1 Outline

In this lecture, we cover

- finite-horizon Markov decision processes (MDPs) with unknown transitions,
- occupancy measure-based model estimation,
- an OMD-based algorithm for learning finite-horizon MDPs with unknown transitions, and
- regret analysis of the OMD-based algorithm.

2 Finite-Horizon Adversarial MDPs with Unknown Transitions

As in the previous lecture, we consider finite-horizon Markov decision processes (MDPs), given by a tuple $(\mathcal{S}, \mathcal{A}, H, \rho, P, \ell)$ where $\mathcal{S}$ is the finite state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the finite action space with $|\mathcal{A}| = A$, H is the finite horizon, $P_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the transition function at step $h \in [H-1]$, and ρ is the initial distribution of the states. Here, $P_h(s' \mid s,a)$ is the probability of transitioning to state s' from state s when the chosen action is a at step $h \in [H-1]$. Equivalently, we may define a single *non-stationary* transition function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \to [0,1]$ with $P(s' \mid s,a,h) = P_h(s' \mid s,a)$ and $P(s' \mid s,a,H) = \rho(s')$ for $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$.

Before an episode begins, the decision-maker prepares a *stochastic policy* $\pi : \mathcal{S} \times [H] \times \mathcal{A} \to [0,1]$ where $\pi(a \mid s,h)$ is the probability of selecting action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at step h. Here, π can be viewed as a *non-stationary policy* as it may change over the horizon, and this is due to the non-stationarity of the transition function P over steps $h \in [H]$. Given a policy π^k for episode $k \in [K]$, the MDP proceeds with trajectory $\{s_h^{P,\pi^k}, a_h^{P,\pi^k}\}_{h\in[H]}$ generated by P.

The loss function of episode $k \in [K]$ is given by $\ell^k : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$, i.e., choosing action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$ at step h generates a loss $\ell^k(s,a,h)$. Here, function ℓ^k is non-stationary over $h \in [H]$. We assume that ℓ^k for episode $k \in [K]$ is chosen *adversarially* after the decision-maker selects the policy π^k for episode k.

Following the setup of [RM19], we assume that ρ and $\{P_h\}_{h=1}^{H-1}$ are *unknown*, in which case P is *unknown*. Moreover, as in the previous lecture and in [ZN13, RM19], we consider the *full-information feedback* setting where at the end of episode k, the decision-maker observes the entire loss function ℓ^k. The goal of the decision-maker is to minimize the regret defined as

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

where the minimum is over all policies π and

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

where the expectation is over the randomness in the trajectory induced by policy π and the MDP dynamics.

3 Occupancy Measures for the Unknown Transition Model

To deal with the unknown transition function P, we need to take a slightly different version of *occupancy measures*. Specifically, we consider the following formulation of occupancy measures [RM19, CKKM21]. Given a policy π and a transition function P, let $\bar{q}^{P,\pi} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \to [0,1]$ be defined as

$$\bar{q}^{P,\pi}(s,a,s',h) = \mathbb{P}\left[s_h^{P,\pi} = s,\ a_h^{P,\pi} = a,\ s_{h+1}^{P,\pi} = s' \mid \pi, P\right] \tag{25.1}$$

for $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Note that any $\bar{q}$ defined as in (25.1) has the following properties:

$$\sum_{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \bar{q}(s,a,s',h) = 1, \quad h \in [H] \tag{C1}$$

$$\sum_{(s',a) \in \mathcal{S} \times \mathcal{A}} \bar{q}(s,a,s',h) = \sum_{(s',a) \in \mathcal{S} \times \mathcal{A}} \bar{q}(s',a,s,h-1), \quad s \in \mathcal{S},\ h = 2,\ldots,H. \tag{C2}$$

The occupancy measure $q^{P,\pi} : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$ associated with policy π and transition function P is defined as

$$q^{P,\pi}(s,a,h) = \sum_{s' \in \mathcal{S}} \bar{q}^{P,\pi}(s,a,s',h). \tag{C3}$$

Then it follows that

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

Hence, if a policy π is chosen, then the occupancy measure for a finite-horizon MDP with transition function P is determined. Conversely, any $q \in \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$ with $\bar{q} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \to [0,1]$ satisfying (C1), (C2), (C3) induces a transition function P^q and a policy π^q given as follows:

$$P^q(s' \mid s,a,h) = \frac{\bar{q}(s,a,s',h)}{\sum_{s'' \in \mathcal{S}} \bar{q}(s,a,s'',h)}, \quad \pi^q(a \mid s,h) = \frac{q(s,a,h)}{\sum_{b \in \mathcal{A}} q(s,b,h)}. \tag{25.2}$$

