# Online Convex Optimization in Adversarial Markov Decision Processes

**Aviv Rosenberg** [1] **Yishay Mansour** [1 2]

## Abstract

We consider online learning in episodic loop-free Markov decision processes (MDPs), where the loss function can change arbitrarily between episodes, and the transition function is not known to the learner. We show $\tilde{O}(L|X|\sqrt{|A|T})$ regret bound, where $T$ is the number of episodes, $X$ is the state space, $A$ is the action space, and $L$ is the length of each episode. Our online algorithm is implemented using entropic regularization methodology, which allows to extend the original adversarial MDP model to handle convex performance criteria (different ways to aggregate the losses of a single episode) , as well as improve previous regret bounds.

## 1. Introduction

Markov Decision processes (Puterman, 1994) have been widely used to model reinforcement learning problems - problems involving sequential decision making in a stochastic environment. In this model both the losses and dynamics of the environment are assumed to be stationary over time. However, in real world applications, the losses might change over time, even throughout the learning process.

The adversarial MDP model (Even-Dar et al., 2009) was proposed to address these issues. In this model, the loss function can change arbitrarily (while still assuming a fixed stochastic transition function). The learner's objective is to minimize its average loss during the learning process, and its performance is measured by the regret - comparing to the best stationary policy in hindsight. These ideas originate from online learning problems (Cesa-Bianchi & Lugosi, 2006) - where, in each round, the learner selects an action before knowing the current loss function.

BGP routing is considered as a motivating example in the full version of the paper.

[1]Tel Aviv University, Israel [2]Google Research, Tel Aviv, Israel. Correspondence to: Aviv Rosenberg <avivros007@gmail.com>, Yishay Mansour <mansour.yishay@gmail.com>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

We propose a novel algorithm for the adversarial MDP model where the transition function is unknown to the learner and the losses change arbitrarily over time. Our algorithm, UC-O-REPS, uses two important ingredients, the first is Online Mirror Descent (OMD) (Shalev-Shwartz, 2012) and the second is UCRL-2 (Auer et al., 2008). A major challenge in this work is to handle convex performance criteria, which model different ways of aggregating the losses of each episode. In order to handle convex performance criteria, we use the methodology of OMD, which is widely used for online convex optimization, and we implement it in the adversarial MDP setting. In order to overcome the unknown dynamics (stochastic transition function) we incorporate techniques from UCRL-2.

Our main contribution is extending the adversarial MDP model to include convex performance criteria, and showing that our algorithm, UC-O-REPS, achieves near-optimal regret bounds in the general model. This is an important extension since different applications have different optimization criteria, other than minimizing the expected average loss. Examples include risk-sensitive objectives and robust objectives (that combine multiple loss functions). In addition, we improve the known regret bound of Neu et al. (2012) for the expected average loss from $\tilde{O}(L|X||A|\sqrt{T})$ to achieve $\tilde{O}(L|X|\sqrt{|A|T})$, which is especially important for large action spaces. Our bounds also hold with high probability, and not only in expectation. Our algorithm builds on a simple entropic regularization method, and the main challenge is the analysis of the regret and computational complexity.

### 1.1. Related Work

The works of Auer et al. (2008) and Bartlett & Tewari (2009) assume an unknown fixed MDP, and achieve a $\tilde{O}(L|X|\sqrt{|A|T})$ regret compared to the optimal policy. A recent work by Azar et al. (2017) achieves $\tilde{O}(\sqrt{L|X||A|T})$ regret for large enough $T$, which is optimal (Auer et al., 2008). We remark that the lower bound of $\Omega(\sqrt{L|X||A|T})$ by Auer et al. (2008) shows that our regret bound is optimal with respect to the number of time steps $T$ and actions $|A|$.

The work of Even-Dar et al. (2009), which presented the adversarial MDP model, assumes full knowledge of the transition function and full information feedback about the losses. They propose an algorithm, MDP-E, which uses an experts

algorithm in each state and achieves $O(\tau^2\sqrt{T\ln|A|})$ regret, where $\tau$ is a bound on the mixing time of the MDP. Another early work in this setting, by Yu et al. (2009), achieves an $O(T^{2/3})$ regret.

In the bandit setting, the learner observes only the losses related to its actions, i.e., a bandit feedback. The work of Neu et al. (2010) achieves an $O(L^2\sqrt{T|A|}/\alpha)$ regret, where $\alpha > 0$ is a lower bound on the steady state probability to reach some state $x$ under some policy $\pi$. Later Neu et al. (2014) eliminate the dependence on $\alpha$ but achieve only $\tilde{O}(T^{2/3})$ regret. A later work, by Zimin & Neu (2013), proposed the O-REPS algorithm which guarantees an $\tilde{O}(\sqrt{L|X||A|T})$ regret.

The only work that considers the setting of unknown transition function in an adversarial MDP is Neu et al. (2012). They propose an algorithm, Follow the Perturbed Optimistic Policy (FPOP), which builds on Follow the Perturbed Leader (Kalai & Vempala, 2003), and achieves $\tilde{O}(L|X||A|\sqrt{T})$ regret.

The rest of the paper is organized as follows. Section 2 presents the formal model and problem. Section 3 presents the concept of occupancy measures, which will enable us to reformulate the problem as an instance of online convex optimization. Section 4 describes our algorithm and its efficient implementation. Section 5 proves our algorithm's regret bound.

## 2. Problem Formulation

An episodic loop-free adversarial MDP is defined by a tuple $M = \left(X, A, P, \{\ell_t\}_{t=1}^T\right)$, where $X$ and $A$ are the finite state and action spaces, and $P : X \times A \times X \to [0,1]$ is the transition function such that $P(x'|x,a)$ is the probability to move to state $x'$ when performing action $a$ in state $x$.

