CS292F Statistical Foundation of Reinforcement Learning **Spring 2021**

Lecture 3: Markov Decision Processes II, April 5

Lecturer: Yu-Xiang Wang *Scribes: Fuheng Zhao*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

3.1 Some Recaps

The following equations are either taken from the lecture slides or from the textbook [1].

3.1.1 Markov Decision Process (MDP)

A discounted Markov Decision Process $M = (S, A, P, r, \gamma, \mu)$ where S denote state space, A denote action space, P denote transition function, r denote reward function, γ denote discounted factor, and μ denote initial state distribution.

3.1.2 Reward function and Value functions

In the following section, π denotes a stationary policy in which the following actions is based on the current state. The expected immediate reward function $r(s, a)$,:

$$r(s, a) = E[R_1|S_1 = s, A_1 = a]$$

$$r^\pi(s) = E_{a\sim\pi(.|s)}[R_1|S_1 = s]$$

The state value function $V^\pi(s)$ denotes expected long-term return when starting in s and following policy π:

$$V^\pi(s) = E_\pi(\sum_{i=1} \gamma^{i-1} R_i|S_1 = s)$$

Similarly, the state-action value function $Q^\pi(s, a)$ denote the expected long-term return when starting in s, performing a, and following π

$$Q^\pi(s, a) = E_\pi(\sum_{i=1} \gamma^{i-1} R_i|S_1 = s, A_1 = a)$$

3.1.3 Bellman consistency equations

Lemma 3.1. *Given π is a stationary policy, for all $s \in S$ and $a \in A$, $V^\pi(s) = Q^\pi(s, \pi(s))$ and $Q^\pi(s, a) = r(s, a) + \gamma E(V^\pi_{s'\sim P(.|s,a)}(s'))$*

By shifting the Markov decision process one step into the future, one can obtain the following equations:

$$V^\pi = r^\pi + \gamma P^\pi V^\pi$$

$$Q^\pi = r + \gamma P V^\pi$$

$$V^\pi = r^\pi + \gamma \hat{P}^\pi V^\pi$$

3.2 Statistic Tools

The law of large number says that as n grows, the probability having the sample mean of independent and identically distributed random variables equal to the expected value goes to 1.

Theorem 3.2. *Law of Large Number. Let X_1, X_2, ... , X_n be independent and identically distributed random variables, and let the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then, as n approach infinity, $P(\bar{X}_n = E(X_1)) = 1$.*

The central limit theorem says that as n grows, the distribution of the sample mean of independent and identically distributed random variables converge to normal distribution.

Theorem 3.3. *Central limit theoreom. Law of Large Number: Let X_1, X_2, ... , X_n be independent and identically distributed random variables. Then as n approach infinity, $\sqrt{n}(\frac{1}{n}\sum_{i=1}^n X_i - E(X_1))$ approach $Normal(0, Var(X_1))$*

The Hoeffding's Inequality is often used to bound the algorithm's failure probability by δ, a small constant $\delta << 1$.

Lemma 3.4. *Hoeffding's Inequality. Let $X_1, \cdots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ and let $S_n = \sum_{i=1}^n X_i)$. Then for any $t > 0$ we have:*

$$Pr[S_n - E[S_n] > t] \leq exp(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2})$$

An alternative easy version is to have $0 < X_i < B$, then with high probablity, $1-\delta$, $|\bar{X} - E(\bar{X})| \leq \sqrt{\frac{B^2}{2n} log(2/\delta)}$. This upper bound can be derived by setting $t = \sqrt{\frac{nB^2}{2} log(2/\delta)}$ and consider both side of the tail to bound the absolute value.

When one desire a even stronger guarantees compare to Hoeffding's Inequality, one can use the Bernstein's Inequality.

Lemma 3.5. *Bernstein's Inequality. Let $X_1, \cdots, X_n$ be independent zero-mean random variables such that $-M \leq X_i \leq M$. Then for any $t > 0$ we have:*

$$Pr[|\sum_{i=1}^n X_i| > t] \leq exp(-\frac{\frac{1}{2}t^2}{\sum_1^n E(X_i^2) + \frac{1}{3}Mt})$$

Similarly to Hoeffiding's inequality, an alternative version of Bernstein's inequality is to have $0 < X_i < B$ such that $Var(X_i) \leq \frac{B^2}{4}$, then with high probability, $1 - \delta$, $|\bar{X} - E(\bar{X})| \leq \sqrt{\frac{2Var(X_1)}{n} log(2/\delta)} + \frac{2Mlog(2/\delta)}{3n}$.

Moreover a generalization of Hoeffding's Inequality, McDiarmid's Inequality by bounding the failure probablity of a special function of the data.

