CMPUT653: POLITEX (2022/02/28)

Tadashi Kozuno
University of Alberta
Edomonton, Alberta, Canada

1 POLITEX QUICK RECAP

Assume that one is given a featurized MDP $(\mathcal{M}, \phi)$, where $\phi : \mathbf{S} \times \mathbf{A} \to \mathbf{R}^d$, access to a simulator, and a G-optimal design $\mathcal{C} \subset \mathbf{S} \times \mathbf{A}$ for ϕ. Politex generates a sequence of policies $\pi_0, \pi_1, \dots$ such that for $k \geq 1$,

$$\pi_k(a|s) \propto \exp\left(\eta \bar{q}_{k-1}(s,a)\right) ,$$

where $\bar{q}_k = \hat{q}_0 + \cdots + \hat{q}_j$ with $\hat{q}_j = \Pi \Phi \hat{\theta}_j$. Here, $\hat{\theta}_j$ for $j \geq 0$ is the parameter vector obtained by running LSPE-G to evaluate policy π_j. Furthermore, $\Pi : \mathbf{R}^{\mathbf{S} \times \mathbf{A}} \to \mathbf{R}^{\mathbf{S} \times \mathbf{A}}$ truncates its argument to the closed interval $[0, H]$:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} \quad (\Pi q)(s,a) = \max(\min(q(s,a), H), 0) .$$

In the last lecture (from API to politex), we considered the mixture policy $\bar{\pi}_k$ and its value functions $v^{\bar{\pi}_k}$. However, I also said we can upper bound $v^* - v^{\pi_k}$ instead of $v^* - v^{\bar{\pi}_k}$. Below, we see how.

Recall the performance difference lemma: for any pair of policies (π, π'),

$$v^\pi - v^{\pi'} = (I - \gamma P_\pi)^{-1}(M_\pi r + \gamma P_\pi v^{\pi'} - v^{\pi'}) = (I - \gamma P_{\pi'})^{-1}(v^\pi - M_{\pi'} r - \gamma P_{\pi'} v^\pi) .$$

For brevity, we let $\Gamma_\pi \coloneqq I - \gamma P_\pi$. Using this notation and the performance difference lemma,

$$\begin{aligned}
k(v^* - v^{\pi_k}) &= k v^* - \sum_{i=0}^{k-1} v^{\pi_i} + \sum_{i=0}^{k-1} v^{\pi_i} - k v^{\pi_k} \\
&= \Gamma_{\pi^*}^{-1} \sum_{i=0}^{k-1} (M_* r + \gamma P_* v^{\pi_i} - v^{\pi_i}) + \Gamma_{\pi_k}^{-1} \sum_{i=0}^{k-1} (v^{\pi_i} - M_{\pi_k} r - \gamma P_{\pi_k} v^{\pi_i}) \\
&= \Gamma_{\pi^*}^{-1} \sum_{i=0}^{k-1} (M_* - M_{\pi_i}) q^{\pi_i} + \Gamma_{\pi_k}^{-1} \sum_{i=0}^{k-1} (M_{\pi_i} - M_{\pi_k}) q^{\pi_i} \\
&= \Gamma_{\pi^*}^{-1} \sum_{i=0}^{k-1} (M_* - M_{\pi_i}) \bar{q}_i + \Gamma_{\pi_k}^{-1} \sum_{i=0}^{k-1} (M_{\pi_i} - M_{\pi_k}) \bar{q}_i \\
&\qquad + \Gamma_{\pi^*}^{-1} \sum_{i=0}^{k-1} (M_* - M_{\pi_i}) \varepsilon_i + \Gamma_{\pi_k}^{-1} \sum_{i=0}^{k-1} (M_{\pi_i} - M_{\pi_k}) \varepsilon_i .
\end{aligned}$$