We assume that the state space can be decomposed into $L$ non-intersecting layers $X_0, \ldots, X_L$ such that the first and the last layers are singletons, i.e., $X_0 = \{x_0\}$ and $X_L = \{x_L\}$. Furthermore, the loop-free assumption means that transitions are only possible between consecutive layers. These assumptions are not necessary, but they simplify some arguments and have a nice interpretation as a game with $L$ steps played for $T$ times.

Let $\{\ell_t\}_{t=1}^T$ be a sequence of loss functions describing the losses at each episode, i.e., $\ell_t : X \times A \times X \to [0,1]^d$. We do not make any statistical assumption on the loss functions, i.e., they can be chosen arbitrarily. Notice that the losses might be multidimensional which can be useful for modeling multiple losses at the same time. Moreover, the learner does not suffer the losses directly, instead they are aggregated using some performance criterion (defined later).

The interaction between the learner and the environment is described in Algorithm 1. It proceeds in episodes, where in each episode the learner starts in state $x_0$ and moves forward across the consecutive layers until it reaches state $x_L$. The learner's task is to select an action at each state it visits. Alternatively, we can say that its task at each episode is to choose a stationary (stochastic) policy ,which is a mapping $\pi : X \times A \to [0,1]$, where $\pi(a|x)$ gives the probability that action $a$ is selected in state $x$.

We denote by $U$ a trajectory through the consecutive layers from $x_0$ to $x_L$, and by $\ell(U)$ the sequence of losses obtained in this trajectory (with respect to loss function $\ell$), i.e.,

$$U = (x_0, a_0, x_1, a_1, \ldots, x_{L-1}, a_{L-1}, x_L)$$
$$\ell(U) = \left\{\ell(x_k, a_k, x_{k+1})\right\}_{k=0}^{L-1}$$

Moreover, we use the notation $\mathbb{E}\left[\ell(U)|P,\pi\right]$ for the expectation of the losses obtained over trajectories that are generated using transition function $P$ and policy $\pi$. That is, action $a_k$ is chosen using $\pi(\cdot|x_k)$ and state $x_{k+1}$ is drawn from distribution $P(\cdot|x_k,a_k)$.

The goal of the learner is to minimize its total loss with respect to some performance criterion $\mathcal{C}$, i.e.,

$$\hat{L}_{1:T}^{\mathcal{C}}(\{\ell_t\}_{t=1}^T) = \sum_{t=1}^{T} \mathcal{C}\left(\mathbb{E}\left[\ell_t(U)|P,\pi_t\right]\right)$$

where $\pi_t$ is the policy chosen by the learner in episode $t$, and $\mathcal{C} : (\mathbb{R}^d)^L \to \mathbb{R}_{\geq 0}$ is the performance criterion, that aggregates the losses of each episode.

**Algorithm 1** Learner-Environment Interaction

**Parameters:** MDP $M = \left(X, A, P, \{\ell_t\}_{t=1}^T\right)$ and performance criterion $\mathcal{C}$
**for** $t = 1$ **to** $T$ **do**
  learner starts in state $x_0^{(t)} = x_0$
  **for** $k = 0$ **to** $L-1$ **do**
    learner chooses action $a_k^{(t)} \in A$
    environment draws new state $x_{k+1}^{(t)} \sim P(\cdot|x_k^{(t)}, a_k^{(t)})$
    learner observes state $x_{k+1}^{(t)}$
  **end for**
  loss function $\ell_t$ is exposed to learner
**end for**

Here are a few interesting and important examples for performance criteria, that our algorithm is able to handle.

**Example 2.1.** *The simplest and most useful example is the total expected loss (TEL) performance criterion, which (to the best of our knowledge) has been the only performance criterion studied so far. Losses are* 1*-dimension, i.e.,* $d = 1$*, and the criterion is defined as follows,*

$$\mathcal{C}^{TEL}\left(\{v_k\}_{k=0}^{L-1}\right) = \sum_{k=0}^{L-1} v_k \qquad (v_k \in \mathbb{R})$$

**Example 2.2.** *We can use the performance criterion to minimize the worst case loss when there are multiple loss functions. Here each dimension of the losses is considered as an individual loss function, and the learner's objective is a min-max criterion, i.e.,*

$$\mathcal{C}^{MM}\left(\{v_k\}_{k=0}^{L-1}\right) = \max_{1\le i\le d}\sum_{k=0}^{L-1} v_k[i] \qquad (v_k \in \mathbb{R}^d)$$

**Example 2.3.** *We can use the performance criterion for a notion of risk-sensitivity. Here losses are 1-dimension and we want to minimize a trade-off between the loss and the risk. Specifically, given a trade-off parameter $0 \le \alpha \le 1$ and a risk parameter $c > 1$, the performance criterion is*

$$\mathcal{C}^{RISK}_{\alpha,c}\left(\{v_k\}_{k=0}^{L-1}\right) = \alpha\left(\sum_{k=0}^{L-1} v_k\right)^c + (1-\alpha)\sum_{k=0}^{L-1}(v_k)^c$$

The performance of the learner will be measured by comparison to the best stationary policy with respect to the chosen performance criterion. For a policy $\pi$ we define its total loss with respect to some performance criterion $\mathcal{C}$ as

$$L^{\mathcal{C}}_{1:T}(\pi;\{\ell_t\}_{t=1}^T) = \sum_{t=1}^T \mathcal{C}\left(\mathbb{E}\left[\ell_t(U)|P,\pi\right]\right)$$

Thus the learner's regret is defined as follows,

$$\hat{R}^{\mathcal{C}}_{1:T} = \hat{L}^{\mathcal{C}}_{1:T}(\{\ell_t\}_{t=1}^T) - \min_{\pi} L^{\mathcal{C}}_{1:T}(\pi;\{\ell_t\}_{t=1}^T)$$

where the minimum is taken over all stationary stochastic policies.

