CSCI 659 Lecture 10

Fall 2022

Instructor: Haipeng Luo

1 Online Reinforcement Learning

In this lecture, we discuss how online learning has been extended to and combined with another important and popular area of machine learning: reinforcement learning (RL). RL can be seen as a generalization of MAB by incorporating the concept of *states* and the transition among them driven by the learner's actions. The reward/loss of the learner depends on not only her chosen action, but also the state that she is currently in. This allows RL to capture many more exciting applications (such as playing games or controlling robots), but also leads to one key challenge compared to other problems discussed so far: long-term planing, that is, a smart learner should not focus too much on the short-term reward generated by the current action, but should instead plan ahead in order to reach a region of states with high rewards in the future.

Formally, RL is often captured by a *Markov Decision Process* (MDP), which is a Markov chain driven by a learner's decisions. There are many variants of MDPs. Here, we focus on a variant called finite-horizon tabular MDPs, defined via a tuple (X, A, P, ℓ):

1. X is a finite state space that can be partitioned into H layers $X_1, \ldots, X_H$ for some fixed parameter H, where X_1 contains only an initial state x_{init}.
2. A is a finite action space containing the available actions for the learner at each state.
3. P is a transition function so that $P(\cdot|x, a)$ (for $x \notin X_H$) is the distribution of the next state after taking action a at state x (note the Markovian assumption here: the next state only depends on the current state and action, but not the past). It is assumed that $P(x'|x, a)$ is not zero only when $x \in X_h$ and $x' \in X_{h+1}$ for some $h < H$, that is, transition can only happen between consecutive layers.
4. Finally, $\ell : X \times A \to [0, 1]$ is a loss function with $\ell(x, a)$ specifying the loss of taking action a at state x.

A policy π is a mapping that maps each state in X to a distribution $\pi(\cdot|x)$ over the actions in A. Starting from the initial state x_{init}, if the learner acts according to π (that is, take an action sampled from $\pi(\cdot|x)$ when at state x), then an *episode* $x_1, a_1, \ldots, x_H, a_H$ of H steps is generated with $x_1 = x_{\text{init}}$, $a_h \sim \pi(\cdot|x_h)$, and $x_{h+1} \sim P(\cdot|x_h, a_h)$. The expected total loss of the learner suffered in this episode is denoted by $V^\pi(x_{\text{init}}; \ell) = \mathbb{E}[\sum_{h=1}^{H} \ell(x_h, a_h)]$. More generally, we denote the expected loss starting from any state x and following π afterwards as $V^\pi(x; \ell)$, which can be written $\mathbb{E}_{a\sim\pi(\cdot|x)}[Q^\pi(x, a; \ell)]$ where

$$Q^\pi(x, a; \ell) = \ell(x, a) + \mathbf{1}\{x \notin X_H\}\mathbb{E}_{x'\sim P(\cdot|x,a)}[V^\pi(x'; \ell)] \tag{1}$$

is the so-called Q-function (specifying the expected loss starting from x, taking action a, and following π afterwards) and the above is known as the *Bellman equation*. While this definition looks recursive, it is really not since the transition is always from one layer to the next one, making $V^\pi(x; \ell)$ well-defined in a backward manner. Traditional RL concerns about finding the (approximately) optimal policy that minimizes $V^\pi(x_{\text{init}}; \ell)$ when P (and sometimes ℓ as well) is unknown, using as few episodes as possible.

We emphasize that assuming a layer structure is really without loss of generality since such a structure can always be created artificially by considering an expanded state space $X \times [H]$ where each state is duplicated H times, each with a different step index. The real restriction of a finite-horizon MDP is that it only considers interaction of a fixed (H) number of steps, after which the learner is reset to the initial state, making any "wrong" actions always recoverable in a sense. There are many other variants of MDP that address this limitation, such as the stochastic shortest path model where the interaction stops only when a certain goal state is reached, or the infinite-horizon model where the interaction never stops and one cares about either the average loss of the learner or her discounted total loss.

Question 1. *Assuming a fixed initial state is also in a sense without loss of generality, and the same model can deal with the case where the initial state is drawn from a fixed distribution. Why?*

Online RL. In online RL, a topic that receives increasing interest in more recent RL studies, the learner deals with a potentially changing MDP with the goal of minimizing regret. Specifically, consider a finite-horizon MDP with a fixed state space X, a fixed action space A, and a fixed transition function P. The learner interacts with the MDP through T episodes, where at each episode $t = 1, \dots, T$:

1. the learner decides a policy π_t while an oblivious adversary decides a loss function $\ell_t : X \times A \to [0, 1]$;
2. the learner starts from the initial state x_{init}, acts according to π_t, and ends at the terminate state x_H after H steps, generating and observing a trajectory $\{(x_{t,h}, a_{t,h}, \ell_t(x_{t,h}, a_{t,h}))\}_{h=1}^{H}$.

Let $V_t^\pi(x)$ be a shorthand for $V^\pi(x; \ell_t)$. The goal of the learner is to minimize her expected regret, defined as

$$\mathbb{E}[\mathcal{R}_T] = \mathbb{E}\left[\sum_{t=1}^{T} V_t^{\pi_t}(x_{\text{init}})\right] - \sum_{t=1}^{T} V_t^{\pi^\star}(x_{\text{init}})$$

where $\pi^\star \in \operatorname{argmin}_\pi \sum_{t=1}^{T} V_t^\pi(x_{\text{init}})$ is the overall optimal policy in hindsight. More concretely, we would like to design an efficient algorithm with (sublinear in T) regret and time/space complexity that are both polynomial in $|X|$, $|A|$, and H.