Since

$$\Gamma_{\pi^*}^{-1} \sum_{i=0}^{k-1} (M_* - M_{\pi_i}) \varepsilon_i \leq \Gamma_{\pi^*}^{-1} \sum_{i=0}^{k-1} (M_* + M_{\pi_i}) \|\varepsilon_i\|_\infty = \frac{2}{1-\gamma} \sum_{i=0}^{k-1} \|\varepsilon_i\|_\infty .$$

Consequently,

$$v^* - v^{\pi_k} = \underbrace{\Gamma_{\pi^*}^{-1} \sum_{i=0}^{k-1} (M_* - M_{\pi_i}) \bar{q}_i}_{\heartsuit} + \underbrace{\Gamma_{\pi_k}^{-1} \sum_{i=0}^{k-1} (M_{\pi_i} - M_{\pi_k}) \bar{q}_i}_{\clubsuit} + \frac{4}{1-\gamma} \sum_{i=0}^{k-1} \|\varepsilon_i\|_\infty . \tag{1}$$

The first term ($\heartsuit$) is regret, and we will explain how it is bounded in Politex in the next section. The second term ($\clubsuit$) will turn out to be less than 0, and we will prove that in the next section too.

2 Online Learning

Consider a series $(y_t)_{t=1}^T$ of vectors in $\mathbf{Y} \subset \mathbf{R}^d$, which the adversary has chosen as described in the lecture note. The aim of online learning is to choose $x_t \in \mathbf{X} \subset \mathbf{R}^d$ sequentially at each time step t given previous vectors $(y_1, \dots, y_{t-1})$ such that the regret,

$$\mathfrak{R}_T := \max_{x \in \mathbf{X}} \sum_{t=1}^T \langle x - x_t, y_t \rangle \,, \tag{2}$$

is as small as possible. A popular algorithm for this is mirror descent (MD).

In this note, we do not describe MD with full generality. Instead, we constrain our description to MD with the KL divergence. Furthermore, we assume $\mathbf{X} = \Delta(A)$ and $\mathbf{Y} = [0, H]^A$. To make it easier to connect the discussion to politex, we denote x_t and y_t as π_t and q_{t+1}, respectively. MD with the KL divergence in this case chooses π_t such that

$$\pi_t = \operatorname*{arg\,max}_{\pi \in \Delta(A)} \left(\eta \langle \pi, q_{t-1} \rangle - \mathbb{KL}(\pi \| \pi_{t-1}) \right) \propto \exp \left(\eta \sum_{u=1}^{t-1} q_u \right)$$

with $q_0 = \mathbf{0}$ and $\pi_0 = (1/A, \dots, 1/A)^\top$.

It is actually easy to prove that $(\pi_t)_{t=1}^T$ can control the regret (2), so we will prove it in this note. In particular, we have the following result.

Theorem 1. *For MD with the KL divergence run with $\eta \in \mathbf{R}_{++}$ the following inequality holds:*

$$\mathfrak{R}_T \leq \frac{\eta}{4} \sum_{t=1}^T \|q_t\|_\infty^2 + \frac{1}{\eta} \log A \,.$$

Before its proof, let us remind you why it is necessary in the analysis of Politex. Recall the $\heartsuit$ term in Equation (1), which is

$$\heartsuit = \Gamma_{\pi^*}^{-1} \sum_{i=0}^{k-1} (M_* - M_{\pi_i}) \bar{q}_i \,.$$

In this expression, $\sum_{i=0}^{k-1} (M_* - M_{\pi_i}) \bar{q}_i \leq \max_\pi \sum_{i=0}^{k-1} (M_\pi - M_{\pi_i}) \bar{q}_i$, and the last expression is the state-wise regret. Therefore, Theorem 1 yields

$$\heartsuit \leq \frac{\eta H}{4} \sum_{i=0}^{k-1} \|q_i\|_\infty^2 + \frac{H}{\eta} \log A \,.$$

We also need to upper-bound the $\clubsuit$ term. From the following lemma, it is readily shown that $\clubsuit \leq 0$.