**Remark 2.1.** *Note that if the dynamics were known to the learner, it would not need to observe the trajectory $U_t$ at each episode $t$, since it could compute its performance criterion using $\ell_t$, $\pi_t$ and $P$. In this case, we actually reduce the problem to online learning in the space of the policies. When the dynamics are unknown, the learner uses the observed trajectories $U_t$ to estimate the transition function $P$, which enables it to estimate its performance criterion.*

## 3. Occupancy Measures

We would like to reformulate the learner's objective in order to approach the problem with techniques from online learning. For this purpose we introduce the concept of occupancy measures (Zimin & Neu, 2013) on the space $X \times A \times X$. For a policy $\pi$ and a transition function $P$ we define the occupancy measure $q^{P,\pi}$ as follows:

$$q^{P,\pi}(x,a,x') = \Pr\left[x_k = x, a_k = a, x_{k+1} = x'|P,\pi\right]$$

where $x \in X_k$ and $x' \in X_{k+1}$. Another notation we will be using is $k(x)$ for the index of the layer that $x$ belongs to.

We start with two basic properties that hold for every occupancy measure $q$. From the loop-free assumption we know that in each episode the learner will go through every layer. Therefore, for every $k = 0,\dots,L-1$,

$$\sum_{x\in X_k}\sum_{a\in A}\sum_{x'\in X_{k+1}} q(x,a,x') = 1 \tag{1}$$

Moreover, the probability to enter a state when coming from the previous layer is exactly the probability to visit that state. Thus, for every $k = 1,\dots,L-1$ and every $x \in X_k$,

$$\sum_{x'\in X_{k+1}}\sum_{a\in A} q(x,a,x') = \sum_{x'\in X_{k-1}}\sum_{a\in A} q(x',a,x) \tag{2}$$

Notice that every occupancy measure $q$ induces a transition function and a policy. We denote them as $P^q$ and $\pi^q$ respectively, and they can be computed as follows:

$$P^q(x'|x,a) = \frac{q(x,a,x')}{\sum_{y\in X_{k(x)+1}} q(x,a,y)}$$

$$\pi^q(a|x) = \frac{\sum_{x'\in X_{k(x)+1}} q(x,a,x')}{\sum_{b\in A}\sum_{x'\in X_{k(x)+1}} q(x,b,x')}$$

We denote the set of all occupancy measures of an MDP $M$ as $\Delta(M)$. The following lemma characterizes $\Delta(M)$ and its proof is straightforward.

**Lemma 3.1.** *For every $q \in [0,1]^{|X|\times|A|\times|X|}$ it holds that $q \in \Delta(M)$ if and only if (1) and (2) hold, and $P^q = P$ (where $P$ is the transition function of $M$).*

We can use occupancy measures to reformulate the regret. We say that a performance criterion $\mathcal{C}$ is convexly-measurable if there exists some convex function $f^{\mathcal{C}} : [0,1]^{|X|\times|A|\times|X|} \to \mathbb{R}_{\ge 0}$, such that

$$\mathcal{C}\left(\mathbb{E}\left[\ell(U)|P,\pi\right]\right) = f^{\mathcal{C}}(q^{P,\pi};\ell)$$

holds for every policy $\pi$ and every transition function $P$. We call $f^{\mathcal{C}}$ the criterion function of $\mathcal{C}$. Since our algorithm requires only the criterion function, performance criteria can also be defined implicitly through criterion functions.

If we redefine the task of the learner from having to select individual actions (or policies) to having to select occupancy measures $q_t \in \Delta(M)$ in each episode $t$, for convexly-measurable performance criteria we can rewrite the regret to obtain an instance of online convex optimization with decision space $\Delta(M)$, i.e.,

$$\begin{aligned}\hat{R}^{\mathcal{C}}_{1:T} &= \hat{L}^{\mathcal{C}}_{1:T}(\{\ell_t\}_{t=1}^T) - \min_{\pi} L^{\mathcal{C}}_{1:T}(\pi;\{\ell_t\}_{t=1}^T)\\ &= \sum_{t=1}^T f^{\mathcal{C}}(q_t;\ell_t) - \min_{q\in\Delta(M)}\sum_{t=1}^T f^{\mathcal{C}}(q;\ell_t)\\ &= \max_{q\in\Delta(M)}\sum_{t=1}^T f^{\mathcal{C}}(q_t;\ell_t) - f^{\mathcal{C}}(q;\ell_t)\end{aligned}$$

The following lemma shows that all performance criterion examples presented in the previous section are indeed convexly-measurable, and gives a way to build more convexly-measurable performance criteria.

**Lemma 3.2.** *If a performance criterion $\mathcal{C}$ has the following form,*

$$\mathcal{C}\left(\{v_k\}_{k=0}^{L-1}\right) = g\left(\left\{\sum_{k=0}^{L-1} h_j(v_k)\right\}_{j=1}^{m}\right)$$

*where $v_k \in \mathbb{R}^d$, $h_j : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ are arbitrary functions and $g : \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ is a convex function, then $\mathcal{C}$ can be modeled as a convexly-measurable performance criterion.*

*Proof.* For any loss function $\ell'$, policy $\pi$ and transition function $P$, we have that

$$\begin{aligned}\mathcal{C}^{TEL}(\mathbb{E}[\ell'(U)|P,\pi]) &= \sum_{k=0}^{L-1}\mathbb{E}\left[\ell'(x_k,a_k,x_{k+1})\middle|P,\pi\right] \\ &= \mathbb{E}\left[\sum_{k=0}^{L-1}\ell'(x_k,a_k,x_{k+1})\middle|P,\pi\right] \\ &= \sum_{x,a,x'} q^{P,\pi}(x,a,x')\ell'(x,a,x') \overset{def}{=} \langle q^{P,\pi},\ell'\rangle\end{aligned}$$