Lemma 25.1. *Let $q : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$. Then q is a valid occupancy measure that induces transition function P if and only if there exists $\bar{q} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H] \to [0,1]$ that satisfies (C1), (C2), (C3), and $P^q = P$.*

Therefore, there is a one-to-one correspondence between the set of policies and the set of occupancy measures that give rise to transition function P. Moreover, the cumulative loss for episode k under loss function ℓ^k, policy π^k, and transition function P can be written in terms of occupancy measure q^{P,π^k} associated with π^k and P.

$$\begin{aligned}
\mathbb{E}\left[\sum_{h=1}^{H} \ell^k\left(s_h^{\pi^k}, a_h^{\pi^k}, h\right) \mid \pi^k, P\right] &= \sum_{(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]} q^{P,\pi^k}(s,a,h)\, \ell^k(s,a,h) \\
&= \sum_{(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]} \bar{q}^{P,\pi^k}\left(s,a,s',h\right) \ell^k(s,a,h).
\end{aligned}$$

We may express occupancy measure $\bar{q}^{P,\pi^k}$ as an $(S \times A \times S \times H)$-dimensional vector $\bar{\boldsymbol{q}}^{\boldsymbol{P},\boldsymbol{\pi}^{\boldsymbol{k}}}$ whose entries are given by $\bar{q}^{P,\pi^k}(s,a,s',h)$ for $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Moreover, we define vector $\boldsymbol{\ell}^{\boldsymbol{k}} \in [0,1]^{S \times A \times S \times H}$ whose entry corresponding to $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$ is given by $\ell^k(s,a,h)$.

Then the right-hand side of the above equation is equal to $\langle \boldsymbol{\ell^k}, \bar{\boldsymbol{q}}^{\boldsymbol{P},\boldsymbol{\pi^k}} \rangle$, the inner product of $\boldsymbol{\ell^k}$ and $\bar{\boldsymbol{q}}^{\boldsymbol{P},\boldsymbol{\pi^k}}$. Then the regret after K episodes can be written as

$$\mathrm{Regret}(K) = \sum_{k=1}^{K} \langle \boldsymbol{\ell^k}, \bar{\boldsymbol{q}}^{\boldsymbol{P},\boldsymbol{\pi^k}} \rangle - \min_{\bar{\boldsymbol{q}} \in \Delta(P)} \sum_{k=1}^{K} \langle \boldsymbol{\ell^k}, \bar{\boldsymbol{q}} \rangle,$$

where

$$\Delta(P) = \left\{ \bar{\boldsymbol{q}} \in [0,1]^{S \times A \times S \times H} : \bar{\boldsymbol{q}} \text{ satisfies (C1), (C2), (C3), } P^q = P \right\}$$

where P^q is defined as in (25.2). Recall that $\Delta(P)$ is a polytope.

4 Estimating the Unknown Transition Model

In contrast to the standard online optimization problems, the feasible set $\Delta(P)$ is unknown as we do not have access to the true transition function P. To remedy this issue, we obtain empirical transition functions $\bar{P}_1, \ldots, \bar{P}_K$ to estimate the true transition function P, based on which we construct *relaxations* $\Delta(P, 1), \ldots, \Delta(P, K)$ of the feasible set $\Delta(P)$ by building *confidence sets* for P.

To construct confidence sets for estimating the non-stationary transition function P, we extend the framework of [JJL+20] developed for the finite-horizon setting and that of [CL21] for the stochastic shortest path setting to our finite-horizon non-stationary setting.

We maintain counters to keep track of the number of visits to each tuple (s, a, h) and tuple (s, a, s', h). For each $k \in [K]$, we define $N_k(s, a, h)$ and $M_k(s, a, s', h)$ as the number of visits to tuple (s, a, h) and the number of visits to tuple (s, a, s', h) up to the first $k - 1$ episodes, respectively, for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$. Given $N_k(s, a, h)$ and $M_k(s, a, s', h)$, we define the empirical transition function $\bar{P}_k$ for episode k as

$$\bar{P}_k(s' \mid s, a, h) = \frac{M_k(s, a, s', h)}{\max\{1, N_k(s, a, h)\}}. \tag{25.3}$$

