pi^*}(\mathbf{r})\right]$ and $\mathbf{r}_0 \in \arg\min_{\mathbf{r}\in\mathcal{R}}\left[v^*(\mathbf{r}) - v^{\pi^*}(\mathbf{r})\right]$. These two definition is valid since $v^*(\cdot)$ and $v^{\pi^*}(\cdot)$ are continuous functions and $\mathcal{R}$ is compact. Let $v'(\mathbf{r}) \triangleq$

$v^{\pi^*}(\mathbf{r})+c_0$, observe that $v^{\pi^*}(\mathbf{r}) < v'(\mathbf{r}) \leq v^*(\mathbf{r})$ holds for all $\mathbf{r} \in \mathcal{R}$, and we also have $v'(\mathbf{r}_0) = v^*(\mathbf{r}_0)$. Note that, $v^{\pi^*}(\mathbf{r})$ is an affine function, so is $v'(\mathbf{r})$ by definition, and we can rewrite

$$v'(\mathbf{r}) = \mathbf{g}^\top \mathbf{r} + \big[v^*(\mathbf{r}_0) - \mathbf{g}^\top \mathbf{r}_0\big].$$

Step 2: To show there exists $\pi' \in \Pi^{HR}$ such that $v^{\pi'}(\mathbf{r}) \geq v'(\mathbf{r})$ for all $\mathbf{r} \in \mathcal{R}$. Let $\mathcal{R} \subseteq \mathbb{R}^m$ and we extend $v^*(\cdot)$ into the whole space, i.e., for $\mathbf{r} \in \mathbb{R}^m$, define

$$v_f^*(\mathbf{r}) \triangleq \max_{\pi \in \Pi^D} P(\pi, \mathbf{r});$$
$$v_o^*(\mathbf{r}) \triangleq \begin{cases} 0 & \text{if } \mathbf{r} \in \mathcal{R}; \\ +\infty & \text{otherwise.} \end{cases}$$

Note that $v'(\mathbf{r}) \leq v^*(\mathbf{r})$ holds for all $\mathbf{r} \in \mathcal{R}$ implies $\mathbf{g}^\top \mathbf{r} + \big[v^*(\mathbf{r}_0) - \mathbf{g}^\top \mathbf{r}_0\big] \leq v_f^*(\mathbf{r}) + v_o^*(\mathbf{r})$ holds for all $\mathbf{r} \in \mathbb{R}^m$. Hence $\mathbf{g}$ is a subgradient to convex function $v_f^*(\mathbf{r}) + v_o^*(\mathbf{r})$ at $\mathbf{r}_0$, denote as $\mathbf{g} \in \partial\big[v_f^*(\mathbf{r}_0) + v_o^*(\mathbf{r}_0)\big]$. Hence there exists $\mathbf{g}_f$, $\mathbf{g}_o$ such that $\mathbf{g}_f \in \partial v_f^*(\mathbf{r}_0)$, $\mathbf{g}_o \in \partial v_o^*(\mathbf{r}_0)$ and $\mathbf{g} = \mathbf{g}_f + \mathbf{g}_o$ (cf Theorem 23.8 of [14]).

Step 2.1 To prove there exists π' such that $v^{\pi'}(\mathbf{r}) = \mathbf{g}_f^\top \mathbf{r} + \big[v^*(\mathbf{r}_0) - \mathbf{g}_f^\top \mathbf{r}_0\big]$ for all $\mathbf{r} \in \mathcal{R}$. Let set $\Pi_0 \triangleq \arg\max_{\pi \in \Pi^D} v^\pi(\mathbf{r}_0)$, i.e., the set of strategies that achieves maximal at $\mathbf{r}_0$. Note that Π_0 is a finite set since Π^D is a finite set. Hence denote $\Pi_0 = \{\pi_1, \cdots, \pi_h\}$. Note that by definition of Π_0, $v^{\pi_i}(\mathbf{r}_0) = v^*(\mathbf{r}_0)$. Hence we can rewrite

$$v^{\pi_i}(\mathbf{r}) = \mathbf{d}_i^\top \mathbf{r} + [v^*(\mathbf{r}_0) - \mathbf{d}_i^\top \mathbf{r}_0],$$

for some $\mathbf{d}_i$ since $v^{\pi_i}(\cdot)$ is a linear function.

Recall $\mathbf{g}_f \in \partial v_f^*(\mathbf{r}_0)$, hence by a standard continuity argument we have in a sufficiently small open ball around $\mathbf{r}_0$, $\mathbf{g}_f^\top \mathbf{r} + \big[v^*(\mathbf{r}_0) - \mathbf{g}_f^\top \mathbf{r}_0\big] \leq \max_{\pi \in \Pi_0} v^{\pi_i}(\mathbf{r})$. Note that the left-hand side is affine, and the right-hand side is piecewise affine, hence this inequality holds for all $\mathbf{r} \in \mathbb{R}^m$. That is

$$\mathbf{g}_f^\top(\mathbf{r} - \mathbf{r}_0) \leq \max_{i \in \{1, \cdots, h\}} \mathbf{d}_i^\top(\mathbf{r} - \mathbf{r}_0), \quad \forall \mathbf{r} \in \mathbb{R}^m.$$

This implies there exists no $\mathbf{y} \in \mathbb{R}^{m+1}$ such that $[\mathbf{g}_f^\top, 1]\mathbf{y} \geq \max_{i \in \{1, \cdots, h\}} [\mathbf{d}_i^\top, 1]\mathbf{y}$, hence no $\mathbf{y}$ satisfy the following conditions

$$\begin{bmatrix} \mathbf{g}_f \\ 1 \end{bmatrix}^\top \mathbf{y} > 0; \quad \begin{bmatrix} \mathbf{d}_i \\ 1 \end{bmatrix}^\top \mathbf{y} \leq 0; \; i = 1, \cdots, h.$$