Therefore the criterion function of $\mathcal{C}^{TEL}$ is $f^{\mathcal{C}^{TEL}}(q;\ell) = \langle q,\ell\rangle$. We can model $\mathcal{C}$ with $m$-dimension losses, such that dimension $j$ features loss function $h_j(\ell)$, and then $\mathcal{C}$ just needs to sum up the $L$ losses and apply $g$. Thus, the criterion function of $\mathcal{C}$ will be

$$f^{\mathcal{C}}(q;\ell) = g\left(\{\langle q, h_j(\ell)\rangle\}_{j=1}^{m}\right)$$

Finally, $f^{\mathcal{C}}$ is convex because the composition of a convex function and a linear function is convex (Boyd & Vandenberghe, 2004). □

## 4. The Algorithm

We call our algorithm, which is presented in algorithms 2 and 3, "Upper Confidence Online Relative Entropy Policy Search" (UC-O-REPS). It is inspired by the O-REPS algorithm (Zimin & Neu, 2013) in the sense that it picks occupancy measures instead of policies. However, unlike our algorithm, O-REPS assumes full knowledge of the transition function. To the best of our knowledge, the only algorithm that handles unknown transition probabilities in adversarial MDPs is FPOP (Neu et al., 2012), which uses a Follow the Pertubed Leader method (Kalai & Vempala, 2003) in the space of the policies.

Recall that the adversarial MDP has a stochastic element - the transition function, and an adversarial element - the loss functions.

To handle the stochastic transition function we use the framework of epochs and confidence sets, first introduced by the UCRL-2 algorithm (Auer et al., 2008). In this framework, the algorithm maintains confidence sets that contain the actual MDP with high probability, but also shrink as time progresses.We translated this method to the occupancy measures space, and the full details can be found in Section 4.1.

The core of the algorithm is the way we choose the occupancy measure for each episode from within the confidence set. This is done by the Online Mirror Descent method (Shalev-Shwartz, 2012) for online linear optimization, since we deal with an arbitrary sequence of loss functions. The full details of adapting OMD to our setting can be found in Section 4.2.

The combination of these two methods is done using an important principle in reinforcement learning - "optimism in face of uncertainty". On the one hand, we keep confidence sets to handle the uncertainty, but on the other hand, within these confidence sets, we solve an OMD optimization problem optimistically (without thinking about the transition function estimation).

### 4.1. Confidence Sets

Since the learner does not know the transition function, it has to estimate $P$ from its experience. Using this estimate we define confidence sets, and choose occupancy measures from within them. Notice that these occupancy measures might not be in $\Delta(M)$, i.e., their induced transition function may differ from $P$. Nevertheless, we can still use them to compute policies and execute those policies.

The algorithm proceeds in epochs of random length, and in the beginning of each epoch the confidence set is updated. The first epoch $E_1$ starts at episode $t = 1$, and each epoch $E_i$ ends when the number of visits at some state-action pair $(x,a)$ is doubled. Let $t_i$ denote the index of the first episode in epoch $E_i$, and $i(t)$ denote the index of the epoch that includes episode $t$. Let $N_i(x,a)$ and $M_i(x'|x,a)$ denote the number of times state-action pair $(x,a)$ was visited and the number of times this event was followed by a transition to $x'$ up to episode $t_i$, respectively. That is

$$N_i(x,a) = \sum_{s=1}^{t_i-1}\mathbb{I}\left\{x_k^{(s)} = x, a_k^{(s)} = a\right\}$$

$$M_i(x'|x,a) = \sum_{s=1}^{t_i-1}\mathbb{I}\left\{x_k^{(s)} = x, a_k^{(s)} = a, x_{k+1}^{(s)} = x'\right\}$$

where $k = k(x)$.

Our estimate $\bar{P}_i$ for the transition function in epoch $E_i$ is

$$\bar{P}_i(x'|x,a) = \frac{M_i(x'|x,a)}{\max\{1, N_i(x,a)\}}$$

and we define our confidence set $\Delta(M,i)$ in epoch $E_i$ to include all the occupancy measures that their induced transition function is "close enough" to $\bar{P}_i$. More formally, given a confidence parameter $\delta > 0$, we define

$$\epsilon_i(x,a) = \sqrt{\frac{2|X_{k(x)+1}| \ln \frac{T|X||A|}{\delta}}{\max\{1, N_i(x,a)\}}}$$

and say that $\Delta(M,i)$ consists of all $q \in [0,1]^{|X|\times|A|\times|X|}$ for which (1) and (2) hold, and

$$\left\| P^q(\cdot|x,a) - \bar{P}_i(\cdot|x,a) \right\|_1 \leq \epsilon_i(x,a) \tag{3}$$

for every $(x,a) \in X \times A$.

Notice that these confidence sets shrink as time progresses, but the following lemma (Auer et al., 2008; Neu et al., 2012) shows that they still contain $\Delta(M)$ with high probability.

**Lemma 4.1.** *For any $0 < \delta < 1$*

$$\left\| P(\cdot|x,a) - \bar{P}_i(\cdot|x,a) \right\|_1 \leq \sqrt{\frac{2|X_{k(x)+1}| \ln \frac{T|X||A|}{\delta}}{\max\{1, N_i(x,a)\}}}$$

*holds with probability at least $1-\delta$ simultaneously for all $(x,a) \in X \times A$ and all epochs.*

### 4.2. Optimization Problem

In order to choose the occupancy measure $q_t$ for episode $t$, the algorithm follows the OMD method. The idea behind this method is to choose an occupancy measure that minimizes the loss in episode $t$, while not straying too far from the previously chosen occupancy measure. Formally, given a parameter $\eta > 0$,

$$q_{t+1} = \arg \min_{q \in \Delta(M, i(t))} \eta \langle q, z_t \rangle + D(q||q_t)$$

where $z_t \in \partial f^{\mathcal{C}}(q_t; \ell_t)$ is a sub-gradient and $D(q||q_t)$ is the unnormalized KL divergence between two occupancy measures defined as

$$D(q||q') = \sum_{x,a,x'} q(x,a,x') \ln \frac{q(x,a,x')}{q'(x,a,x')} - q(x,a,x') + q'(x,a,x')$$