Lemma 1. *For MD with the KL divergence run with $\eta \in \mathbf{R}_{++}$ the following inequality holds:*

$$\sum_{t=1}^T \langle \pi_t, q_t \rangle \leq \sum_{t=1}^T \langle \pi_T, q_t \rangle \,.$$

Remark 1. *This lemma seems to be important and follows from the follow-the-regularized-leader (FTRL) view of MD with the KL divergence. So, maybe we can derive a different variant of Politex with a different regularizer than the entropy.*

2.1 Proof of Theorem 1

For the proof, we need the following lemma.

Lemma 2 (Lemma 2.2 of Cesa-Bianchi & Lugosi (2006))**.** *Let X be a real-valued random variable with $a \leq X \leq b$. Then, for any $\eta \in \mathbf{R}$,*

$$\log \mathbb{E} \left[\exp \left(\eta X \right) \right] \leq \eta \mathbb{E} X + \frac{\eta^2 (b-a)^2}{8} \,.$$

Now, let us begin the proof. The following quantity plays a key role:

$$W_t := \sum_{a \in \mathbf{A}} w_t(a) \text{ , where } w_t := \exp\left(\eta \sum_{u=1}^{t} q_u(a)\right) .$$

Observe that

$$\log \frac{W_T}{W_0} = \log \sum_{a \in \mathbf{A}} w_T(a) - \log A \geq \eta \max_{a \in \mathbf{A}} \sum_{t=1}^{T} q_t(a) - \log A = \eta \max_{\pi \in \Delta(A)} \sum_{t=1}^{T} \langle \pi, q_t \rangle - \log A .$$

On the other hand, for each $t = 1, \dots, T$,

$$\begin{aligned} \log \frac{W_t}{W_{t-1}} = \log \frac{\sum_{a \in \mathbf{A}} w_t(a)}{\sum_{b \in \mathbf{A}} w_{t-1}(b)} &= \log \frac{\sum_{a \in \mathbf{A}} w_{t-1}(a) \exp\left(\eta q_t(a)\right)}{\sum_{b \in \mathbf{A}} w_{t-1}(b)} \\ &= \log \sum_{a \in \mathbf{A}} \pi_t(a) \exp\left(\eta q_t(a)\right) \\ &\leq \eta \langle \pi_t, q_t \rangle + \frac{\eta^2 \|q_t\|_\infty^2}{4} , \end{aligned}$$

where Lemma 2 is used at the last line. As a result, we have that

$$\log \frac{W_T}{W_0} = \sum_{t=1}^{T} \log \frac{W_t}{W_{t-1}} \leq \eta \sum_{t=1}^{T} \langle \pi_t, q_t \rangle + \frac{\eta^2}{4} \sum_{t=1}^{T} \|q_t\|_\infty^2$$

and

$$\log \frac{W_T}{W_0} \geq \eta \max_{\pi \in \Delta(A)} \sum_{t=1}^{T} \langle \pi, q_t \rangle - \log A .$$

Combining these two inequalities, we deduce that

$$\frac{\eta}{4} \sum_{t=1}^{T} \|q_t\|_\infty^2 + \frac{1}{\eta} \log A \geq \max_{\pi \in \Delta(A)} \sum_{t=1}^{T} \langle \pi - \pi_t, q_t \rangle = \mathfrak{R}_T .$$

This is the desired result.

2.2 Proof of Lemma 1

It is known that MD's policy update,

$$\pi_t = \arg\max_{\pi \in \Delta(A)} \left(\eta \langle \pi, q_{t-1} \rangle - \mathbb{KL}(\pi \| \pi_{t-1})\right) ,$$

is equivalent to

$$\pi_t = \arg\max_{\pi \in \Delta(A)} \left(\eta \left\langle \pi, \sum_{u=0}^{t-1} q_u \right\rangle + \mathbb{H}(\pi)\right) .$$