By Farkas Lemma, this means there exists $\lambda_1, \cdots, \lambda_h$ such that $\lambda_i \geq 0$ and

$$\begin{bmatrix} \mathbf{g}_f \\ 1 \end{bmatrix} = \sum_{i=1}^h \lambda_i \begin{bmatrix} \mathbf{d}_i \\ 1 \end{bmatrix}.$$

This implies $\sum_{i=1}^h \lambda_i = 1$ and $\sum_{i=1}^h \lambda_i \mathbf{d}_i = \mathbf{g}_f$. Now construct a strategy π' as taking strategy π_i with probability λ_i, and we have

$$\begin{aligned} v^{\pi'}(\mathbf{r}) &= \sum_{i=1}^h \lambda_i v^{\pi_i}(\mathbf{r}) \\ =& \sum_{i=1}^h \lambda_i \Big\{\mathbf{d}_i^\top \mathbf{r} + [v^*(\mathbf{r}_0) - \mathbf{d}_i^\top \mathbf{r}_0]\Big\} \\ =& \mathbf{g}_f^\top \mathbf{r} + \big[v^*(\mathbf{r}_0) - \mathbf{g}_f^\top \mathbf{r}_0\big]. \end{aligned}$$

Step 2.2: To show that $v^{\pi'}(\mathbf{r}) \geq v'(\mathbf{r})$ for all $\mathbf{r} \in \mathcal{R}$. By definition of $v_o^*(\cdot)$ and $\mathbf{g}_o \in \partial v_o^*(\mathbf{r}_0)$ we have

$$\mathbf{g}_o^\top \mathbf{r} + [v_o^*(\mathbf{r}_0) - \mathbf{g}_o^\top \mathbf{r}_0] \leq 0, \quad \forall \mathbf{r} \in \mathcal{R}.$$

Recall $\mathbf{r}_0 \in \mathcal{R}$, which implies $v_o^*(\mathbf{r}_0) = 0$. Hence substitute this into $\mathbf{g} = \mathbf{g}_f + \mathbf{g}_o$ leads to

$$\begin{aligned} & v'(\mathbf{r}) = \mathbf{g}^\top \mathbf{r} + [v^*(\mathbf{r}_0) - \mathbf{g}^\top \mathbf{r}_0] \\ =& \mathbf{g}_o^\top \mathbf{r} + [v_o^*(\mathbf{r}_0) - \mathbf{g}_o^\top \mathbf{r}_0] + \mathbf{g}_f^\top \mathbf{r} + [v_f^*(\mathbf{r}_0) - \mathbf{g}_f^\top \mathbf{r}_0] \\ \leq& \mathbf{g}_f^\top \mathbf{r} + [v_f^*(\mathbf{r}_0) - \mathbf{g}_f^\top \mathbf{r}_0] \\ =& v^{\pi'}(\mathbf{r}). \quad \forall \mathbf{r} \in \mathcal{R}. \end{aligned}$$

Hence we proved Step 2. Combining two steps, we establish the proposition.

References

[1] M. L. Puterman, *Markov Decision Processes*. John Wiley & Sons, INC, 1994.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[3] J. Hannan, “Approximation to Bayes risk in repeated play,” *Contributions to the Theory of Games*, vol. 3, pp. 97–139, 1957.

[4] A. Nilim and L. El Ghaoui, “Robust control of Markov decision processes with uncertain transition matrices,” *Operations Research*, vol. 53, no. 5, pp. 780–798, September 2005.

[5] A. Bagnell, A. Ng, and J. Schneider, “Solving uncertain Markov decision processes,” Carnegie Mellon University, Tech. Rep. CMU-RI-TR-01-25, August 2001.

[6] C. White III and H. K. El-Deib, “Parameter imprecision in finite state, finite action dynamic programs,” *Operations Research*, vol. 34, no. 1, pp. 120–128, January 1986.

[7] G. Iyengar, “Robust dynamic progamming,” *Mathematics of Operations Research*, vol. 30, no. 2, pp. 257–280, 2005.

[8] E. Delage and S. Mannor, “Percentile optimization for Markov decision processes with parameter uncertainty,” To appear in *Operations Research*, 2009.

[9] H. Levy and H. Markowtiz, “Approximating expected utility by a function of mean and variance,” *American Economic Review*, vol. 69, no. 3, pp. 308–17, 1979.

[10] S. Singh, T. Jaakkola, and M. I. Jordan, “Reinforcement learning with soft state aggregation,” in *Advances in Neural Information Processing Systems*, G. Tesauro, D. Touretzky, and T. Leen, Eds., vol. 7. The MIT Press, 1995, pp. 361–368.

[11] A. L. Soyster, “Convex programming with set-inclusive constraints and applications to inexact linear programming,” *Operations Research*, vol. 21, pp. 1154–1157, 1973.

[12] A. Ben-Tal and A. Nemirovski, “Robust solutions of uncertain linear programs,” *Operations Research Letters*, vol. 25, no. 1, pp. 1–13, August 1999.

[13] C. Papadimitriou, *Computational Complexity*. Addison Wesley, 1994.

[14] R. Rockafellar, *Convex Analysis*. Princeton University Press, Princeton, N.J., 1970.

[15] K. G. Murty, *Linear Programming*. John Wiley & Sons, 1983.

[16] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*. Athena Scientific, 1997.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[18] J. MacQueen, “A modified dynamic programming method for Markov decision problems,” *Journal of Mathematical Analysis and Application*, vol. 14, pp. 38–43, 1966.

[19] E. Even-Dar, S. Mannor, and Y. Mansour, “Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems,” *Journal of Machine Learning Research*, vol. 7, pp. 1079–1105, 2006.