We now proceed to show that this optimization problem can be solved efficiently. From the theory of OMD it is known that we can split this problem as follows: we start by solving the unconstrained problem, and then project the unconstrained minimizer into the feasible set, namely,

$$\begin{aligned} \tilde{q}_{t+1} &= \arg\min_q \eta \langle q, z_t \rangle + D(q||q_t) \\ q_{t+1} &= \arg \min_{q \in \Delta(M, i(t))} D(q||\tilde{q}_{t+1}) \end{aligned} \tag{4}$$

The unconstrained problem can be easily solved by setting $\tilde{q}_{t+1}(x,a,x') = q_t(x,a,x') e^{-\eta z_t(x,a,x')}$ for every $(x,a,x') \in X \times A \times X_{k(x)+1}$. Theorem 4.2 shows that the second optimization problem can be reduced to a convex optimization problem with only non-negativity constraints (and no constraints about the relations between the variables), which can be solved efficiently using iterative methods (Boyd & Vandenberghe, 2004).

Before stating the theorem we consider some definitions that will simplify its formulation. Let $v : X \times A \times X \to \mathbb{R}$ be a value function and $e : X \times A \times X \to \mathbb{R}$ be an error function. We use $v$ and $e$ to define an estimated Bellman error.

**Definition 4.1.** *For every $t = 1, \ldots, T$ define the estimated Bellman error for episode $t$, given value function $v$ and error function $e$, as*

$$B_t^{v,e}(x,a,x') = e(x,a,x') + v(x,a,x') - \eta z_t(x,a,x') - \sum_{y \in X_{k(x)+1}} \bar{P}_{i(t)}(y|x,a) v(x,a,y)$$

We would like to define a parameterization to $v$ and $e$ using variables that will later be known as Lagrange multipliers. Let $\beta : X \to \mathbb{R}$ and let $\mu = (\mu^+, \mu^-)$ such that $\mu^+, \mu^- : X \times A \times X \to \mathbb{R}_{\geq 0}$. We define the following parameterization to $v$ and $e$ using $\beta$ and $\mu$.

$$\begin{aligned} v^\mu(x,a,x') &= \mu^-(x,a,x') - \mu^+(x,a,x') \\ e^{\mu,\beta}(x,a,x') &= (\mu^+(x,a,x') + \mu^-(x,a,x'))\epsilon_{i(t)}(x,a) \\ &\quad + \beta(x') - \beta(x) \end{aligned}$$

Now we are ready to state the theorem.

**Theorem 4.2.** *Let $t > 1$ and define the function*

$$Z_t^k(v,e) = \sum_{x \in X_k} \sum_{a \in A} \sum_{x' \in X_{k+1}} q_t(x,a,x') e^{B_t^{v,e}(x,a,x')}$$

*Then the solution to optimization problem (4) is*

$$q_{t+1}(x,a,x') = \frac{q_t(x,a,x') e^{B_t^{v^{\mu_t}, e^{\mu_t,\beta_t}}(x,a,x')}}{Z_t^{k(x)}(v^{\mu_t}, e^{\mu_t,\beta_t})}$$

*where*

$$\beta_t, \mu_t = \arg \min_{\beta, \mu \geq 0} \sum_{k=0}^{L-1} \ln Z_t^k(v^\mu, e^{\mu,\beta}) \tag{5}$$

*Proof.* First of all we would like to reformulate optimization problem (4) as a convex optimization problem. Notice that the target function is convex (since it is the KL-divergence) and so are constraints (1), (2) of $\Delta(M,i)$ (where $i = i(t)$). As for constraint (3), we will need to write it differently.

**Algorithm 2** UC-O-REPS Algorithm

**Input:** state space $X$, action space $A$, time horizon $T$, convexly-measurable performance criterion $\mathcal{C}$ with its criterion function $f^{\mathcal{C}}$, optimization parameter $\eta$ and confidence parameter $\delta$.

**Initialization:**
start first epoch: $i(1) \leftarrow 1 \quad ; \quad t_1 \leftarrow 1$
initialize counters $\forall(x,a,x')$:

$$n_1(x,a) \leftarrow 0 \quad ; \quad N_1(x,a) \leftarrow 0$$
$$m_1(x'|x,a) \leftarrow 0 \quad ; \quad M_1(x'|x,a) \leftarrow 0$$

initialize first policy $\forall(x,a)$: $\pi_1(a|x) \leftarrow \frac{1}{|A|}$
initialize first occupancy measure $\forall k \quad \forall(x,a,x') \in X_k \times A \times X_{k+1}$: $q_1(x,a,x') \leftarrow \frac{1}{|X_k||A||X_{k+1}|}$

**for** $t = 1$ **to** $T$ **do**
traverse trajectory $U_t$ using policy $\pi_t$
observe loss function $\ell_t$
update epoch counters $\forall k$:

$$n_{i(t)}(x_k^{(t)}, a_k^{(t)}) \leftarrow n_{i(t)}(x_k^{(t)}, a_k^{(t)}) + 1$$
$$m_{i(t)}(x_{k+1}^{(t)}|x_k^{(t)}, a_k^{(t)}) \leftarrow m_{i(t)}(x_{k+1}^{(t)}|x_k^{(t)}, a_k^{(t)}) + 1$$

