1 Outline

In this lecture, we cover

- finite-horizon Markov decision processes (MDPs),
- an online algorithm for learning finite-horizon MDPs.

2 Finite-Horizon MDPs with Advesarial Losses

A finite-horizon MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, H, \rho, \mathbb{P}, \ell)$ where

- $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$,
- $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$,
- H is the horizon length,
- $\rho \in \Delta(\mathcal{S}) := \{\rho \in [0,1]^S : \mathbf{1}^\top \rho = 1\}$ is the initial state distribution,
- $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^{H-1}$ where $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability at step h, i.e., $\mathbb{P}_h(s' \mid s, a)$ is the probability of transitioning to state s' when taking action a in state s at step h,
- $\ell = \{\ell_h\}_{h=1}^{H}$ where $\ell_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the loss function at step h, i.e., $\ell_h(s,a)$ is the loss incurred when taking action a in state s at step h.

In this lecture, we consider an online learning problem setting due to Zimin and Neu [ZN13] described as follows. The learning proceeds in episodes $k = 1, 2, \ldots, K$. While the initial state distrbituion ρ and the transition probabilities $\mathbb{P}$ are fixed and known to the learner across episodes, the environment chooses the loss functions $\ell^k = \{\ell_h^k\}_{h=1}^H$ adversarially. The interaction between the learner and the environment is as follows. In each episode k,

- the learner chooses a policy $\pi^k = (\pi_1^k, \ldots, \pi_H^k)$ where $\pi_h^k : \mathcal{S} \to \Delta(\mathcal{A})$ maps states to distributions over actions, i.e., at step h, the learner takes action a_h according to $a_h \sim \pi_h^k(\cdot \mid s_h)$,
- the environment chooses loss functions $\ell_h^k : \mathcal{S} \times \mathcal{A} \to [0,1]$ for each step $h = 1, \ldots, H$,
- the learner interacts with the MDP using policy π^k and observes the trajectory

$$(s_1^k, a_1^k, s_2^k, a_2^k, \ldots, s_H^k, a_H^k)$$

  where $s_1^k \sim \rho(\cdot)$ is the initial state and $a_h^k \sim \pi_h^k(\cdot \mid s_h^k)$, $s_{h+1}^k \sim \mathbb{P}(\cdot \mid s_h^k, a_h^k)$ for $h = 1, \ldots, H$,
- the learner incurs cumulative loss $\sum_{h=1}^H \ell_h^k(s_h^k, a_h^k)$ in episode k.

Given a policy π, the expected cumulative for episode k is given by

$$L^k(\pi) = \mathbb{E}\left[\sum_{h=1}^{H} \ell_h^k(s_h, a_h) \mid \pi, \mathbb{P}, \rho\right],$$

where the expectation is over the randomness in the trajectory induced by policy π and the MDP dynamics. The goal of the learner is to minimize the regret defined as

$$\text{Regret}(K) = \sum_{k=1}^{K} L^k(\pi^k) - \min_{\pi} \sum_{k=1}^{K} L^k(\pi),$$

where the minimum is over all policies π. In fact, the problem of minimizing the regret can be cast as an instance of online linear optimization over the space of *occupancy measures*. The occupancy measure corresponding to a policy π is defined as

$$q^\pi(s, a, h) = \mathbb{P}\left[s_h = s, a_h = a \mid \pi, \mathbb{P}, \rho\right],$$

i.e., $q^\pi(s, a, h)$ is the probability of being in state s and taking action a at step h when following policy π in the MDP. In general, an accupancy measure $q \in \mathcal{S} \times \mathcal{A} \times [H]$ satisfies the following constraints:

- (initial distribution) for all $s \in \mathcal{S}$,

$$\sum_{a \in \mathcal{A}} q(s, a, 1) = \rho(s),$$

- (transition) for all $s' \in \mathcal{S}$ and $h = 1, \ldots, H-1$,

$$\sum_{a \in \mathcal{A}} q(s', a, h+1) = \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mathbb{P}_h(s' \mid s, a) q(s, a, h),$$