Lemma 3.6. *McDiarmid's Inequality. Let $X_1, \cdots, X_n$ be independent random variables, and lef a function f statisfy coordinate uniform stability condition where for all $i \in \{1, 2..., n\}$ and for all $x_1, x_2, ..., x_n, x_{i'} \in X$, $|f(x_1, ..., x_i, .., x_n) - f(x_1, ..., x_{i'}, .., x_n)| \leq c_i$ Then for any $t > 0$ we have:*

$$Pr[f(X_1, ..., X_n) - E(f(X_1, ..., X_n)) \geq \epsilon] \leq exp(-\frac{2\epsilon^2}{\sum_1^n c_i^2})$$

A special trick often used to bound the failure probability of a set of events is called the union bound. More formally, for a countable sequence of events, $P(\cup_i A_i) \leq \sum_i P(A_i)$.

Lemma 3.7. *Concentration for Discrete Distributions: Let z be a discrete random variable that takes values in 1,...,d, distributed according to q. We write q as a vector where $q = [Pr(z = j)]_{j=1}^d$. Assume we have N iid samples, and that our empirical estimate of q is $[q]_j = \sum_{i=1}^N \frac{z_i = j}{N}$, then for all $\epsilon > 0$):*

$$Pr(\|\hat{q} = q\|_1 \geq \sqrt{d}(\frac{1}{\sqrt{N}} + \epsilon)) \leq exp(-N\epsilon^2)$$

3.3 Simulation Lemma and model-based approach

The transition matrix P is a very large matrix, $|S||A|$ x $|S|$, can we use a sparse matrix $\hat{P}$, a sampled transition to approximation of P to reduce the computational complexity but still get the desired result. How many samples do we need to draw to obtain an ϵ-optimal policy (sample size should be less than $|S|$).

Define $\hat{P}(s'|s, a) = \frac{count(s', s, a)}{N}$ in which $count(s', s, a) = \sum_{i=1}^N 1(S'(S'_{i,s,a} = s')$. Since the sample size $N << |S|$, we expect many entry in $\hat{P}$ is 0. The time complexity of computing $\hat{P}$ is $O(N|S||A|)$ and the space complexity is $O(N|S||A|)$.

Let's define the approximate Markov decision process $\hat{M} = (S, A, \hat{P}, r, \gamma, \mu)$. Then, one can run value iteration or policy iteration on $\hat{M}$ to obtain $\hat{\pi}^* = argmax \hat{Q}^*(s, a)$ where $\hat{Q}^*$ is the optimal state value function under $\hat{M}$. To show $\hat{M}$ is ϵ-optimal, we want to bound:

$$Q^{\pi*} - Q^{\hat{\pi}*} = Q^{\pi*} + (\hat{Q}^{\pi*} - \hat{Q}^{\pi*}) + (\hat{Q}^{\hat{\pi}*} - \hat{Q}^{\hat{\pi}*}) - Q^{\hat{\pi}*} \leq 2\epsilon \tag{3.1}$$

To show Equation 3.1 holds, it is equivalent to show the following two equations holds.

$$Q^{\pi*} - \hat{Q}^{\pi*} \leq \epsilon$$
$$\hat{Q}^{\hat{\pi}*} - Q^{\hat{\pi}*} \leq \epsilon$$

Lemma 3.8. *Simulation Lemma: For all π we have that: $Q^\pi - \hat{Q}^\pi = \gamma(1 - \gamma\hat{P}^\pi)^{-1}(P - \hat{P})V^\pi$*

Proof. $Q^\pi - \hat{Q}^\pi = (I - \gamma\hat{P}^\pi)^{-1}(I - \gamma\hat{P}^\pi)Q^\pi - (I - \gamma\hat{P}^\pi)^{-1}(I - \gamma P^\pi)^{-1}Q^\pi$
$= (I - \gamma\hat{P}^\pi)^{-1}((I - \gamma\hat{P}^\pi) - (I - \gamma P^\pi)^{-1})Q^\pi$
$= \gamma(I - \gamma\hat{P}^\pi)^{-1}(P^\pi - \hat{P}^\pi)Q^\pi$
$= \gamma(I - \gamma\hat{P}^\pi)^{-1}(P - \hat{P})V^\pi$

□

Lemma 3.9. *For any π, M, and $x \in R^{\{|S|,|A|\}}$, $\|(I - \gamma P^\pi)^{-1}x\|_\infty \leq \frac{\|x\|_\infty}{1-\gamma}$*

Proof. Let $x = (I - \gamma P^\pi)(I - \gamma P^\pi)^{-1}x = (I - \gamma P^\pi)y$ where $y = (I - \gamma P^\pi)^{-1}x$. By triangle inequality:

$$\|x\| = \|(I - \gamma P^\pi)y\| \geq \|y\|_{\text{inf}} - \gamma\|P^\pi y\|_{\text{inf}} \geq \|y\|_{\text{inf}} - \gamma\|y\|_{\text{inf}}$$

□

3.3.1 Applying Simulation Lemma

Firstly, we can show that using $O(S^2A)$ space is sufficient to provide accurate model using uniform convergence via simulation lemma, such that for all policies π, $\|Q^\pi - \hat{Q}^\pi\|_\infty \leq \epsilon$.