**if** $\exists(x,a) \in X \times A. \quad n_{i(t)}(x,a) \geq N_{i(t)}(x,a)$ **then**
start new epoch:

$$i(t+1) \leftarrow i(t) + 1 \quad ; \quad t_{i(t+1)} \leftarrow t+1$$

initialize epoch counters $\forall(x,a,x')$:

$$n_{i(t+1)}(x,a) \leftarrow 0 \quad ; \quad m_{i(t+1)}(x'|x,a) \leftarrow 0$$

update total counters $\forall(x,a,x')$:

$$N_{i(t+1)}(x,a) \leftarrow N_{i(t)}(x,a) + n_{i(t)}(x,a)$$
$$M_{i(t+1)}(x'|x,a) \leftarrow M_{i(t)}(x'|x,a) + m_{i(t)}(x'|x,a)$$

compute probability estimate $\forall(x,a,x')$:

$$\bar{P}_{i(t+1)}(x'|x,a) \leftarrow \frac{M_{i(t+1)}(x'|x,a)}{\max\left\{1, N_{i(t+1)}(x,a)\right\}}$$

**else**
continue in the same epoch: $i(t+1) \leftarrow i(t)$
**end if**
compute policy for next episode:

$$q_{t+1}, \pi_{t+1} \leftarrow \texttt{Comp-Policy}(q_t, \bar{P}_{i(t+1)}, \ell_t, f^{\mathcal{C}})$$

**end for**

**Algorithm 3** Comp-Policy Procedure

**Input:** previous occupancy measure $q_t$, transition function estimate $\bar{P}_{i(t+1)}$, current loss function $\ell_t$ and convex criterion function $f^{\mathcal{C}}$.

obtain sub-gradient $z_t \in \partial f^{\mathcal{C}}(q_t; \ell_t)$
solve optimization problem (5):

$$\beta_t, \mu_t = \arg\min_{\beta,\mu \geq 0} \sum_{k=0}^{L-1} \ln Z_t^k(v^{\mu}, e^{\mu,\beta})$$

compute next occupancy measure $\forall(x,a,x')$:

$$q_{t+1}(x,a,x') = \frac{q_t(x,a,x') e^{B^{v^{\mu_t}, e^{\mu_t,\beta_t}}(x,a,x')}}{Z_t^{k(x)}(v^{\mu_t}, e^{\mu_t,\beta_t})}$$

compute next policy $\forall(x,a)$:

$$\pi_{t+1}(a|x) = \frac{\sum_{x' \in X_{k(x)+1}} q_{t+1}(x,a,x')}{\sum_{b \in A} \sum_{x' \in X_{k(x)+1}} q_{t+1}(x,b,x')}$$

---

Let $(x,a) \in X \times A$, we can replace

$$\left\| \frac{q(x,a,\cdot)}{\sum_{y \in X_{k(x)+1}} q(x,a,y)} - \bar{P}_i(\cdot|x,a) \right\|_1 \leq \epsilon_i(x,a)$$

with $|X_{k(x)+1}| + 1$ constraints as follows. For each $x' \in X_{k(x)+1}$ we bound the difference in the transition probability with a new variable $\epsilon'(x,a,x')$ and then we bound their sum with the original bound $\epsilon_i(x,a)$. That is

$$\left| \frac{q(x,a,x')}{\sum_{y \in X_{k(x)+1}} q(x,a,y)} - \bar{P}_i(x'|x,a) \right| \leq \epsilon'(x,a,x')$$
$$\sum_{x' \in X_{k(x)+1}} \epsilon'(x,a,x') \leq \epsilon_i(x,a)$$

Now we can get rid of the denominator by multiplying the equation and then replacing $\epsilon'(x,a,x')$ with a different variable $\epsilon(x,a,x') = \epsilon'(x,a,x') \sum_{y \in X_{k(x)+1}} q(x,a,y)$. Moreover, we will discard the absolute value by replacing it with two linear constraints. The resulting constraints are,

$$q(x,a,x') - \bar{P}_i(x'|x,a) \sum_{y \in X_{k(x)+1}} q(x,a,y) \leq \epsilon(x,a,x')$$
$$\bar{P}_i(x'|x,a) \sum_{y \in X_{k(x)+1}} q(x,a,y) - q(x,a,x') \leq \epsilon(x,a,x')$$
$$\sum_{x' \in X_{k(x)+1}} \epsilon(x,a,x') \leq \epsilon_i(x,a) \sum_{x' \in X_{k(x)+1}} q(x,a,x')$$

This gives us a convex optimization problem with linear constraints. This problem obtains strong duality because: (1) The target function is bounded from below because KL-divergence is non-negative, (2) The target function and all constraints are convex, (3) Slater condition holds (easy to check).

Thus we can use the method of Lagrange multipliers, and we are ensured that the solution we get is optimal and finite. The full derivation can be found in the supplementary material and yields the aforementioned result. □

## 5. Analysis

In this section we bound the regret of the UC-O-REPS algorithm, by combining ideas from the regret analyses of OMD and UCRL-2. First we partition the regret into two terms: $\hat{R}^{APP}_{1:T}$ - which includes the error that comes from the estimation of the unknown transition function, and $\hat{R}^{ON}_{1:T}$ - which includes the error that comes from choosing sub-optimal policies. Formally,

$$\begin{aligned}
\hat{R}^{\mathcal{C}}_{1:T} &= \hat{L}^{\mathcal{C}}_{1:T}(\{\ell_t\}_{t=1}^T) - \min_{\pi} L^{\mathcal{C}}_{1:T}(\pi;\{\ell_t\}_{t=1}^T) \\
&= \sum_{t=1}^T \mathcal{C}(\mathbb{E}\left[\ell_t(U)|P,\pi_t\right]) - \sum_{t=1}^T \mathcal{C}(\mathbb{E}\left[\ell_t(U)|P,\pi\right]) \\
&= \left(\sum_{t=1}^T \mathcal{C}(\mathbb{E}\left[\ell_t(U)|P,\pi_t\right]) - \mathcal{C}(\mathbb{E}\left[\ell_t(U)|P_t,\pi_t\right])\right) \\
&\quad + \left(\sum_{t=1}^T \mathcal{C}(\mathbb{E}\left[\ell_t(U)|P_t,\pi_t\right]) - \mathcal{C}(\mathbb{E}\left[\ell_t(U)|P,\pi\right])\right) \\
&\overset{def}{=} \hat{R}^{APP}_{1:T} + \hat{R}^{ON}_{1:T}
\end{aligned}$$

where $P_t = P^{q_t}$ and $\pi_t = \pi^{q_t}$.