Next, for some confidence parameter $\delta \in (0, 1)$, we define the confidence radius $\epsilon_k(s' \mid s, a, h)$ for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$ and $k \in [K]$ as

$$\epsilon_k(s' \mid s, a, h) = 2\sqrt{\frac{\bar{P}_k(s' \mid s, a, h) \log(HSAK/\delta)}{\max\{1, N_k(s, a, h) - 1\}}} + \frac{14 \log(HSAK/\delta)}{3 \max\{1, N_k(s, a, h) - 1\}}. \tag{25.4}$$

Based on the empirical transition function and the radius, we define the confidence set $\mathcal{P}_k$ for episode k as

$$\mathcal{P}_k = \left\{ \widehat{P} : \left| \widehat{P}(s' \mid s, a, h) - \bar{P}_k(s' \mid s, a, h) \right| \leq \epsilon_k(s' \mid s, a, h) \quad \forall (s, a, s', h) \right\}. \tag{25.5}$$

Then, by the empirical Bernstein inequality due to [MP09], we show the following.

Lemma 25.2. *With probability at least $1 - 4\delta$, the true transition function P is contained in the confidence set $\mathcal{P}_k$ for every episode $t \in [T]$.*

For episode $k \in [K]$, we define $\Delta(P, k)$ as

$$\Delta(P, k) = \left\{ \bar{\boldsymbol{q}} \in [0,1]^{S \times A \times S \times H} : \exists \bar{\boldsymbol{q}} \text{ satisfies (C1), (C2), (C3), } P^q \in \mathcal{P}_k \right\}. \tag{25.6}$$

The following is a direct consequence of Theorem 25.2.

Lemma 25.3. *With probability at least $1 - 4\delta$, $\Delta(P) \subseteq \Delta(P, k)$ for every episode $k \in [K]$.*

5 OMD-Based Algorithm

We present Algorithm 1, an online mirror descent (OMD)-based algorithm for learning finite-horizon MDPs with unknown transition functions. The algorithm maintains counters to keep track of the

---

Algorithm 1 Online Mirror Descent for Learning Episodic Finite-horizon MDPs

---

Initialize: episode counter $k = 1$, counters $N(s, a, h) = 0$ and $M(s, a, s', h) = 0$ for $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$, and step size $\eta > 0$.

(0. Initialization)

Set the initial occupancy measure $\widehat{q}^1$ as $\widehat{q}^1(s, a, s', h) = 1/(SAS)$ for $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$.

Deduce initial policy π^1 associated with $\widehat{q}^1$ as in (25.2).

for $k = 1, \dots, K$ **do**

  (1. Policy execution)

  Sample state s_1 from distribution $p(\cdot)$.

  for $h = 1, \dots, H$ **do**

    Sample a_h from $\pi^k(\cdot \mid s_h, h)$.

    Observe loss function ℓ^k.

    Observe s_{h+1} determined by $P(\cdot \mid s_h, a_h, h)$.

    Update counters $N(s, a, h) \leftarrow N(s, a, h) + 1$ and $M(s, a, s', h) \leftarrow M(s, a, s', h) + 1$.

  end for

  (2. Confidence set construction)

  Set counters $N_k \leftarrow N$ and $M_k \leftarrow M$.

  Compute $\bar{P}_k$, ϵ_k, and $\mathcal{P}_k$ as in (25.3), (25.4), and (25.5).

  Take $\Delta(P, k)$ defined in (25.6).

  (3. Policy update)

  Compute $\tilde{q}^{k+1}$ as

$$\boldsymbol{\tilde{q}^{k+1}} \in \operatorname*{argmin}_{\boldsymbol{\bar{q}} \in \mathbb{R}^{S \times A \times S \times H}} \left\{ \eta \langle \boldsymbol{\ell^k}, \boldsymbol{\bar{q}} \rangle + D_\psi(\bar{q}, \widehat{q}^k) \right\}.$$

  Deduce policy π_{k+1} associated with $\widehat{q}^{k+1}$ as in (25.2) where

$$\boldsymbol{\widehat{q}^{k+1}} \in \operatorname*{argmin}_{\boldsymbol{\bar{q}} \in \Delta(P,k)} \left\{ D_\psi(\bar{q}, \tilde{q}^{k+1}) \right\}.$$

end for

---

number of visits to each tuple (s, a, h) and tuple (s, a, s', h), and at the beginning of each episode $k \in [K]$, it constructs the empirical transition function $\bar{P}_k$ as in (25.3) and the confidence set $\mathcal{P}_k$ as in (25.5) based on the counters. Then, it defines the relaxation $\Delta(P, k)$ of the feasible set $\Delta(P)$ as in (25.6). Next, it updates the occupancy measure $\boldsymbol{\widehat{q}^{k+1}}$ by performing an OMD step over the relaxation $\Delta(P, k)$ using the observed loss function ℓ^k. Finally, it deduces policy π^{k+1} from occupancy measure $\boldsymbol{\widehat{q}^{k+1}}$ as in (25.2).