- (probability measure) for all $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$,

$$q(s, a, h) \geq 0,$$

  and for all $h \in [H]$,

$$\sum_{s \in \mathcal{S}, a \in \mathcal{A}} q(s, a, h) = 1.$$

An occupancy measure q satisfying the above constraints corresponds to a valid policy $\pi = (\pi_1, \ldots, \pi_H)$ via

$$\pi_h(a \mid s) = \frac{q(s, a, h)}{\sum_{a' \in \mathcal{A}} q(s, a', h)}$$

for $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$. In fact, there is a one-to-one correspondence between policies and occupancy measures satisfying the above constraints. Moreover, the expected cumulative loss for episode k can be expressed as a linear function of the occupancy measure:

$$L^k(\pi) = \sum_{h=1}^{H} \sum_{s \in \mathcal{S}, a \in \mathcal{A}} q^\pi(s, a, h) \ell_h^k(s, a).$$

Representing the occupancy measure q^π as a vector in $\mathbb{R}^{S\times A\times H}$ and the loss functions ℓ^k as a vector in $\mathbb{R}^{S\times A\times H}$, we have

$$L^k(\pi) = {q^\pi}^\top \ell^k.$$

Therefore, the regret minimization problem can be cast as an instance of online linear optimization over the set of occupancy measures, i.e.,

$$\text{Regret}(K) = \sum_{k=1}^K {q^{\pi^k}}^\top \ell^k - \min_{q\in\mathcal{Q}} \sum_{k=1}^K q^\top \ell^k = \max_{q\in\mathcal{Q}} \sum_{k=1}^K \langle q^{\pi^k} - q, \ell^k\rangle,$$

where $\mathcal{Q}$ is the set of all occupancy measures satisfying the constraints above.

3 An OMD-based Algorithm for Learning Online MDPs

In the previous section, we have observed that the problem of computing policies for minimizing the regret in online finite-horizon MDPs can be cast as an instance of online linear optimization over the set of occupancy measures. Based on this observation, we may design an online learning algorithm based on online mirror descent (OMD). In particular, we choose the (unnormalized) negative Shannon entropy as the mirror map:

$$\psi(q) = \sum_{s\in\mathcal{S},a\in\mathcal{A},h\in[H]} q(s,a,h)\log q(s,a,h) - \sum_{s\in\mathcal{S},a\in\mathcal{A},h\in[H]} q(s,a,h).$$

The corresponding Bregman divergence is the (unnormalized) Kullback-Leibler (KL) divergence:

$$D_\psi(q,q') = \sum_{s\in\mathcal{S},a\in\mathcal{A},h\in[H]} q(s,a,h)\log\frac{q(s,a,h)}{q'(s,a,h)} - \sum_{s\in\mathcal{S},a\in\mathcal{A},h\in[H]} \big(q(s,a,h) - q'(s,a,h)\big).$$

Next we describe the OMD-based algorithm for learning online finite-horizon MDPs. To initialize the algorithm, for episode 1, we take

$$\tilde{q}^1(s,a,h) = \frac{1}{SA}$$

for all $s\in\mathcal{S}, a\in\mathcal{A}, h\in[H]$. Then we compute

$$q^{\pi^1} = \text{argmin}_{q\in\mathcal{Q}} D_\psi(q,\tilde{q}^1),$$

which corresponds to a policy π^1 for episode 1.