Note that the feedback on the loss function is bandit-type, that is, the learner does not observe the entire ℓ_t at the end of episode t, but only its value for those visited state-action pairs. In fact, it is clear that adversarial MAB is exactly a special case of this setup with $H = 1$.

One might wonder why we only allow the loss function to be changing over time, but not the transition P as well. It turns out that, if we allow P to also be adversarially chosen, then in the worst case the regret can be as bad as $\Omega(2^H\sqrt{T})$; see [Tian et al., 2021, Lemma 1]. Thus, we assume a fixed P throughout. In fact, for simplicity, we will also by default assume that P is known to the learner in the following discussions and only briefly mention what modifications are needed to handle unknown P. Note that unlike the traditional RL setup where the MDP is fixed and thus knowing P significantly simplifies the problem, online RL is still highly non-trivial even with known P due to the changing loss functions.

2 FTRL over the Occupancy Measure Space

It is not difficult to see that $V_t^\pi(x_{\text{init}})$ is in fact nonconvex in π (try to convince yourself), making it difficult to directly perform FTRL over the policy space. It turns out that, however, it is possible to view this as a linear problem over a different space. Specifically, any policy π induces a corresponding *occupancy measure* $q^\pi : X \times A \to [0, 1]$ such that $q^\pi(x, a)$ is the probability of visiting state-action pair (x, a) when the learner starts from the initial state and acts according to π. By seeing q^π and a loss function ℓ as $|X||A|$-dimensional vectors, we can then write $V^\pi(x_{\text{init}}; \ell)$ as $\langle q^\pi, \ell \rangle$ and thus the expected regret as

$$\mathbb{E}[\mathcal{R}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \langle q_t - q^\star, \ell_t \rangle\right]$$

where $q_t = q^{\pi_t}$ and $q^\star = q^{\pi^\star}$. Importantly, this translates the problem into an online linear optimization problem!

This naturally suggests performing FTRL over the space of occupancy measures to directly come up with q_t. Specifically, let $\Omega = \{q^\pi : \pi \text{ is a policy}\}$ be the space of all valid occupancy measures. While it might not be clear at first glance, this set is in fact a relatively simple polytope with $\mathcal{O}(|X|)$ linear constraints, as shown by the following lemma.

Lemma 1. *The set of valid occupancy measures Ω can be equivalently written as*

$$\Omega = \Bigg\{ q \in [0,1]^{|X|\times|A|} : q(x_{\mathit{init}}) = 1 \text{ and } \sum_{x\in X_h}\sum_{a\in A} q(x,a)P(x'|x,a) = q(x'),\ \forall x' \in X_{h+1} \text{ and } h \in [H-1], \Bigg\}$$

where $q(x)$ is a shorthand for $\sum_{a\in A} q(x,a)$.

Proof. For any π, the induced occupancy measure q^π must belong to this polytope since the first constraint trivially holds, and the second constraint says that the probability of visiting state x' is the sum of probabilities of visiting a state-action pair (x,a) in the last layer and then transiting to x', which is also trivially true by definition.

On the other hand, for some q that belongs to this polytope, it corresponds to the occupancy measure induced by the policy π with $\pi(a|x) \propto q(x,a)$ (if $q(x) \neq 0$, this means $\pi(a|x) = q(x,a)/q(x)$; otherwise, $\pi(\cdot|x)$ can be an arbitrary distribution over A). To verify this, we need to argue $q^\pi(x,a) = q(x,a)$ for any (x,a) pair, which can be done via a simple induction on the layer index: when $x = x_{\text{init}}$, this holds since $q^\pi(x_{\text{init}},a) = \pi(a|x_{\text{init}}) = q(x_{\text{init}},a)/q(x_{\text{init}}) = q(x_{\text{init}},a)$ using the first constraint $q(x_{\text{init}}) = 1$; now assuming $q^\pi(x,a) = q(x,a)$ for any $(x,a) \in X_h \times A$, then for any $(x',a') \in X_{h+1} \times A$, we have

$$\begin{aligned}
q^\pi(x',a') &= \left(\sum_{x\in X_h}\sum_{a\in A} q^\pi(x,a)P(x'|x,a)\right)\pi(a'|x') && \text{(definition of occupancy measure)}\\
&= \left(\sum_{x\in X_h}\sum_{a\in A} q(x,a)P(x'|x,a)\right)\pi(a'|x') && \text{(inductive assumption)}\\
&= q(x')\pi(a'|x') && \text{(second constraint of the polytope)}\\
&= q(x',a'), && \text{(definition of } \pi)
\end{aligned}$$

which finishes the proof. □

Note that the proof also tells us how to extract a policy π from its induced occupancy measure q^π by simply normalizing $q^\pi(x,\cdot)$ at each state x. Thus, after obtaining q_t from FTRL, we can extract the policy π_t in this way. It remains to construct loss estimators for ℓ_t since we only observe its value for H state-action pairs. Due to the similarity with MAB, it is natural to again use the inverse importance weighting idea to construct $\widehat{\ell}_t$ as

$$\widehat{\ell}_t(x,a) = \frac{\ell_t(x,a)}{q_t(x,a)}\mathbf{1}_t(x,a) \tag{2}$$

where $\mathbf{1}_t(x,a)$ is 1 if (x,a) is visited during episode t and 0 otherwise (so there are at most H non-zero entries in $\widehat{\ell}_t$, one for each layer). The following properties are direct generalization of those from MAB.