In other words, MD with the KL divergence is equivalent to FTRL with the entropy regularizer. This equivalence allows us to deduce that

$$\begin{aligned} \eta \sum_{t=1}^{T} \langle \pi_T, q_t \rangle &= \eta \left\langle \pi_{T-1}, q_T \right\rangle - \mathbb{H}(\pi_T) + \eta \left\langle \pi_T, \sum_{t=1}^{T-1} q_t \right\rangle + \mathbb{H}(\pi_T) \\ &\geq \eta \left\langle \pi_T, q_T \right\rangle - \mathbb{H}(\pi_T) + \eta \left\langle \pi_{T-1}, \sum_{t=1}^{T-1} q_t \right\rangle + \mathbb{H}(\pi_{T-1}) . \end{aligned}$$

Repeating the same argument, it is easy to deduce that

$$\eta \sum_{t=1}^{T} \langle \pi_T, q_t \rangle \geq \eta \sum_{t=1}^{T} \langle \pi_t, q_t \rangle - \mathbb{H}(\pi_T) + \mathbb{H}(\pi_1) .$$

By definition, $\mathbb{H}(\pi_1) = \log A \geq \mathbb{H}(\pi_T)$, and thus,

$$\eta \sum_{t=1}^{T} \langle \pi_T, q_t \rangle \geq \eta \sum_{t=1}^{T} \langle \pi_t, q_t \rangle ,$$

which is the desired result.

3 A Simple Proof of API's Lower Bound

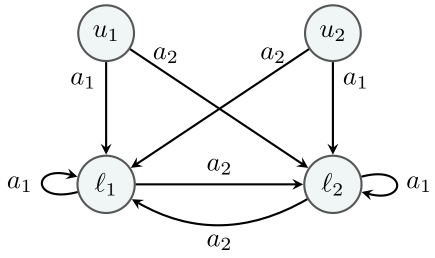

Figure 1: An environment for the proof of API's lower bound. a_1 is the action to move to the other side, and a_2 is the action to stay on the same side. The states u_i are upper states, which cannot be reached from other states. The states l_i are lower states. The sole purpose of upper states are to introduce some function approximation error and confuse the agent's decision. See the main text for details.

Suppose the MDP in Figure 1. The state and action spaces are $\mathbf{S} := \{u_1, l_1, u_2, l_2\}$. We define $\mathbf{S}_i = \{u_i, l_i\}$, $\mathbf{U} = \{u_1, u_2\}$, and $\mathbf{L} = \{l_1, l_2\}$. We often denote s_i to mean an element of $\mathbf{S}_i$. The reward function is the following

$$r(s, a_1) = \begin{cases} -\varepsilon + 2\delta & \text{if } s \in \mathbf{U} \\ -\varepsilon & \text{if } s \in \mathbf{L} \end{cases} \quad \text{and} \quad r(s, a_2) = 0\,,$$

where c is a positive scalar we specify later. [1] Clearly, the optimal policy is unique and taking a_2 everywhere. For any policy π, its action-value functions satisfy the following:

$$q^\pi(s, a_1) = \begin{cases} -\varepsilon + 2\delta\mathbb{I}\{s \in \mathbf{U}\} + \gamma v^\pi(l_1) & \text{if } s \in \mathbf{S}_1 \\ -\varepsilon + 2\delta\mathbb{I}\{s \in \mathbf{U}\} + \gamma v^\pi(l_2) & \text{if } s \in \mathbf{S}_2 \end{cases} \quad q^\pi(s, a_2) = \begin{cases} \gamma v^\pi(l_2) & \text{if } s \in \mathbf{S}_1 \\ \gamma v^\pi(l_1) & \text{if } s \in \mathbf{S}_2 \end{cases}.$$

The feature map we consider is a state-aggregation feature map:

$$\phi(s, a) = \begin{pmatrix} \mathbb{I}\{s \in \mathbf{S}_1\}\mathbb{I}\{a = a_1\} \\ \mathbb{I}\{s \in \mathbf{S}_1\}\mathbb{I}\{a = a_2\} \\ \mathbb{I}\{s \in \mathbf{S}_2\}\mathbb{I}\{a = a_1\} \\ \mathbb{I}\{s \in \mathbf{S}_2\}\mathbb{I}\{a = a_2\} \end{pmatrix}.$$