Lemma 24.1. *For any occupancy measure $q\in\mathcal{Q}$, it holds that*

$$D_\psi(q,\tilde{q}^1) \le H\log(SA).$$

Proof. Note that

$$D_\psi(q,\tilde{q}^1) = \sum_{s\in\mathcal{S},a\in\mathcal{A},h\in[H]} q(s,a,h)\log\left(q(s,a,h)SA\right) \le \sum_{s\in\mathcal{S},a\in\mathcal{A},h\in[H]} q(s,a,h)\log(SA) = H\log(SA)$$

where the equualities hold because

$$\sum_{s\in\mathcal{S},a\in\mathcal{A},h\in[H]} q(s,a,h) = H$$

for any occupancy measure $q\in\mathcal{Q}$. □

Suppose that we have finished episode k with policy π^k and occupancy measure q^{π^k}. We now describe how to compute the policy π^{k+1} for episode $k+1$. Suppose that for episode k, we have obtained the trajectory

$$(s_1^k, a_1^k, s_2^k, a_2^k, \ldots, s_H^k, a_H^k)$$

and incurred the losses

$$(\ell_1^k(s_1^k, a_1^k), \ldots, \ell_H^k(s_H^k, a_H^k)).$$

We then form an unbiased estimate of the loss function ℓ^k as follows:

$$\hat{\ell}_h^k(s,a) = \begin{cases} \frac{\ell_h^k(s,a)}{q^{\pi^k}(s,a,h)} & \text{if } s = s_h^k, a = a_h^k, \\ 0 & \text{otherwise,} \end{cases}$$

for $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$. Note that

$$q^{\pi^k}(s,a,h) = \mathbb{P}[s_h^k = s, a_h^k = a \mid \pi^k, \mathbb{P}, \rho].$$

This implies that for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$,

$$\begin{aligned} \mathbb{E}\left[\hat{\ell}_h^k(s,a) \mid \pi^k, \mathbb{P}, \rho\right] &= \mathbb{P}[s_h^k = s, a_h^k = a \mid \pi^k, \mathbb{P}, \rho] \cdot \frac{\ell_h^k(s,a)}{q^{\pi^k}(s,a,h)} \\ &= q^{\pi^k}(s,a,h) \cdot \frac{\ell_h^k(s,a)}{q^{\pi^k}(s,a,h)} \\ &= \ell_h^k(s,a). \end{aligned}$$

Then, we perform the OMD update:

$$\tilde{q}^{k+1} = \operatorname{argmin}_{q \in \mathbb{R}^{S \times A \times H}} \left\{ \eta \langle q, \hat{\ell}^k \rangle + D_\psi(q, q^{\pi^k}) \right\},$$

where $\eta > 0$ is the step size. The solution to the above optimization problem can be computed in closed form:

$$\tilde{q}^{k+1}(s,a,h) = q^{\pi^k}(s,a,h) \exp(-\eta \hat{\ell}_h^k(s,a))$$

for $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$. Finally, we need to project $\tilde{q}^{k+1}$ back to the set of occupancy measures $\mathcal{Q}$:

$$q^{\pi^{k+1}} = \arg\min_{q \in \mathcal{Q}} D_\psi(q, \tilde{q}^{k+1}).$$

The complete algorithm is summarized in Algorithm 1.

We may prove the following lemma, due to our OMD update rule.

Lemma 24.2 (Lemma 13 in [Rak09]). *It holds that for any occupancy measure $q \in \mathcal{Q}$,*

$$\sum_{k=1}^K \langle q^{\pi^k} - q, \hat{\ell}^k \rangle \le \sum_{k=1}^K \langle q^{\pi^k} - \tilde{q}^{k+1}, \hat{\ell}^k \rangle + \frac{1}{\eta} D_\psi(q, \tilde{q}^1).$$

Based on this lemma, we can further prove the following regret bound for Algorithm 1.