Lemma 2. *The estimator defined in Eq. (2) satisfies $\mathbb{E}_t[\widehat{\ell}_t(x,a)] = \ell_t(x,a)$ (unbiasedness) and $\mathbb{E}_t[\widehat{\ell}_t(x,a)^2] = \frac{\ell_t(x,a)^2}{q_t(x,a)} \leq \frac{1}{q_t(x,a)}$.*

Proof. Both are based on direct calculations using the fact $\mathbb{E}_t[\mathbf{1}_t(x,a)] = q_t(x,a)$. □

This concludes the design of the algorithm, formally shown below with a general regularizer. To come up with a concrete regularizer, we can again take inspiration from MAB. For example, we can use the generalized Shannon entropy regularizer $\psi(q) = \sum_{(x,a)\in X\times A} q(x,a)\ln q(x,a)$. The resulting algorithm was proposed by [Zimin and Neu, 2013] and ensures the following regret bound.

Algorithm 1: FTRL for Finite-Horizon MDPs

Input: learning rate $\eta > 0$ and a regularizer ψ defined on the occupancy measure space Ω
for $t = 1, \dots, T$ **do**
  compute $q_t = \operatorname{argmin}_{q \in \Omega} \left\langle q, \sum_{s<t} \widehat{\ell}_s \right\rangle + \frac{1}{\eta}\psi(q)$
  extract policy π_t such that $\pi_t(a|x) \propto q_t(x,a)$
  execute policy π_t and observe trajectory $\{(x_{t,h}, a_{t,h}, \ell_t(x_{t,h}, a_{t,h}))\}_{h=1}^{H}$
  construct estimator $\widehat{\ell}_t$ based on Eq. (2)

Theorem 1. *Algorithm 1 with regularizer $\psi(q) = \sum_{(x,a) \in X \times A} q(x,a) \ln q(x,a)$ ensures $\mathbb{E}[\mathcal{R}_T] \le \frac{H \ln(|X||A|)}{\eta} + \eta |X||A|T$, which is $\widetilde{\mathcal{O}}(\sqrt{H|X||A|T})$ after picking the optimal η.*

Proof. By the exact same analysis as Hedge, one can show (we have seen several approaches by now and you should verify this yourself):

$$\sum_{t=1}^{T} \left\langle q_t - q^\star, \widehat{\ell}_t \right\rangle \le \frac{B_\psi}{\eta} + \eta \sum_{t=1}^{T} \sum_{(x,a) \in X \times A} q_t(x,a) \widehat{\ell}_t(x,a)^2$$

where B_ψ is the range of ψ and is at most

$$\max_{q \in \Omega} \sum_{(x,a) \in X \times A} q(x,a) \ln \frac{1}{q(x,a)} = \max_{q \in \Omega} \sum_{h \in [H]} \left(\sum_{(x,a) \in X_h \times A} q(x,a) \ln \frac{1}{q(x,a)} \right)$$
$$\le \sum_{h \in [H]} \ln(|X_h||A|) \le H \ln(|X||A|)$$

where the first inequality is by realizing $\sum_{(x,a) \in X_h \times A} q(x,a) = 1$ (based on the definition of occupancy measure) and treating $\sum_{(x,a) \in X_h \times A} q(x,a) \ln \frac{1}{q(x,a)}$ as the Shannon entropy of a distribution over $|X_h||A|$ elements, which is thus at most $\ln(|X_h||A|)$. The rest of the proof is simply by taking expectation and applying Lemma 2. □

Ignoring logarithmic terms, this regret bound is known to be optimal. When $H = 1$ (and thus $|X| = 1$ as well), the algorithm/regret also recovers Exp3. Moreover, the algorithm can be implemented in time/space polynomial in all the parameters since the FTRL optimization is a convex problem defined over a simple polytope with $\mathcal{O}(|X|)$ constraints.

Question 2. *What happens if we use a direct generalization of Tsallis entropy, such as $\psi(q) = -2 \sum_{(x,a) \in X \times A} \sqrt{q(x,a)}$, as the regularizer?*

Handling Unknown Transition. When P is unknown (the more interesting situation), we need to handle two issues in Algorithm 1: first, the decision set Ω is no longer known ahead of time, and second, the exact loss estimator defined in Eq. (2) is also no longer constructable since q_t, the occupancy measure of π_t, is unknown. To deal with both issues, we need to apply the optimism principle again. Indeed, the challenge coming from unknown transition is very much similar to that in solving stochastic MAB, since both are about adaptively getting samples from fixed unknown distributions. (So interestingly, this problem combines the challenges of both stochastic and adversarial MAB.)