Notice that $\mathcal{C}(\mathbb{E}\left[\ell_t(U)|P_t,\pi_t\right]) = f^{\mathcal{C}}(q_t;\ell_t)$ but it isn't the case with $\mathcal{C}(\mathbb{E}\left[\ell_t(U)|P,\pi_t\right])$ because $q_t$ is not necessarily an occupancy measure of $M$. Theorems 5.2 and 5.3 bound each of these terms, which yields our main result.

**Theorem 5.1.** *Let $M = \left(X, A, P, \{\ell_t\}_{t=1}^T\right)$ be an episodic loop-free adversarial MDP, and let $\mathcal{C}$ be a convexly-measurable performance criterion such that $f^{\mathcal{C}}$ is $F$-Lipschitz. Then, with probability at least $1-2\delta$, UC-O-REPS with $\eta = \sqrt{\frac{\ln \frac{|X|^2|A|}{L^2}}{F^2 T}}$ achieves the following regret,*

$$\hat{R}^{\mathcal{C}}_{1:T} \leq 15FL|X|\sqrt{T|A|\ln\frac{T|X||A|}{\delta}}$$

An immediate corollary of this theorem is the regret bound in the classical case of total expected loss performance criterion.

**Corollary 5.1.** *Running UC-O-REPS in an episodic loop-free adversarial MDP $M = \left(X, A, P, \{\ell_t\}_{t=1}^T\right)$ yields the following regret with respect to the total expected loss, when setting $\delta = \frac{|X||A|}{T}$,*

$$\hat{R}^{\mathcal{C}^{TEL}}_{1:T} \leq 25L|X|\sqrt{T|A|\ln T}$$

*Proof.* For the total expected loss performance criterion we have that $f^{\mathcal{C}^{TEL}}(q_t;\ell_t) = \langle q_t, \ell_t\rangle$ and therefore the gradient of $f^{\mathcal{C}}$ is $z_t = \ell_t$. Since the losses are bounded by 1, we have that $f^{\mathcal{C}}$ is 1-Lipschitz, i.e., $F = 1$.

Recall that in this case the regret is an expectation. With probability at least $1-2\delta$ it is bounded using Theorem 5.1, and with probability at most $2\delta$ we have a worst case bound of $TL$. Substituting $\delta$ and using the law of total expectation finishes the proof. □

### 5.1. Bounding $\hat{R}^{APP}_{1:T}$

The term $\hat{R}^{APP}_{1:T}$ is a result of the learner's lack of knowledge about the environment's dynamics. Since the dynamics are stochastic the learner estimates the transition probabilities to build confidence sets. It then selects occupancy measures from within these confidence sets, but they are not exactly occupancy measures of $M$.

In this section we bound the difference between the loss of the learner's chosen policies in $M$ and the loss of these policies in the "optimistic" MDP (the one induced by the occupancy measure $q_t$), i.e.,

$$\hat{R}^{APP}_{1:T} = \sum_{t=1}^T \mathcal{C}(\mathbb{E}\left[\ell_t(U)|P,\pi_t\right]) - \mathcal{C}(\mathbb{E}\left[\ell_t(U)|P_t,\pi_t\right])$$

The way the algorithm minimizes this difference is through shrinking of the confidence sets. The following bound on $\hat{R}^{APP}_{1:T}$ is adapted from arguments in the regret analysis of UCRL-2, and the proof can be found in the supplementary material.

**Theorem 5.2.** *Let $M = \left(X, A, P, \{\ell_t\}_{t=1}^T\right)$ be an episodic loop-free adversarial MDP, and let $\mathcal{C}$ be a convexly-measurable performance criterion such that $f^{\mathcal{C}}$ is $F$-Lipschitz. Then, with probability at least $1-2\delta$, UC-O-REPS obtains,*

$$\hat{R}^{APP}_{1:T} \leq 3FL|X|\left(2\sqrt{T\ln\frac{L}{\delta}} + 3\sqrt{T|A|\ln\frac{T|X||A|}{\delta}}\right)$$

### 5.2. Bounding $\hat{R}^{ON}_{1:T}$

The term $\hat{R}^{ON}_{1:T}$ is a result of the learner's lack of knowledge about the loss functions. Since the sequence of loss functions can be arbitrary, the learner handles it with tools from online convex optimization.

In this section we ignore the fact that the occupancy measures chosen by the learner are not exactly occupancy measures of $M$, since this issue was already addressed in the previous section bounding $\hat{R}_{1:T}^{APP}$. Here we are only interested in the following difference

$$\hat{R}_{1:T}^{ON} = \sum_{t=1}^{T} \mathcal{C}(\mathbb{E}\left[\ell_t(U)|P_t, \pi_t\right]) - \mathcal{C}(\mathbb{E}\left[\ell_t(U)|P, \pi\right])$$

First we use the connection between $\mathcal{C}$ and $f^{\mathcal{C}}$, and the convexity of $f^{\mathcal{C}}$ to obtain

$$\hat{R}_{1:T}^{ON} = \sum_{t=1}^{T} f^{\mathcal{C}}(q_t; \ell_t) - f^{\mathcal{C}}(q; \ell_t) \leq \sum_{t=1}^{T} \langle q_t - q, z_t \rangle$$

where $z_t \in \partial f^{\mathcal{C}}(q_t; \ell_t)$.

Now we can use arguments from online linear optimization. Specifically, the following theorem is an adaptation of OMD regret analysis to our setting.