$\|Q^{\pi*} - \hat{Q}^{\pi*}\|_\infty = \|\gamma(I - \gamma\hat{P}^\pi)^{-1}(P - \hat{P}V^\pi)\|_\infty$
$\leq \frac{\gamma}{1-\gamma}\|(P - \hat{P})\|_\infty$
$\leq \frac{\gamma}{1-\gamma}(max_{s,a}\|P(.|s,a) - \hat{P}(.|s,a)\|_1)\|V^\pi\|_\infty$
$\leq \frac{\gamma}{1-\gamma}max_{s,a}\|P(.|s,a) - \hat{P}(.|s,a)\|_1.$

By lemma 3.7, for some constant c, sample size m, and fixed s, a, with high success probability, i.e, $1-\delta$, $|P(.|s,a) - \hat{P}(.|s,a)\|_1 \leq c\sqrt{\frac{|S|log(1/\delta)}{m}}$. We can then union bound of $m|S\|A|$ samples by decrease the failure probability to $\frac{\delta}{|S\|A|}$.

Hence, setting $m \geq \frac{2\gamma^2(log(\frac{2SA}{\delta})+S)}{(1-\gamma)^4\epsilon^2}$ is sufficent to bound $\|Q^\pi - \hat{Q}^\pi\|_\infty \leq \epsilon$. However, this $O(m)$ grows linear to S, meaning we will still need S^2A matrix.

3.3.2 Bounding the value function instead

Lemma 3.10. *Q-error amplification:* $V^{\pi Q} \geq V^* - \frac{2\|Q-Q*\|_\infty}{1-\gamma}$

If we can bound $\|\hat{Q}^* - Q^*\|$ with error independent to S, then we automatically improve upon the previous bound.

Lemma 3.11. $\|\hat{Q}^* - Q^*\|_\infty \leq \frac{\gamma}{1-\gamma}\|(P - \hat{P}V^*)\|_\infty$

Proof. $\|\hat{Q}^* - Q^*\|_\infty = \|Q^* - \hat{\tau}Q^* + \hat{\tau}Q^* - \hat{\tau}\hat{Q}^*\|_\infty$
$\leq \|Q^* - \hat{\tau}Q^*\|_\infty + \|\hat{\tau}Q^* - \hat{\tau}\hat{Q}^*\|_\infty$
$\leq \|\gamma + \gamma PV^* - (\gamma + \gamma\hat{P}\frac{max_{s,a}Q^*(s,a)}{V^*})\|_\infty + \gamma\|Q^* - \hat{Q}^*\|_\infty$
$= \|\gamma(P - \hat{P}V^*)\|_\infty + \gamma\|Q^* - \hat{Q}^*\|_\infty$
$\leq \frac{\gamma}{1-\gamma}\|(P - \hat{P}V^*)\|_\infty$ □

Instead of solving S dimensional concentration with bounded l_1 norm, we used inner produce to collapse the S dimension into 1 dimension, such that $\|(P - \hat{P}V^*)\|_\infty = max_{s,a}|E_{s'\sim P(.|s,a)}[V * s'] - E_{s'\sim\hat{P}(.|s,a)}[V^*(s')]| = max_{s,a}|E_{s'\sim P(.|s,a)}[V^*(s')] - \frac{1}{N}\sum_{i=1}^{N} V^*S_{i,s,a}^{'}|$. We can then apply Hoeffding's inequality to bound the equation by $\frac{1}{1-\gamma}\sqrt{\frac{log(1/\delta)}{2N}}$.

Since $V^* - V^{\hat{\pi}^*} \leq \frac{2\|Q^* - \hat{Q}^{\hat{\pi}^*}\|_\infty}{1-\gamma} \leq \frac{2}{(1-\gamma)^3}\sqrt{log(c/\delta')/2m} = \epsilon$ where $\delta^{'} = \frac{\delta}{SA}$, we obtain $m \geq \frac{2\gamma}{(1-\gamma)^3}\frac{log(cSA/\delta)}{\epsilon^2}$ and the computational complexity is $O(SA(m + \#VI))$.

3.3.3 Optimal sample complexity

In 2013, Azar et al, [2] proved the matching lower bound $m = \theta(\frac{1}{(1-\gamma)^3}\frac{log(cSA/\delta)}{\epsilon^2})$ sample complexity of estimating the optimal action-value function. Very recently, a group of researchers from UC Santa Barbara, Yin, et al, [3] proposed off-policy double variance reduction approach to achieve the optimal sample complexity for offline RL in stationary transition setting. It remains an open problem whether model-based plug-in is optimal for all ϵ.

References

[1] Alekh Agarwal, Nan Jiang, and Sham M Kakade. "Reinforcement learning: Theory and algorithms." In: *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep* (2019).

[2] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model." In: *Machine learning* 91.3 (2013), pp. 325–349.

[3] Ming Yin, Yu Bai, and Yu-Xiang Wang. "Near-Optimal Offline Reinforcement Learning via Double Variance Reduction." In: *arXiv preprint arXiv:2102.01748* (2021).