Specifically, at each time t, we build a confidence set $\mathcal{P}_t$ that contains the true transition P with high probability. We omit the concrete form of $\mathcal{P}_t$ but point out that it is based on standard concentration inequalities and in spirit similar to the confidence set of the loss of each arm in stochastic MAB. Let $q^{\pi, \widehat{P}}$ be the occupancy measure induced by policy π for an MDP with transition function $\widehat{P}$ (so the earlier notation q^π is simply a shorthand for $q^{\pi, P}$). Then, we make the following two modifications to Algorithm 1:

- replace the unknown decision set Ω in FTRL with $\Omega_t = \{q^{\pi, \widehat{P}} : \widehat{P} \in \mathcal{P}_t, \pi \text{ is a policy}\}$, the set of all plausible occupancy measures, which incorporates optimism into the FTRL update.

- similarly, when constructing the loss estimator, replace the unknown $q_t(x,a) = q^{\pi_t,P}(x,a)$ with $\max_{\widehat{P}\in\mathcal{P}_t} q^{\pi_t,\widehat{P}}(x,a)$, the largest plausible probability of visiting (x,a), thus incorporating optimism into the loss estimators as well.

It turns out that the resulting algorithm is still implementable in polynomial time. By carefully analyzing the error coming from the transition estimation, Jin et al. [2020] showed that this algorithm ensures $\mathcal{O}(H|X|\sqrt{|A|T})$ regret, which only exhibits a small gap compared to the best known lower bound $\Omega(H\sqrt{|X||A|T})$. Closing this gap (with any algorithm, even inefficient ones) is still open.

3 Policy Optimization Methods

Earlier, we mentioned that $V^\pi(x_{\text{init}};\ell)$ is nonconvex in π, but this in fact does not immediately rule out the possibility of optimizing directly over the policy space. Such approaches are generally called policy optimization methods. Here, we introduce one such method, which in a sense decomposes the online RL problem into $|X|$ different instances of the adversarial MAB problem, one for each state. This is enabled by the following simple yet fundamental *performance difference lemma*.

Lemma 3. *For any loss function ℓ and any two policies π and $\pi^\star$, the difference of their expected total losses starting from the initial state can be decomposed as*

$$V^\pi(x_{init};\ell) - V^{\pi^\star}(x_{init};\ell) = \sum_{x\in X} q^\star(x) \sum_{a\in A} (\pi(a|x) - \pi^\star(a|x)) Q^\pi(x,a;\ell)$$

where $q^\star$ is a shorthand for $q^{\pi^\star}$.

Proof. This is by direct calculations (below, $\sum_x$ means $\sum_{x\in X}$ and $\sum_{x,a}$ means $\sum_{x\in X, a\in A}$):

$$\begin{aligned}
&V^\pi(x_{\text{init}};\ell) - V^{\pi^\star}(x_{\text{init}};\ell) \\
&= V^\pi(x_{\text{init}};\ell) - \sum_{x,a} q^\star(x,a)\ell(x,a) \\
&= V^\pi(x_{\text{init}};\ell) - \sum_{x,a} q^\star(x,a) \left(Q^\pi(x,a;\ell) - \mathbf{1}\{x\notin X_H\} \sum_{x'} P(x'|x,a) V^\pi(x';\ell) \right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Bellman equation (1))} \\
&= V^\pi(x_{\text{init}};\ell) + \sum_{x'} \sum_{x,a} \mathbf{1}\{x\notin X_H\} q^\star(x,a) P(x'|x,a) V^\pi(x';\ell) - \sum_{x,a} q^\star(x,a) Q^\pi(x,a;\ell) \\
&= V^\pi(x_{\text{init}};\ell) + \sum_{x'\neq x_{\text{init}}} q^\star(x') V^\pi(x';\ell) - \sum_{x,a} q^\star(x,a) Q^\pi(x,a;\ell) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\textstyle\sum_{x,a} \mathbf{1}\{x\notin X_H\} q^\star(x,a) P(x'|x,a) = q^\star(x') \text{ if } x'\neq x_{\text{init}}) \\
&= \sum_{x'} q^\star(x') V^\pi(x';\ell) - \sum_{x,a} q^\star(x,a) Q^\pi(x,a;\ell) \qquad\qquad (q^\star(x_{\text{init}}) = 1) \\
&= \sum_{x,a} q^\star(x)\pi(a|x) Q^\pi(x,a;\ell) - \sum_{x,a} q^\star(x)\pi^\star(a|x) Q^\pi(x,a;\ell),
\end{aligned}$$

where the last step changes the name of the variable x' to x and uses the definition $V^\pi(x;\ell) = \sum_a \pi(a|x) Q^\pi(x,a;\ell)$. □

Below is a direct consequence of the performance difference lemma.

Corollary 1. *Denote $Q^{\pi_t}(x,a;\ell_t)$ by $Q_t(x,a)$. Then the expected regret of the learner can be decomposed as*

$$\mathbb{E}[\mathcal{R}_T] = \mathbb{E}\left[\sum_{x\in X} q^\star(x) \sum_{t=1}^T \langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), Q_t(x,\cdot)\rangle \right].$$

Note that for each x, the term $\sum_{t=1}^{T} \langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), Q_t(x,\cdot)\rangle$ is exactly the regret of a $|A|$-armed bandit problem, with $Q_t(x,\cdot)$ being the "loss vector" for this MAB instance. Thus, Corollary 1 indicates that the regret of the online RL problem can be written as the weighted average of some MAB regret over all states, where the weight for each state is its (unknown) probability of being visited by $\pi^\star$.

