CS292F StatRL Lecture 3
MDP with a generative model

Instructor: Yu-Xiang Wang

Spring 2021

UC Santa Barbara

Recap: Markov Decision processes (MDP)

- Infinite horizon / discounted setting

$$\mathcal{M}(\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$$

Transition kernel: $P: \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ i.e. $P(s'|s,a)$

(Expected) reward function: $r: \mathcal{S} \times \mathcal{A} \to \mathbb{R} / [0, R_{max}]$ $\mathbb{E}[R_t \mid S_t = s, A_t = a] =: r(s,a)$

Initial state distribution $\mu_0 \in \Delta(\mathcal{S})$

Discounting factor: γ

Recap: Reward function and Value functions

- Immediate reward function r(s,a)
  - expected **immediate** reward

$$r(s,a) = \mathbb{E}[R_1 | S_1 = s, A_1 = a]$$
$$r^{\pi}(s) = \mathbb{E}_{a \sim \pi(a|s)}[R_1 | S_1 = s]$$

- state value function: $V^{\pi}(s)$
  - expected **long-term** return when starting in *s* and following π

$$V^{\pi}(s) = \mathbb{E}_{\pi}[R_1 + \gamma R_2 + ... + \gamma^{t-1} R_t + ... | S_1 = s]$$

- state-action value function: $Q^{\pi}(s,a)$
  - expected **long-term** return when starting in *s*, performing *a*, and following π

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}[R_1 + \gamma R_2 + ... + \gamma^{t-1} R_t + ... | S_1 = s, A_1 = a]$$

Recap: Bellman equations

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

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

$$Q^\pi = r + \gamma \breve{P}^\pi Q^\pi \,.$$

$V^\pi = r + \gamma (P^\pi)^\top V^\pi$

$\breve{V}^\pi = \breve{r}^\pi + \gamma (\breve{P}^\pi)^\top \breve{V}^\pi$

Recap: Duality and LP-formulation

- Primal LP:

$$\begin{aligned} \min \quad & \sum_s \mu(s) V(s) \\ \text{subject to} \quad & V(s) \geq r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \quad \forall a \in \mathcal{A},\ s \in \mathcal{S} \end{aligned}$$

- Dual LP:

$$\begin{aligned} \max_{\nu} \quad & \sum_{s,a} \nu(s,a) r(s,a) \\ \text{subject to} \quad & \nu \geq 0 \\ & \sum_z \nu(s,a) = \mu(s) + \gamma \sum_{s',a'} P(s|s',a') \nu(s',a') \end{aligned}$$

Recall: Bellman optimality equation and a stationary and deterministic optimal policy

$$V^*(s) = \max_a \sum_{s'} P(s'|s,a)[r(s,a,s') + \gamma V^*(s')]$$

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$

Value iterations (VI) aim at finding the fixed point
by recursively applying the Bellman (optimality) operator.

Lemma 1. The Bellman operator is a γ-contraction.

$$\textit{For any two vectors } Q, Q' \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|},$$

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

Recap: computational complexity

	Value Iteration	Policy Iteration	LP-Algorithms
Poly?	$\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert\frac{L(P,r,\gamma)\log\frac{1}{1-\gamma}}{1-\gamma}$	$(\lvert\mathcal{S}\rvert^3+\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert)\frac{L(P,r,\gamma)\log\frac{1}{1-\gamma}}{1-\gamma}$	$\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert L(P,r,\gamma)$
Strongly Poly?	✗	$(\lvert\mathcal{S}\rvert^3+\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert)\cdot\min\left\{\frac{\lvert\mathcal{A}\rvert^{\lvert\mathcal{S}\rvert}}{\lvert\mathcal{S}\rvert},\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert\log\frac{\lvert\mathcal{S}\rvert^2}{1-\gamma}}{1-\gamma}\right\}$	$\lvert\mathcal{S}\rvert^4\lvert\mathcal{A}\rvert^4\log\frac{\lvert\mathcal{S}\rvert}{1-\gamma}$

A trivial lower bound: Ω(SA) needed to store the Q* function.

Question: If we allow randomness, can we further improve the computational complexity?

- Large MDPs
  - Backgammon: 10^{20}
  - Chess: 10^{47}
  - Game of Go: 10^{174}
- The transition kernel requires S^2A parameters to describe, and to apply.
  - VI, PI, LP all depends at least S^2A
- What if we can sample transition in O(1)?

Access to a simulator or a generative model $S' \sim P(\cdot|s,a)$

- Popularized by Kakade (2003)

- Examples when this is a meaningful model:
  - Games
  - Robotics simulation
  - RL for Science

- Not the most realistic if
  - The simulator is a crude approximation of the world
  - You cannot take a snapshot and restart

How many generative model oracle calls do we need to obtain an ε-optimal policy?