Theorem 24.3 ([ZN13]). *Setting the step size*

$$\eta = \sqrt{\log(SA)/(SAK)},$$

the regret of Algorithm 1 satisfies

$$\mathbb{E}\left[\text{Regret}(K)\right] \le 2H\sqrt{SAK \log(SA)}.$$

Algorithm 1 OMD-based Algorithm for Online Finite-Horizon MDPs

Input: step size $\eta > 0$, initial state distribution ρ, transition probabilities $\mathbb{P}$
Initialize occupancy measure for step 1: $\tilde{q}^1(s,a,h) = 1/(SA)$ for $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$
Take $q^{\pi^1} = \operatorname{argmin}_{q \in \mathcal{Q}} D_\psi(q, \tilde{q}^1)$
for episode $k = 1$ to K **do**
    Derive policy π^k from occupancy measure q^{π^k}
    Interact with the MDP using policy π^k and observe trajectory $(s_1^k, a_1^k, \ldots, s_H^k, a_H^k)$
    Observe losses $(\ell_1^k(s_1^k, a_1^k), \ldots, \ell_H^k(s_H^k, a_H^k))$
    Form unbiased estimate of loss function $\hat{\ell}^k$
    Perform OMD update to obtain $\tilde{q}^{k+1}$
    Project $\tilde{q}^{k+1}$ back to set of occupancy measures to obtain $q^{\pi^{k+1}}$
end for

Proof. Recall that

$$\tilde{q}^{k+1}(s,a,h) = q^{\pi^k}(s,a,h)\exp(-\eta\hat{\ell}_h^k(s,a)).$$

Since $\exp(x) \geq 1 + x$, we have

$$\tilde{q}^{k+1}(s,a,h) \geq q^{\pi^k}(s,a,h)(1 - \eta\hat{\ell}_h^k(s,a)).$$

This implies that

$$\begin{aligned}
\langle q^{\pi^k} - \tilde{q}^{k+1}, \hat{\ell}^k \rangle &\leq \eta \sum_{s\in\mathcal{S}, a\in\mathcal{A}, h\in[H]} q^{\pi^k}(s,a,h)(\hat{\ell}_h^k(s,a))^2 \\
&\leq \eta \sum_{s\in\mathcal{S}, a\in\mathcal{A}, h\in[H]} q^{\pi^k}(s,a,h)\frac{\ell_h^k(s,a)}{q^{\pi^k}(s,a,h)}\hat{\ell}_h^k(s,a) \\
&\leq \eta \sum_{s\in\mathcal{S}, a\in\mathcal{A}, h\in[H]} \hat{\ell}_h^k(s,a)
\end{aligned}$$

where the second inequality holds because

$$\hat{\ell}_h^k(s,a) \leq \frac{\ell_h^k(s,a)}{q^{\pi^k}(s,a,h)}$$

and the third inequality follows from the fact that $\ell_h^k(s,a) \leq 1$. Since $\hat{\ell}_h^k$ is an unbiased estimator of ℓ_h^k, we have that

$$\mathbb{E}\left[\langle q^{\pi^k} - \tilde{q}^{k+1}, \hat{\ell}^k \rangle \mid \pi^k, \mathbb{P}, \rho\right] \leq \eta \sum_{s\in\mathcal{S}, a\in\mathcal{A}, h\in[H]} \ell_h^k(s,a) \leq \eta HSA.$$

Then it follows that

$$\sum_{k=1}^K \mathbb{E}\left[\langle q^{\pi^k} - q, \hat{\ell}^k \rangle \mid \pi^k, \mathbb{P}, \rho\right] \leq \eta HSAK + \frac{1}{\eta} D_\psi(q, \tilde{q}^1) \leq \eta HSAK + \frac{H\log(SA)}{\eta}$$

where the first inequality is from Lemma 24.2 and the second inequality is from Lemma 24.1. Therefore, we have

$$\mathbb{E}\left[\sum_{k=1}^K \langle q^{\pi^k} - q, \hat{\ell}^k \rangle\right] \leq 2H\sqrt{SAK\log(SA)},$$

as required. □

References

[Rak09] Alexander Rakhlin. Lecture notes on online learning. Technical report, MIT, 2009. 24.2

[ZN13] Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 2, 24.3