This suggests the following algorithmic idea: simply run an adversarial MAB algorithm on each state x to decide $\pi_t(\cdot|x)$. We emphasize that it is critical to use an *adversarial* MAB algorithm here, even if the loss functions $\ell_1, \ldots, \ell_T$ are stochastic (thus the online RL problem is a stochastic one). This is because $Q_t(x,\cdot)$ is the Q-function with respect to π_t, and π_t itself is changing over time in a potentially complicated way, making the corresponding MAB instance non-stochastic.

One slight difference compared to a standard MAB problem is that after selecting an action a, we do not exactly observe $Q_t(x,a)$. However, we can still naturally construct an estimator $\widehat{Q}_t$ for Q_t using some loss estimator $\widehat{\ell}_t$ and the transition function P (again assumed to be known for now) via: $\widehat{Q}_t(x,a) = Q^{\pi_t}(x,a;\widehat{\ell}_t)$. For example, if we again use the unbiased loss estimator of Eq. (2), then clearly $\widehat{Q}_t$ is also an unbiased estimator for Q_t.

The issue of this estimator is that its variance is quite different from that of $\widehat{\ell}_t$, making it difficult to control the local-norm from the MAB regret. For example, if we use Exp3 to solve the MAB problem at each state, then the local-norm for time t is $\mathbb{E}[\sum_{a\in A} \pi_t(a|x)\widehat{Q}_t(x,a)^2]$, which does not enjoy the critical variance cancellation effect any more. In fact, it is even difficult to control the magnitude of $\widehat{Q}_t(x,a)$ in this case, since the importance weight $q_t(x,a)$ in Eq. (2) can be arbitrarily small (making $\widehat{\ell}_t(x,a)$ arbitrarily large). Also note that unlike some other problems we have seen (such as Problem 3(b)(ii) of HW2), there is no way to modify the algorithm to enforce a lower bound on $q_t(x,a)$, since it might be the case that the transition of the MDP is such that regardless of how the learner behaves, the probability of visiting state x is always tiny.

To fix this issue, we have to sacrifice a little bit of the unbiasedness and add a small value to the importance weight. This leads to the following algorithm, taken from [Shani et al., 2020].

Algorithm 2: Policy Optimization

Input: parameter $\gamma > 0$, learning rate $\eta > 0$, and a regularizer ψ defined over $\Delta(A)$
for $t = 1, \ldots, T$ **do**

for each state x, define $\pi_t(\cdot|x) = \operatorname{argmin}_{p\in\Delta(A)} \left\langle p, \sum_{s<t} \widehat{Q}_s(\cdot,x)\right\rangle + \frac{1}{\eta}\psi(p)$
execute policy π_t and observe trajectory $\{(x_{t,h}, a_{t,h}, \ell_t(x_{t,h},a_{t,h}))\}_{h=1}^{H}$
construct Q-function estimator $\widehat{Q}_t$ such that $\widehat{Q}_t(x,a) = Q^{\pi_t}(x,a;\widehat{\ell}_t)$ where

$$\widehat{\ell}_t(x,a) = \frac{\ell_t(x,a)}{q_t(x,a)+\gamma}\mathbf{1}_t(x,a) \tag{3}$$

The following lemma summarizes the bias and variance of the estimators.

Lemma 4. *The loss estimator defined in Eq. (3) satisfies* $0 \le \ell_t(x,a) - \mathbb{E}_t[\ell_t(x,a)] \le \frac{\gamma\ell_t(x,a)}{q_t(x,a)}$ *and* $V^{\pi_t}(x_{init};\ell_t) - \mathbb{E}_t[V^{\pi_t}(x_{init};\widehat{\ell}_t)] \le \gamma|X||A|$. *Also, the corresponding Q-function estimator* $\widehat{Q}_t(x,a) = Q^{\pi_t}(x,a;\widehat{\ell}_t)$ *satisfies* $\mathbb{E}_t[\widehat{Q}_t(x,a)^2] \le \frac{H^2}{\gamma}$.

Proof. To prove the first statement, note that

$$\ell_t(x,a) - \mathbb{E}_t[\ell_t(x,a)] = \ell_t(x,a)\left(1 - \frac{q_t(x,a)}{q_t(x,a)+\gamma}\right) = \frac{\gamma\ell_t(x,a)}{q_t(x,a)+\gamma},$$

which is clearly nonnegative and at most $\frac{\gamma\ell_t(x,a)}{q_t(x,a)}$. Using this upper bound, we can prove the second statement:

$$V^{\pi_t}(x_{\text{init}};\ell_t) - \mathbb{E}_t[V^{\pi_t}(x_{\text{init}};\widehat{\ell}_t)] = \left\langle q_t, \ell_t - \mathbb{E}_t[\widehat{\ell}_t]\right\rangle \le \sum_{x,a} q_t(x,a)\frac{\gamma\ell_t(x,a)}{q_t(x,a)} \le \gamma|X||A|.$$

To prove the last statement, note that the range of $\widehat{\ell}_t(x,a)$ is $[0, 1/\gamma]$, and thus $\widehat{Q}_t(x,a)$, being the Q-function with respect to $\widehat{\ell}_t$, is in the range $[0, H/\gamma]$. Therefore, $\mathbb{E}_t[\widehat{Q}_t(x,a)^2] \le \frac{H}{\gamma}\mathbb{E}_t[\widehat{Q}_t(x,a)] \le \frac{H}{\gamma}Q_t(x,a) \le \frac{H^2}{\gamma}$. □

This lemma tells us that the loss estimator is underestimating the true loss, but even if the amount of underestimation can be very large for a specific (x,a) pair, importantly, the overall underestimation of the expected loss of π_t itself is only $\gamma|X||A|$. On the other hand, what this bias buys us is an explicit control on the variance of the Q-function estimator (which is at most H^2/γ). Together, these facts allow us to prove the following regret bound.