Under Algorithm 1, regret can be expressed as

$$\text{Regret}(K) = \underbrace{\sum_{k=1}^{K} \langle \boldsymbol{\ell^k}, \boldsymbol{\widehat{q}^k} - \boldsymbol{q^{P,\pi^*}} \rangle}_{\text{(I)}} + \underbrace{\sum_{k=1}^{K} \langle \boldsymbol{\ell^k}, \boldsymbol{q^{P,\pi^k}} - \boldsymbol{\widehat{q}^k} \rangle}_{\text{(II)}}$$

where $\pi^* = \operatorname{argmin}_\pi \sum_{k=1}^{K} L^k(\pi)$ is the best policy in hindsight. The first term can be bounded using the OMD analysis, while the second term can be bounded using the properties of the confidence sets.

5.1 Bounding the Regret Term (I)

By Lemma 25.3, with probability at least $1-4\delta$, we have $\boldsymbol{q^{P,\pi^*}} \in \Delta(P,k)$ for every episode $k \in [K]$. Therefore, we can apply the standard OMD analysis to bound term (I) as follows.

We may argue that based on Lemma 13 in [Rak09],

$$\text{(I)} \le \sum_{k=1}^{K} \langle \boldsymbol{\ell^k}, \boldsymbol{\widehat{q}^k} - \boldsymbol{\tilde{q}^{k+1}} \rangle + \frac{1}{\eta} D_\psi(q, \widehat{q}^1) \le \sum_{k=1}^{K} \langle \boldsymbol{\ell^k}, \boldsymbol{\widehat{q}^k} - \boldsymbol{\tilde{q}^{k+1}} \rangle + \frac{H \log(SAS)}{\eta}.$$

Moreover, note that

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

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

$$\tilde{q}^{k+1}(s,a,s',h) \ge \widehat{q}^k(s,a,s',h)(1-\eta\ell^k(s,a,h)).$$

This implies that

$$\langle \boldsymbol{\widehat{q}^k} - \boldsymbol{\tilde{q}^{k+1}}, \boldsymbol{\ell^k} \rangle \le \eta \sum_{s,s' \in \mathcal{S}, a \in \mathcal{A}, h \in [H]} \widehat{q}^k(s,a,s',h)(\ell^k(s,a,h))^2 \le \eta H.$$

Therefore, it follows that

$$\text{(I)} \le \eta H K + \frac{H \log(SAS)}{\eta}.$$

5.2 Bounding the Regret Term (II)

The regret term (II) comes from the error in estimating the true transition function P. We can bound this term using the properties of the confidence sets $\mathcal{P}_k$ defined in (25.5). We may prove that with probability at least $1-4\delta$,

$$\text{(II)} = O\left(\left(H^{3/2} S\sqrt{AK} + H^{5/2} S^2 A\right)\left(\log \frac{HSAK}{\delta}\right)^2\right).$$

5.3 Regret Bound

Combining the bounds for terms (I) and (II), we obtain the following regret bound for Algorithm 1.

Theorem 25.4. *Let $\delta \in (0,1)$ and set the step size $\eta = \sqrt{\log(SAS)/K}$. Then, with probability at least $1-4\delta$, the regret of Algorithm 1 after K episodes is bounded as*

$$\text{Regret}(K) = O\left(\left(H^{3/2} S\sqrt{AK} + H^{5/2} S^2 A\right)\left(\log \frac{HSAK}{\delta}\right)^2\right).$$

References

[CKKM21] Alon Cohen, Haim Kaplan, Tomer Koren, and Yishay Mansour. Online markov decision processes with aggregate bandit feedback. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1301–1329. PMLR, 15–19 Aug 2021. 3

[CL21] Liyu Chen and Haipeng Luo. Finding the stochastic shortest path with low regret: the adversarial cost and unknown transition case. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1651–1660. PMLR, 18–24 Jul 2021. 4

[JJL+20] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4860–4869. PMLR, 13–18 Jul 2020. 4

[MP09] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. 4

[Rak09] Alexander Rakhlin. Lecture notes on online learning. Technical report, MIT, 2009. 5.1

[RM19] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5478–5486. PMLR, 09–15 Jun 2019. 2, 3

[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