- (oracle) computational complexity
  - Assume O(1) time to draw sample $S' \sim P(\cdot|s,a)$

- But also can be viewed as a simplified version of the sample complexity of RL
  - without worrying about exploration.
  - Let's get N samples **for each** (s,a) pairs.
  - How is N related to ε

How are we using the simulator?

- We will consider the dumbest way of using it
  - Sampling N rounds. Each round go over each (s,a) pair.
- A total of NSA oracle calls.
  - We have N samples for each SA, but often N<<S
- It is possible to do better than this, but not in the worst case, so we will study this algorithm first.

Plug-in estimator of P

$$\widehat{P}(s'|s,a) = \frac{\text{count}(s',s,a)}{N} \text{ where } \text{count}(s',s,a) = \sum_{i=1}^{N} \mathbf{1}(S'_{i,s,a} = s').$$

- How many parameters does P have?

- Often in large MDP, N<<S

Key question of interest:

Do we need to estimate P accurately to obtain near optimal policies?

Outline of the lecture today

- Simulation Lemma and model-based approach
- Review of statistical tools we need:
  - Hoeffding's inequality
  - Bernstein inequality
  - McDiarmid's inequality
- Sample complexity bounds

Model-based approach

- Approximate MDP
  - Run VI, PI on the approximate MDP

- From uniform convergence to suboptimality bound

Computational complexity of the model-based approach

- To construct the approximate transition kernel

- To compute empirically optimal policy
  - via value iteration

Attempt 1: Simulation Lemma (Kearns and Singh, 2002)

Lemma 2.2. *(Simulation Lemma) For all π we have that:*

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

- Proof using closed-form solution for Q

$$\begin{aligned} Q^\pi - \widehat{Q}^\pi &= (I - \gamma P^\pi)^{-1} r - (I - \gamma \widehat{P}^\pi)^{-1} r \\ &= (I - \gamma \widehat{P}^\pi)^{-1}((I - \gamma \widehat{P}^\pi) - (I - \gamma P^\pi))Q^\pi \\ &= \gamma(I - \gamma \widehat{P}^\pi)^{-1}(P^\pi - \widehat{P}^\pi)Q^\pi \\ &= \gamma(I - \gamma \widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi \end{aligned}$$

Uniform convergence via the Simulation Lemma

$$\|Q^\pi - \widehat{Q}^\pi\|_\infty = \|\gamma(I - \gamma\widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi\|_\infty \leq \frac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\pi\|_\infty \quad (1)$$

$$\leq \frac{\gamma}{1-\gamma}\left(\max_{s,a}\|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1\right)\|V^\pi\|_\infty \quad (2)$$

$$\leq \frac{\gamma}{(1-\gamma)^2}\max_{s,a}\|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1 \quad (3)$$

- We proved (1) when we prove the invertability in Lecture 2

- Key observation: RHS doesn't depend on the policy.

All (you need to know) about Statistics in one slide, two theorems.

- Statistics is about using **samples** from a distribution to infer the properties of the distribution itself (**population**)
  - $X_1, X_2, X_3, \ldots, X_n \sim P$

- Law of large number
  - Average ---> Mean

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}[X_1]$$

- Central limit theorem
  - The rate is sqrt(1/n)

$$\sqrt{n} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_1] \right) \to N(0, \mathrm{Var}(X_1))$$

Concentration inequalities --- finite-sample bounds of LLN and CLT

- **Hoeffding's inequality**: Assume X_1, …, X_n are independent and their support bounded:

$$S_n = X_1 + \cdots + X_n$$

$$\mathrm{P}(S_n - \mathrm{E}[S_n] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$

- Easy version, if $0<X_i<B$, with probability $1-\delta$:

$$\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \leq \sqrt{\frac{B^2}{2n}\log(2/\delta)}$$

Concentration inequalities --- finite-sample bounds of LLN and CLT

- **Bernstein inequality**: Assume $X_1, \ldots, X_n$ are independent, **zero-mean**, and their absolute value bounded by M, then

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] + \frac{1}{3}Mt}\right).$$

- Easy version for the iid case, with probability 1-δ:

$$\left|\bar{X} - \mathbb{E}[X_1]\right| \leq \sqrt{\frac{2\mathbf{Var}[X_1]}{n}\log(2/\delta)} + \frac{2M\log(2/\delta)}{3n}$$

A generalization of Hoeffding's inequality to McDiarmid's Inequality

McDiarmid's inequality: Assume X_1, …, X_n are independent, and function f satisfies the following

Coordinatewise Uniform Stability condition:

$$\sup_{x_1,\cdots,x_{i-1},x_i,x_i',x_{i+1},\cdots,x_n} |f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\cdots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\cdots,x_n)| \leq c_i.$$