Theorem 2. *With the entropy regularizer $\psi(p) = \sum_{a\in A} p(a)\ln p(a)$, Algorithm 2 ensures $\mathbb{E}[\mathcal{R}_T] \le \gamma|X||A|T + \frac{H\ln|A|}{\eta} + \frac{\eta H^3 T}{\gamma}$, which is $\widetilde{\mathcal{O}}((T^2H^4|X||A|)^{1/3})$ after picking the optimal γ and η.*

Proof. First, we decompose the regret as:

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_T] &= \mathbb{E}\left[\sum_{t=1}^{T}(V_t^{\pi_t}(x_{\text{init}}) - V_t^{\pi^\star}(x_{\text{init}}))\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}(V^{\pi_t}(x_{\text{init}};\ell_t) - V^{\pi_t}(x_{\text{init}};\widehat{\ell}_t)) + \sum_{t=1}^{T}(V^{\pi_t}(x_{\text{init}};\widehat{\ell}_t) - V^{\pi^\star}(x_{\text{init}};\ell_t))\right] \\
&\le \gamma|X||A|T + \mathbb{E}\left[\sum_{t=1}^{T}(V^{\pi_t}(x_{\text{init}};\widehat{\ell}_t) - V^{\pi^\star}(x_{\text{init}};\widehat{\ell}_t))\right]
\end{aligned}$$

where the last step uses Lemma 4. We then apply the performance difference lemma to the second term:

$$\sum_{t=1}^{T}(V^{\pi_t}(x_{\text{init}};\widehat{\ell}_t) - V^{\pi^\star}(x_{\text{init}};\widehat{\ell}_t)) = \sum_{x\in X} q^\star(x)\sum_{t=1}^{T}\left\langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), \widehat{Q}_t(x,\cdot)\right\rangle.$$

For each state x, note that $\pi_t(\cdot|x)$ is obtained by Hedge with the Q-function estimators as inputs. Thus, we can apply the regret bound of Hedge to obtain:

$$\sum_{t=1}^{T}\left\langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), \widehat{Q}_t(\cdot,x)\right\rangle \le \frac{\ln|A|}{\eta} + \eta\sum_{t=1}^{T}\sum_{a\in A}\pi_t(a|x)\widehat{Q}_t(x,a)^2,$$

which is at most $\frac{\ln|A|}{\eta} + \frac{\eta H^2 T}{\gamma}$ after taking expectation on both sides and using Lemma 4. Finally, using $\sum_{x\in X} q^\star(x) = H$ and combining everything finishes the proof. □

As one can see in this analysis, because we are not able to make use of the critical variance cancellation effect from the local-norm term and instead directly enforce an H^2/γ upper bound on the variance, the final regret bound we obtain is of order $T^{2/3}$ instead of $\sqrt{T}$. However, compared to Algorithm 1, policy optimization methods like Algorithm 2 are even more efficient since it does not require solving a convex problem over the occupancy measure space. In fact, it is not even necessary to actually run an MAB algorithm for every state at every round; instead, it suffices to compute $\pi_t(\cdot|x)$ only when x is actually visited. Due to this property, policy optimization methods can even be generalized to problems with an infinite number of states.

3.1 A Different Q-Function Estimator and Extra Bonuses

Is there a way to enjoy both the nice properties of policy optimization methods and the $\sqrt{T}$-type regret bound simultaneously? As mentioned, the variance cancellation is critical if we want $\sqrt{T}$-type regret, and that requires a more careful treatment of the variance of the Q-function estimator. To this end, we study a different (and in a sense simpler) estimator for some parameter $\gamma \ge 0$:

$$\widehat{Q}_t(x,a) = \frac{L_{t,h}}{q_t(x,a)+\gamma}\mathbf{1}_t(x,a) \quad \text{where } h \text{ is s.t. } x\in X_h \text{ and } L_{t,h} = \sum_{k=h}^{H}\ell_t(x_{t,k},a_{t,k}). \tag{4}$$

Note that $L_{t,h}$ is the total loss suffered (and observed) by the learner starting from step h at round t, which by definition has conditional expectation $Q_t(x_{t,h}, a_{t,h})$ (given everything up to step h). Therefore, when $\gamma = 0$, $\widehat{Q}_t(x,a)$ is indeed an unbiased estimator for $Q_t(x,a)$. The advantage of this estimator is that its variance is in a simple form similar to its bias, as summarized below.

Lemma 5. *The Q-function estimator defined in Eq. (4) satisfies* $0 \le Q_t(x,a) - \mathbb{E}_t[\widehat{Q}_t(x,a)] \le \frac{\gamma H}{q_t(x,a)+\gamma}$ *and* $\mathbb{E}_t[\widehat{Q}_t(x,a)^2] \le \frac{H^2}{q_t(x,a)+\gamma}$.