Therefore, the class of greedy policies wrt $\theta^\top \phi$ contains the optimal policy. We denote this class of policies by Π_ϕ. The action-value function is estimated by

$$\theta = \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \sum_{(s,a) \in \mathbf{S} \times \mathbf{A}} \rho(s, a) \left(\theta^\top \phi(s, a) - q^\pi(s, a)\right)^2 ,$$

where $\rho(s, a)$ is the uniform distribution over $\mathbf{S} \times \mathbf{A}$. By a quick calculation, it is easy to show that

$$\theta = \frac{1}{2} \begin{pmatrix} q^\pi(u_1, a_1) + q^\pi(l_1, a_1) \\ q^\pi(u_1, a_2) + q^\pi(l_1, a_2) \\ q^\pi(u_2, a_1) + q^\pi(l_2, a_1) \\ q^\pi(u_2, a_2) + q^\pi(l_2, a_2) \end{pmatrix}.$$

On the other hand, note that $\max_{\pi \in \Pi_\phi} \min_{\theta \in \mathbf{R}^d} \|\theta^\top \phi - q^\pi\|_\infty = \delta$.

3.1 Proof

Now let us begin the proof. Consider the policy $\pi \in \Pi_\phi$ that takes a_1 at $\mathbf{S}_1$ and a_2 at $\mathbf{S}_2$. [2] Then,

$$q^\pi(s, a_1) = \begin{cases} -\dfrac{\varepsilon}{1-\gamma} + 2\delta\mathbb{I}\{s \in \mathbf{U}\} & \text{if } s \in \mathbf{S}_1 \\ -\dfrac{\varepsilon}{1-\gamma} + 2\delta\mathbb{I}\{s \in \mathbf{U}\} + \gamma\varepsilon & \text{if } s \in \mathbf{S}_2 \end{cases} \quad q^\pi(s, a_2) = \begin{cases} -\dfrac{\gamma^2\varepsilon}{1-\gamma} & \text{if } s \in \mathbf{S}_1 \\ -\dfrac{\gamma\varepsilon}{1-\gamma} & \text{if } s \in \mathbf{S}_2 \end{cases}.$$

Therefore,

$$\theta^\top = \left(-\frac{\varepsilon}{1-\gamma} + \delta\,, -\frac{\gamma^2\varepsilon}{1-\gamma}\,, -\frac{\varepsilon}{1-\gamma} + \delta + \gamma\varepsilon\,, -\frac{\gamma\varepsilon}{1-\gamma}\right),$$

[1] Importantly, the reward is larger at u_i than that of l_i, which causes the value overestimation at $\mathbf{L}$.

[2] By symmetry, the policy $\pi \in \Pi_\phi$ that takes a_2 at $\mathbf{S}_1$ and a_1 at $\mathbf{S}_2$ shows a similar behavior.

and thus,

$$q(s,a_1) - q(s,a_2) = \begin{cases} -(1+\gamma)\varepsilon + \delta & \text{if } s \in \mathbf{S}_1 \\ -(1-\gamma)\varepsilon + \delta & \text{if } s \in \mathbf{S}_2 \end{cases} .$$

Accordingly, if $\delta = (1-\gamma)\varepsilon$ and ties are broken such that a_2 is taken everywhere, then the next policy takes a_2 at $\mathbf{S}_1$ and a_1 at $\mathbf{S}_2$. By symmetry, the same argument allows us to conclude that after two policy updates, the policy become π again.

Note that

$$v^*(l_1) - v^\pi(l_1) = \frac{\varepsilon}{1-\gamma} = \delta H^2 = H^2 \max_{\pi \in \Pi_\phi} \min_{\theta \in \mathbf{R}^d} \|\theta^\top \phi - q^\pi\|_\infty .$$

This concludes the proof.

References

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, USA, 2006. ISBN 0521841089.