**Theorem 5.3.** *Let $M = \left(X, A, P, \{\ell_t\}_{t=1}^{T}\right)$ be an episodic loop-free adversarial MDP, and let $\mathcal{C}$ be a convexly-measurable performance criterion such that $f^{\mathcal{C}}$ is $F$-Lipschitz. Then, with probability at least $1-\delta$, UC-O-REPS obtains the following for every $q \in \Delta(M)$.*

$$\hat{R}_{1:T}^{ON} \leq \sum_{t=1}^{T} \langle q_t - q, z_t \rangle \leq \eta F^2 LT + \frac{L \ln \frac{|X|^2|A|}{L^2}}{\eta}$$

*and setting $\eta = \sqrt{\frac{\ln \frac{|X|^2|A|}{L^2}}{F^2 T}}$ yields*

$$\hat{R}_{1:T}^{ON} \leq 2FL\sqrt{2T \ln \frac{|X||A|}{L}}$$

*where $q_t$ is the occupancy measure chosen by UC-O-REPS in episode $t$, and $z_t \in \partial f^{\mathcal{C}}(q_t; \ell_t)$.*

*Proof.* By standard arguments of OMD regret analysis (the full proof can be found in the full version of the paper) we have that

$$\sum_{t=1}^{T} \langle q_t - q, z_t \rangle \leq \sum_{t=1}^{T} \langle q_t - \tilde{q}_{t+1}, z_t \rangle + \frac{D(q||q_1)}{\eta}$$

However these arguments assume that $q_t$ are chosen from within $\Delta(M)$ so we need to show that they are still valid. From Lemma 4.1 we know that $\Delta(M) \subseteq \Delta(M, i)$ for every $i$ with probability at least $1-\delta$. Therefore, by choosing approximate occupancy measures we can only improve the regret so the arguments are indeed valid.

Using the exact form of $\tilde{q}_{t+1}$ and the fact that $e^x \geq 1 + x$, we get that

$$\tilde{q}_{t+1}(x, a, x') \geq q_t(x, a, x') - \eta q_t(x, a, x') z_t(x, a, x')$$

and therefore

$$\begin{aligned}\sum_{t=1}^{T} \langle q_t - \tilde{q}_{t+1}, z_t \rangle &\leq \eta \sum_{t=1}^{T} \sum_{x,a,x'} q_t(x, a, x') z_t^2(x, a, x') \\ &\leq \eta F^2 \sum_{t=1}^{T} \sum_{x,a,x'} q_t(x, a, x') = \eta F^2 LT\end{aligned}$$

For the second term, $D(q||q_1)/\eta$, we use the fact that the unnormalized KL divergence is the Bregman divergence associated with the unnormalized negative entropy, defined as follows.

$$R(q) = \sum_{x,a,x'} q(x, a, x') \ln q(x, a, x') - q(x, a, x')$$

Now from standard arguments we obtain

$$\begin{aligned}D(q||q_1) &\leq R(q) - R(q_1) \\ &\leq \sum_{x \in X} \sum_{a \in A} \sum_{x' \in X_{k(x)+1}} q_1(x, a, x') \ln \frac{1}{q_1(x, a, x')} \\ &\leq \sum_{k=0}^{L-1} \ln |X_k||A||X_{k+1}| \leq L \ln \frac{|X|^2|A|}{L^2}\end{aligned}$$

Putting these two bounds together completes the proof. □

## 6. Conclusions and Future Work

In this paper we considered online learning in adversarial MDPs where the transition function is not known to the learner and the losses can change arbitrarily between episodes, and showed an algorithm that achieves $\tilde{O}(L|X|\sqrt{T|A|})$ regret. The algorithm is based on a combination of the OMD method for online convex optimization, and the UCRL-2 algorithm for reinforcement learning. Moreover, we extended the adversarial MDP model to include convex performance criteria, and showed that our algorithm achieves near-optimal regret bounds in this model as well.

The natural open problem is whether the lower bound of $\Omega(\sqrt{L|X||A|T})$ (Auer et al., 2008) can be achieved in this model. An algorithm that achieves this will have to build upon a different method than UCRL-2, and it will be interesting to see if the techniques of Azar et al. (2017) can be implemented here. Another interesting open question is to consider bandit feedback when the transition function is unknown. This question seems to be difficult because the natural approach of building an unbiased estimator for the losses cannot be implemented easily, since the natural construction of inverse probability estimator requires knowledge of the transition probabilities.

## Acknowledgements

This work was supported in part by a grant from the Israel Science Foundation (ISF) and by the Tel Aviv University Yandex Initiative in Machine Learning.

## References

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 89–96, 2008.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 263–272, 2017.

Bartlett, P. L. and Tewari, A. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 35–42, 2009.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521833787.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, 2006. ISBN 978-0-521-84108-5.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Online Markov Decision Processes. *Math. Oper. Res.*, 34(3): 726–736, 2009. (preliminary version NIPS 2004).

Kalai, A. and Vempala, S. Efficient algorithms for online decision problems. In *16th Annual Conference on Computational Learning Theory (COLT)*, pp. 26–40, 2003.

Neu, G., György, A., and Szepesvári, C. The online loop-free stochastic shortest-path problem. In *Conference on Learning Theory (COLT)*, pp. 231–243, 2010.

Neu, G., György, A., and Szepesvári, C. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, (AISTATS)*, pp. 805–813, 2012.

Neu, G., György, A., Szepesvári, C., and Antos, A. Online Markov Decision Processes under bandit feedback. *IEEE Trans. Automat. Contr.*, 59(3):676–691, 2014.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994. ISBN 0471619779.

Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

Yu, J. Y., Mannor, S., and Shimkin, N. Markov Decision Processes with arbitrary reward processes. *Math. Oper. Res.*, 34(3):737–757, 2009.

Zimin, A. and Neu, G. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, pp. 1583–1591, 2013.