Proof. To analyze the bias, note that

$$\mathbb{E}_t[\widehat{Q}_t(x,a)] = \mathbb{E}_t[\mathbf{1}_t(x,a)]\mathbb{E}_t\left[\frac{L_{t,h}}{q_t(x,a)+\gamma}\middle|\mathbf{1}_t(x,a)=1\right] = q_t(x,a)\frac{Q_t(x,a)}{q_t(x,a)+\gamma},$$

and thus $0 \le Q_t(x,a) - \mathbb{E}_t[\widehat{Q}_t(x,a)] = \frac{\gamma Q_t(x,a)}{q_t(x,a)+\gamma} \le \frac{\gamma H}{q_t(x,a)+\gamma}$. Similarly, the variance is

$$\mathbb{E}_t[\widehat{Q}_t(x,a)^2] = \mathbb{E}_t[\mathbf{1}_t(x,a)]\mathbb{E}_t\left[\frac{L_{t,h}^2}{(q_t(x,a)+\gamma)^2}\middle|\mathbf{1}_t(x,a)=1\right] \le \frac{q_t(x,a)H^2}{(q_t(x,a)+\gamma)^2} \le \frac{H^2}{q_t(x,a)+\gamma},$$

proving the second statement. □

If we run Hedge at each state with this Q-function estimator as inputs, then based on Corollary 1, Lemma 4, and the Hedge regret bound, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_T] &= \mathbb{E}\left[\sum_{x\in X} q^\star(x)\sum_{t=1}^T \langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), Q_t(x,\cdot)\rangle\right] \\
&\le \mathbb{E}\left[\sum_{x\in X} q^\star(x)\sum_{t=1}^T \left(\left\langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), \widehat{Q}_t(x,\cdot)\right\rangle + \left\langle \pi_t(\cdot|x), Q_t(x,\cdot) - \widehat{Q}_t(x,\cdot)\right\rangle\right)\right] \\
&\le \frac{H\ln|A|}{\eta} + \mathbb{E}\left[\sum_{x\in X} q^\star(x)\sum_{t=1}^T\sum_{a\in A}\frac{\eta H^2\pi_t(a|x)}{q_t(x,a)+\gamma}\right] + \mathbb{E}\left[\sum_{x\in X} q^\star(x)\sum_{t=1}^T\sum_{a\in A}\frac{\gamma H\pi_t(a|x)}{q_t(x,a)+\gamma}\right]. \quad (5)
\end{aligned}$$

By taking $\gamma = \eta H$ and defining a "bonus" function $b_t(x) = \sum_{a\in A}\frac{2\gamma H\pi_t(a|x)}{q_t(x,a)+\gamma}$ for each state, the regret bound above can be simplified as $\frac{H\ln|A|}{\eta} + \sum_{t=1}^T \mathbb{E}[V^{\pi^\star}(x_{\text{init}}; b_t)]$ (by seeing b_t as function on $X\times A$ so that $b_t(x,a) = b_t(x)$ for all a). While $V^{\pi^\star}(x_{\text{init}}; b_t)$ itself is still not well bounded due to the mismatch between $\pi^\star$ and π_t (in the definition of b_t), note that $V^{\pi_t}(x_{\text{init}}; b_t)$ is on the other hand nicely bounded: $V^{\pi_t}(x_{\text{init}}; b_t) = \sum_{x\in X} q_t(x)b_t(x) = \sum_{x\in X}\sum_{a\in A}\frac{2\gamma H q_t(x,a)}{q_t(x,a)+\gamma} \le 2\gamma H|X||A|$. So is there a way to somehow move from $V^{\pi^\star}(x_{\text{init}}; b_t)$ to $V^{\pi_t}(x_{\text{init}}; b_t)$?

In fact, we have seen the opposite of such techniques in Section 2 of Lecture 3 where by adding an extra penalty term to the losses fed to Hedge, we were able to turn a bound $\eta\sum_i p_t(i)(\ell_t(i)-m_t(i))^2$ (in terms of the learner's strategy p_t) to a bound $\eta(\ell_t(i^\star) - m_t(i^\star))^2$ (in terms of the optimal action $i^\star$). This suggests that if we want the opposite effect, we need to subtract a bonus term from the loss ℓ_t, and that bonus term is exactly b_t. This intuitively encourages more exploration to less frequently visited states (since b_t is large when $q_t(x)$ is small), and algorithmically amounts to subtracting a bonus $B_t(x,a) = Q^\pi(x,a;b_t)$ from the Q-function estimator fed to Hedge (or any other MAB algorithm). We summarize the algorithm and its regret guarantee below.

Theorem 3. *With the entropy regularizer* $\psi(p) = \sum_{a\in A} p(a)\ln p(a)$, *Algorithm 3 ensures* $\mathbb{E}[\mathcal{R}_T] \le \frac{H\ln|A|}{\eta} + 2\eta H^2|X||A|T + 4\eta H^5 T$, *which is* $\widetilde{\mathcal{O}}(\sqrt{H^3|X||A|T} + H^3\sqrt{T})$ *using the optimal* η.