Then we have:

$$\mathrm{P}(f(X_1,X_2,\cdots,X_n) - \mathbb{E}[f(X_1,X_2,\cdots,X_n)] \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right),$$

High probability bound for estimating probability distribution in L1

- Apply McDiarmid inequality

- Calculate the expectation

The “Union bound” trick

- “Union bound”: For a countable sequence of events $A_1, A_2, A_3, \ldots$

$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}(A_i).$$

- Typical use case:
  - Bound low-probability events together

Continue with the uniform convergence via the “Simulation Lemma”

$$\|Q^\pi - \widehat{Q}^\pi\|_\infty = \|\gamma(I - \gamma\widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi\|_\infty \leq \frac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\pi\|_\infty \quad (1)$$

$$\leq \frac{\gamma}{1-\gamma}\left(\max_{s,a}\|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1\right)\|V^\pi\|_\infty \quad (2)$$

$$\leq \frac{\gamma}{(1-\gamma)^2}\max_{s,a}\|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1 \quad (3)$$

Summary of Attempt 1: “simulation lemma” + uniform convergence

- Sample complexity

- Computational complexity

Exercise: Try the alternatives

- Try applying Hoeffding's inequality coordinatewise, then union bound over s'
  - Could you recover the same bound?

- Try applying Bernstein's inequality coordinatewise, then union bound over s'
  - Do you need additional assumptions to get the same bound?

Attempt 2: Bounding the value function instead

- Recall:

Lemma 1.11 AJKS (Q-error amplification):

$$V^{\pi_Q} \geq V^\star - \frac{2\|Q - Q^\star\|_\infty}{1-\gamma}\mathbb{1}.$$

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

Bounding the value function

- Key lemma: $\|Q^\star - \widehat{Q}^\star\|_\infty \leq \frac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\star\|_\infty$

- Proof: Use the contraction of Bellman (optimality) operator.

The key trick for knocking off an S factor is the following:

- **Key lemma:** $\|Q^\star - \widehat{Q}^\star\|_\infty \leq \frac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\star\|_\infty$

$$\|(P - \widehat{P})V^\star\|_\infty = \max_{s,a} \left| E_{s' \sim P(\cdot|s,a)}[V^\star(s')] - E_{s' \sim \widehat{P}(\cdot|s,a)}[V^\star(s')] \right|$$

$$= \max_{s,a} \left| E_{s' \sim P(\cdot|s,a)}[V^\star(s')] - \frac{1}{N}\sum_{i=1}^{N} V^*(S'_{i,s,a}) \right|$$

Apply Hoeffding's inequality!

Summary of Attempt 2: “Q-amplification” + Bellman operator

- Sample complexity

- Computational complexity

Optimal sample complexity (Azar et al., 2013)

$$N = \Theta\left(\frac{1}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{\epsilon^2}\right)$$

Recent literature:

- (Sidford et al, 2018) A variance reduced approx. value iteration-based approach for $\varepsilon < 1$
- (Agarwal et al., 2019) Proven the same for model-based approach for $\varepsilon < \sqrt{1/(1-\gamma)}$
- (Li et al., 2020) optimal rates for all values of $\varepsilon < 1/(1-\gamma)$ for a perturbed model-based approach.
- (Yin, Bai, W., 2020) optimal rates for **the finite horizon case** with model-based plug-in method. $\varepsilon < \sqrt{H}$
- (Yin, Bai, W., 2021) double variance reduction, all values of $\varepsilon < H$, finite horizon case (and $\varepsilon < 1/(1-\gamma)$ for the infinite horizon case too)

It remains an open problem whether model-based plug-in is optimal for all ε

References on the minimax sample complexity

Azar, M. G., Munos, R., & Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, *91*(3), 325-349.

Sidford, A., Wang, M., Wu, X., Yang, L. F., & Ye, Y. (2018). Near-optimal time and sample complexities for solving Markov decision processes with a generative model. *Advances in Neural Information Processing Systems*

Agarwal, A., Kakade, S., & Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory* (pp. 67-83). PMLR.

Li, G., Wei, Y., Chi, Y., Gu, Y., & Chen, Y. (2020). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, *33*.

Yin, M., Bai, Y., & Wang, Y. X. (2020). Near optimal provable uniform convergence in offline policy evaluation for reinforcement learning. *In AISTATS'2021*.

Yin, M., Bai, Y., & Wang, Y. X. (2021). "Near-Optimal Offline Reinforcement Learning via Double Variance Reduction." *arXiv preprint arXiv:2102.01748* (2021).

Next lecture

- Notes on finite horizon MDP
- Some ideas behind how to improve the dependence on H or $1/(1-\gamma)$.
- RL algorithms:
  - Temporal difference learning
  - TD-learning with function approximation