# Online Convex Optimization in Adversarial MDPs

——陈宏俊

# Outline

- Problem Formulation
- Occupancy Measures
- The Algorithm
- Analysis
- Conclusion and Future Work

多智体人工智能实验室
Multi-Agent Artificial Intelligence Laboratory

# Problem Formulation

多智体人工智能实验室
Multi-Agent Artificial Intelligence Laboratory

- Definition of the episodic loop-free adversarial MDP

$$M = (X, A, P, \{\ell_t\}_{t=1}^T)$$

- P: X×A×X → [0,1] is the transition function, the probability to move to state ***x'*** when performing action ***a*** in state ***x***

$$P(x'|x, a)$$

- Assuming that the state space can be decomposed into ***L*** non-intersecting layers $X_0, \ldots, X_L$ such that the first and last layers are singletons.
- Furthermore, the loop-free assumption means that transitions are only possible between consecutive layers.
- Let $\{\ell_t\}_{t=1}^T$ be a sequence of loss function describing the losses at each episode

# Learner-Environment Interaction

多智体人工智能实验室
Multi-Agent Artificial Intelligence Laboratory

**Algorithm 1** Learner-Environment Interaction

**Parameters:** MDP $M = (X, A, P, \{\ell_t\}_{t=1}^{T})$ 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**

# The goal of the Learner

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.

$$U = (x_0, a_0, x_1, a_1, \ldots, x_{L-1}, a_{L-1}, x_L)$$

$$\ell(U) = \left\{\ell(x_k, a_k, x_{k+1})\right\}_{k=0}^{L-1}$$

# Performance Criterion

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

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

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

# Regret of the Learner



- Total loss:

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

- Learner's regret:

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

*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.*

# Occupancy Measures

- To reformulate the learner's objective for online learning, it was supposed to introduce the definition of occupancy measure:

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

- Two basic properties:

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

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

# Transition Function & Policy

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

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

**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 $f^{\mathcal{C}} : [0,1]^{|X|\times|A|\times|X|} \to \mathbb{R}_{\geq 0}$, 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 $P$.

# Regret Rewriting

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

**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

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

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

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

# Confidence Sets

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

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

where $k = k(x)$.

# Confidence Sets

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.*

# Optimization Problem



- With the 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

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

# Problem Splitting



- 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 $(x,a,x') \in X \times A \times X_{k(x)+1}$.

# Bellman error

多智体人工智能实验室
Multi-Agent Artificial Intelligence Laboratory

**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 $\mu$.

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

# Proof

多智体人工智能实验室
Multi-Agent Artificial Intelligence Laboratory

$$
\begin{aligned}
&\min_{q,\epsilon} D(q||\tilde{q}_{t+1}) \\
&s.t. \sum_{x\in X_k}\sum_{a\in A}\sum_{x'\in X_{k+1}} q(x,a,x') = 1 && \forall k=0,\ldots,L-1 \\
&\quad \sum_{x'\in X_{k+1}}\sum_{a\in A} q(x,a,x') = \sum_{x'\in X_{k-1}}\sum_{a\in A} q(x',a,x) && \forall k=1,\ldots,L-1 \quad \forall x\in X_k \\
&\quad q(x,a,x') - \bar{P}_i(x'|x,a)\sum_{y\in X_{k+1}} q(x,a,y) \le \epsilon(x,a,x') && \forall k=0,\ldots,L-1 \quad \forall (x,a,x')\in X_k\times A\times X_{k+1} \\
&\quad \bar{P}_i(x'|x,a)\sum_{y\in X_{k+1}} q(x,a,y) - q(x,a,x') \le \epsilon(x,a,x') && \forall k=0,\ldots,L-1 \quad \forall (x,a,x')\in X_k\times A\times X_{k+1} \\
&\quad \sum_{x'\in X_{k+1}} \epsilon(x,a,x') \le \epsilon_i(x,a)\sum_{x'\in X_{k+1}} q(x,a,x') && \forall k=0,\ldots,L-1 \quad \forall (x,a)\in X_k\times A \\
&\quad q(x,a,x') \ge 0 && \forall k=0,\ldots,L-1 \quad \forall (x,a,x')\in X_k\times A\times X_{k+1}
\end{aligned}
$$

# Lagrangian Form

$$
\begin{aligned}
\mathcal{L}(q,\epsilon) = D(q||\tilde{q}_{t+1}) &+ \sum_{k=0}^{L-1} \lambda_k \left( \sum_{x\in X_k} \sum_{a\in A} \sum_{x'\in X_{k+1}} q(x,a,x') - 1 \right) \\
&+ \sum_{k=1}^{L-1} \sum_{x\in X_k} \beta(x) \left( \sum_{a\in A} \sum_{x'\in X_{k+1}} q(x,a,x') - \sum_{a\in A} \sum_{x'\in X_{k-1}} q(x',a,x) \right) \\
&+ \sum_{k=0}^{L-1} \sum_{x\in X_k} \sum_{a\in A} \sum_{x'\in X_{k+1}} \mu^+(x,a,x') \left( q(x,a,x') - \bar{P}_i(x'|x,a) \sum_{y\in X_{k+1}} q(x,a,y) - \epsilon(x,a,x') \right) \\
&+ \sum_{k=0}^{L-1} \sum_{x\in X_k} \sum_{a\in A} \sum_{x'\in X_{k+1}} \mu^-(x,a,x') \left( \bar{P}_i(x'|x,a) \sum_{y\in X_{k+1}} q(x,a,y) - q(x,a,x') - \epsilon(x,a,x') \right) \\
&+ \sum_{k=0}^{L-1} \sum_{x\in X_k} \sum_{a\in A} \mu(x,a) \left( \sum_{x'\in X_{k+1}} \epsilon(x,a,x') - \epsilon_i(x,a) \sum_{x'\in X_{k+1}} q(x,a,x') \right)
\end{aligned}
$$

Let $(x,a,x') \in X \times A \times X_{k(x)+1}$ and consider the derivative with respect to $\epsilon(x,a,x')$.

$$
\frac{\partial \mathcal{L}}{\partial \epsilon(x,a,x')} = -\mu^+(x,a,x') - \mu^-(x,a,x') + \mu(x,a)
$$

$$
\mu(x,a) = \mu^+(x,a,x') + \mu^-(x,a,x')
$$

# Lagrangian Form



**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



*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.

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).

# UC-O-REPS

**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 $\forall(x,a,x')$:

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

**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 $\forall(x, a)$:

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