Proof. First, decompose the regret in the following way

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_T] &= \mathbb{E}\left[\sum_{t=1}^T (V^{\pi_t}(x_{\text{init}}; \ell_t) - V_t^{\pi^\star}(x_{\text{init}}; \ell_t))\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T (V^{\pi_t}(x_{\text{init}}; \ell_t - b_t) - V_t^{\pi^\star}(x_{\text{init}}; \ell_t - b_t))\right] + \mathbb{E}\left[\sum_{t=1}^T (V^{\pi_t}(x_{\text{init}}; b_t) - V_t^{\pi^\star}(x_{\text{init}}; b_t))\right]
\end{aligned}$$

Algorithm 3: Policy Optimization with Bonuses

Input: parameter $\gamma = \eta H$, learning rate $\eta \le \frac{1}{2H^2}$, and a regularizer ψ defined over $\Delta(A)$
for $t = 1, \ldots, T$ **do**
  for each state x, define $\pi_t(\cdot|x) = \operatorname{argmin}_{p \in \Delta(A)} \left\langle p, \sum_{s<t} (\widehat{Q}_s(\cdot, x) - B_s(x, a)) \right\rangle + \frac{1}{\eta}\psi(p)$
  execute policy π_t and observe trajectory $\{(x_{t,h}, a_{t,h}, \ell_t(x_{t,h}, a_{t,h}))\}_{h=1}^H$
  construct Q-function estimator $\widehat{Q}_t$ as in Eq. (4) and a bonus Q-function B_t with $B_t(x,a) = Q^{\pi_t}(x, a; b_t)$ and $b_t(x, a') = \sum_{a \in A} \frac{2\gamma H \pi_t(a|x)}{q_t(x,a)+\gamma}$ for all $a' \in A$

$$\le \mathbb{E}\left[\sum_{x \in X} q^\star(x) \sum_{t=1}^T \langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), Q_t(x, \cdot) - B_t(x, \cdot)\rangle\right] + 2\gamma H|X||A|T - \mathbb{E}\left[\sum_{t=1}^T V_t^{\pi^\star}(x_{\text{init}}; b_t)\right]$$

where the last step uses Corollary 1 and the earlier calculation $V^{\pi_t}(x_{\text{init}}; b_t) \le 2\gamma H|X||A|$. Importantly, the last negative term helps us cancel the large bias and variance coming from the first term. Indeed, we can further upper bound the first term by (since $\widehat{Q}_t$ is an underestimator)

$$\mathbb{E}\left[\sum_{x \in X} q^\star(x) \sum_{t=1}^T \left(\left\langle \pi_t(\cdot|x) - \pi^\star(\cdot|x), \widehat{Q}_t(x, \cdot) - B_t(x, \cdot)\right\rangle + \left\langle \pi_t(\cdot|x), Q_t(x, \cdot) - \widehat{Q}_t(x, \cdot)\right\rangle\right)\right]. \tag{6}$$

Since $b_t(x, a) \le 2H$ and consequently $B_t(x, a) \le 2H^2$, under the condition $\eta \le \frac{1}{2H^2}$, we have $\eta(\widehat{Q}_t(x, a) - B_t(x, a)) \ge -1$, satisfying the condition for Hedge's local-norm regret bound. Applying it, we see that compared to Eq. (5), which is $\frac{H \ln |A|}{\eta} + \sum_{t=1}^T \mathbb{E}[V^{\pi^\star}(x_{\text{init}}; b_t)]$, the only extra term in Eq. (6), coming from the local-norm term of Hedge, is

$$\eta \mathbb{E}\left[\sum_{x \in X} q^\star(x) \sum_{t=1}^T \sum_{a \in A} \pi_t(a|x) B_t(x, a)^2\right] \le 4\eta H^4 \mathbb{E}\left[\sum_{x \in X} q^\star(x) \sum_{t=1}^T \sum_{a \in A} \pi_t(a|x)\right] = 4\eta H^5 T.$$

Combining everything (and importantly canceling $V^{\pi^\star}(x_{\text{init}}; b_t)$ with $-V^{\pi^\star}(x_{\text{init}}; b_t)$) shows $\mathbb{E}[\mathcal{R}_T] \le \frac{H \ln |A|}{\eta} + 2\gamma H|X||A|T + 4\eta H^5 T$. Plugging in $\gamma = \eta H$ finishes the proof. □

Therefore, Algorithm 3 not only enjoys the nice properties of policy optimization methods, but also achieves a regret bound that is only worse by some H factors compared to the optimal bound $\mathcal{O}(\sqrt{H|X||A|T})$ (achieved by Algorithm 1). In fact, the term $\widetilde{\mathcal{O}}(H^3\sqrt{T})$ can be further improved to $\widetilde{\mathcal{O}}(H^4)$ by slightly enlarging the bonus function. For this improvement, as well as how to deal with unknown transition (which uses similar ideas discussed earlier) and how to generalize the algorithm to MDPs with an infinite number states and a certain linear structure, see [Luo et al., 2021].

Question 3. *Once again, in both Algorithms 2 and 3, what happens if we use Tsallis entropy, such as $\psi(p) = -2\sum_{a \in A} \sqrt{p(a)}$, as the regularizer?*

References

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2020.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34: 22931–22942, 2021.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613. PMLR, 2020.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown markov games. In *International conference on machine learning*, pages 10279–10288. PMLR, 2021.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. *Advances in neural information processing systems*, 26, 2013.