Parametric Regret in Uncertain Markov Decision Processes

Huan Xu, and Shie Mannor

We consider Markov Decision Processes (MDPs) where the reward parameters are not known in advance, and 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. This setup is motivated by applications such as portfolio optimizations, in which the objective is often to "beat the market", i.e., to perform favorably than a strategy that holds index stocks. A natural measurement of strategies in such setup is the so-called *parametric regret*: the gap between the performance of a strategy and that of the optimal one.

We investigate two related formulations: the first one takes a minimax approach, i.e., the true parameters can be any element of a known set, and strategies are evaluated based on the parametric regret under the worst possible parameter realization. The second one regard the true parameters as random variables. Thus, given a strategy, its parametric regret is a random variable whose probability distribution can be obtained. And the optimal strategy is the one that maximizes its mean-variance tradeoff. See below for the details.

Although the parametric regret formulations are motivated by MDPs with a comparative objective, it can also be used for the standard *planning problem*, i.e., the parameters are uncertainty and the decision goal is the accumulated reward-to-go. In such a case, the parametric regret formulations provide decision criteria that handle uncertainty in a less conservative way compared to the robust (i.e., worst-case analysis) approach. It would be interesting to investigate how these formulations can be incorporated in the *learning problem* where the decision agent can actively obtain better estimation of the parameters.

Notations: 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; S is the state set and A is the action set; $\mathbf{p}$ is the transition probability; and $\mathcal{R}$ is the admissible set of reward parameter, which is termed *uncertainty set*. We focus on the case when the uncertainty set $\mathcal{R}$ is a polytope. All history-dependent randomized strategies are admissible, and we denote that set by Π^{HR}. Given a strategy $\pi \in \Pi^{HR}$ and a parameter realization $\mathbf{r} \in \mathcal{R}$, its expected performance is denoted by $P(\pi, \mathbf{r})$, that is $P(\pi, \mathbf{r}) \triangleq \mathbb{E}_\pi\{\sum_{i=1}^{T} \gamma^{i-1} r(s_i, a_i)\}$.

A. Minimax regret

First, we consider the minimax approach.

Definition 1:

(I) The *parametric regret* of a strategy π w.r.t. $\mathbf{r}_0 \in \mathcal{R}$ is defined as

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

(II) The *Maximum Regret* of a strategy π is defined as

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

(III) The *MiniMax Regret* (MMR) strategy is defined as

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

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 2: 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.

We propose first a subgradient method to find the MMR solution. The subgradient for each step is indeed the reward parameter that achieves the maximum regret.

Algorithm 3:

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 $\mathbf{r}^* := \arg\max_{\mathbf{r} \in \mathcal{R}} \left\{ \max_{\mathbf{x}' \in \mathcal{X}} (\mathbf{r}^\top \mathbf{x}' - \mathbf{r}^\top \mathbf{x}^*) \right\}$.
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.

Here $\mathcal{X}$ is the state-action frequency polytope. This algorithm is based on the fact that minimizing the maximum regret is indeed a convex program. Furthermore, based on a "large M" method, we propose a MIP formulation that finds the subgradient $\arg\max_{\mathbf{r}\in\mathcal{R}}\left\{\max_{\mathbf{x}'\in\mathcal{X}}(\mathbf{r}^\top\mathbf{x}'-\mathbf{r}^\top\mathbf{x}^*)\right\}$. This method is non-polynomial, due to the inherent NP-hardness of the problem.

Two special types of uMDP in which the MMR strategy is polynomial time solvable are then investigated. The first one considers the case where the uncertainty set $\mathcal{R}$ has a small number of vertices. The second one considers the case where the set of "efficient" strategies, i.e., deterministic strategies that are optimal to at least one $\mathbf{r}\in\mathcal{R}$, contains a small number of elements. In either case, MMR can be obtained by solving a linear program.

B. Mean variance tradeoff of regret

So far the true parameters as deterministic but unknown. Next 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 4:

(I) 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

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

2) the *regret variance* is

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

(II) 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, let $P^*(\mathbf{r})\triangleq\max_{\pi\in\Pi^{HR}}P(\pi,\mathbf{r})$, i.e., the optimal reward-to-go given $\mathbf{r}$. Note that $P^*(\mathbf{r})$ is easy to compute. We have the following theorem that solves the OMVTR strategy.

Theorem 5: For $\lambda\in[0,\,1)$, let $\mathbf{x}_\lambda$ be the optimal solution to the following convex quadratic program

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

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.

The coefficients of Problem (1) are not readily known. However, we can use Monte Carlo methods to solve Problem (1). 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 6: Let π^* and $\overline{\pi}$ be the OMVTR and the solution to the approximated problem using n i.i.d. samples respectively. Denote $\hat{T}\triangleq\sum_{i=1}^{T}\gamma^{i-1}$; $V\triangleq|S|\times|A|$ and $\hat{R}\triangleq\sup_{\mathbf{r}\in\mathcal{R}}\max_{s\in S,a\in A}|r(s,a)|$. Then, the following holds:

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