CS292F StatRL Lecture 11 Exploration in Linear MDP & Introduction to offline RL

Instructor: Yu-Xiang Wang Spring 2021 UC Santa Barbara

Logistics

- Project midterm milestone due
 - Important as I need to allocate space for student presentation
- For those who haven't submitted HW1
 - You don't have to solve everything, just submit what you have
 - HW1 is long I am thinking of adjusting grading criteria
- HW2 is not as long
 - Don't wait

Recap: Lecture 10

- Exploration in Linear MDPs
- Properties of Linear MDPs
- Algorithm: UCB-VI for Linear MDPs
- Regret analysis

Recap: Impossibility results

- What are the assumptions to make?
 - Q*(s,a) approximately linear?
 - $Q^{\pi}(s,a)$ is approximately linear for all π ? $\exists e$
 - Q*(s,a) is exactly linear?

Weisz et al (ALT-2020): http://proceedings.mlr.press/v1 32/weisz21a.html

• $Q^{\pi}(s,a)$ is exactly linear for all π ?

Exponential sample complexity / regret lower bounds for the approximate case...

Open pullen

(Du, Kakade, Wang, Yang, 2019) Is a good representation sufficient for sample efficient reinforcement learning?

Recap: Linear MDPs

- Exists feature map $\phi: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$
 - Such that:

$$r_{h}(s,a) = \theta_{h}^{\star} \cdot \phi(s,a), \quad P_{h}(\cdot|s,a) = \mu_{h}^{\star}\phi(s,a), \forall h$$

$$\int_{h}(\cdot|\cdot,\cdot) \in |\mathcal{I}^{\mathsf{S}} \times |\mathcal{I}^{\mathsf{S}}| \qquad |\mathcal{I}^{\mathsf{S}} \times |\mathcal{I}^$$

(Jin et al., 2020) Provably efficient reinforcement learning with linear function approximation

Recap: UCB-VI for Linear MDPs

- In every round:
 - 1. Run Ridge regression for estimating the model

$$\widehat{\mu}_{h}^{n} = \operatorname{argmin}_{\mu \in \mathbb{R}^{|\mathcal{S}| \times d}} \sum_{i=0}^{n-1} \left\| \mu \phi(s_{h}^{i}, a_{h}^{i}) - \delta(s_{h+1}^{i}) \right\|_{2}^{2} + \lambda \|\mu\|_{F}^{2}$$

$$\widehat{\mu}_{h}^{n} = \sum_{i=0}^{n-1} \delta(s_{h+1}^{i}) \phi(s_{h}^{i}, a_{h}^{i})^{\top} (\Lambda_{h}^{n})^{-1}$$

2. Construct the exploration bonuses

$$b_h^n(s,a) = \beta \sqrt{\phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s,a)},$$

3. Run optimistic value iterations, and update greedy policy

Recap: Regret bound

• Choose
$$\beta = Hd \left(\sqrt{\ln \frac{H}{\delta}} + \sqrt{\ln(W + H)} + \sqrt{\ln B} + \sqrt{\ln d} + \sqrt{\ln N} \right)$$
$$\lambda = 1$$
$$b_h^n(s, a) = \beta \sqrt{\phi(s, a)^\top (\Lambda_h^n)^{-1} \phi(s, a)},$$
$$b_h^n(s, a) = \beta \sqrt{\phi(s, a)^\top (\Lambda_h^n)^{-1} \phi(s, a)},$$
$$f(s, a) = \beta \sqrt{\phi(s, a)^\top (\Lambda_h^n)^{-1} \phi(s, a)},$$
$$f(s, a) = \beta \sqrt{\phi(s, a)^\top (\Lambda_h^n)^{-1} \phi(s, a)},$$
$$f(s, a) = \beta \sqrt{\phi(s, a)^\top (\Lambda_h^n)^{-1} \phi(s, a)},$$
$$f(s, a) = \beta \sqrt{\phi(s, a)^\top (\Lambda_h^n)^{-1} \phi(s, a)},$$
$$f(s, a) = \beta \sqrt{\phi(s, a)^\top (\Lambda_h^n)^{-1} \phi(s, a)},$$
$$f(s, a) = \beta \sqrt{\phi(s, a)^\top (\Lambda_h^n)^{-1} \phi(s, a)},$$
$$f(s, a) = \beta \sqrt{\phi(s, a)^\top (\Lambda_h^n)^{-1} \phi(s, a)},$$
$$f(s, a) = \beta \sqrt{\phi(s, a)^\top (\Lambda_h^n)^{-1} \phi(s, a)},$$
$$f(s, a) = \beta \sqrt{\phi(s, a)^\top (\Lambda_h^n)^{-1} \phi(s, a)},$$
$$f(s, a) = \beta \sqrt{\phi(s, a)^\top (\Lambda_h^n)^{-1} \phi(s, a)},$$

Recap: Regret analysis



Regret of episode t n

Cimilation

Pegret =

Æ

- Optimism / simulation lemma
- Sum them up, to get total regret, the help the first total regret, the help the first total regret, the help the first total with the lend
 - Same information-gain bound from linear bandits

$-p(\cdot|s_{a}) + p(\cdot|s_{a}) = \hat{\mu} \cdot \phi(s_{a}) - \hat{\mu}^{*} \cdot \phi(s_{a})$ Recap: It remains to prove (S_{G}) - 1. Uniform convergence bound 52 • 2. "Opti The same induction argument as in the UCB-VI for tabular MDP (Read Lemma 7.10 in AJKS)

• 3. "Information gain" bound

The same argument as in the Linear Bandits case. (Read Lemma 7.12 in AJKS) Se the fact that $\delta(s) \mid V = V(s)$. Thus the operator $P_h^n(\cdot \mid s, a) \cdot V$ simply requires storing all data d via simple linear algebra and the computation complexity is simply poly(d, n)—no poly dependence.



• The quantity of interest is a inner product with this:



Challenge: we cannot use union bound because we have an infinite number of value functions $\sup_{x \in \mathcal{A}_{i}} \sum_{x \in \mathcal{A}_{i}$

- A covering number argument.



What is a finite set to cover this class such that for every f in this set, there is a function in the finite set, such that they are ε -close in sup-norm?

$$[Lemma: De N_{\mathcal{E}}(\{x \in \mathbb{R}^{q} \mid \|x\|_{2} \in \mathbb{B}^{2}) = O((\mathbb{R}^{q}))$$



From covering number to a uniform convergence bound

$$\begin{split} & \sum_{f \in \mathcal{F}} \left\| \sum_{i=1}^{\infty} (G_{i} \circ_{i}) - \sum_{i=1}^{n} f_{i} \right\|_{H_{h}^{n}-1} \lesssim \frac{S_{h}}{f_{e}} \left\| \sum_{f \in \mathcal{F}} (S_{i} \circ_{i}) - \sum_{i=1}^{n} f_{e} + \frac{S_{h}}{f_{e}} \right\|_{H_{h}^{n}-1} \lesssim \frac{S_{h}}{f_{e}} \left\| \sum_{f \in \mathcal{F}} (S_{i} \circ_{i}) - \sum_{i=1}^{n} f_{e} + \frac{S_{h}}{f_{e}} \right\|_{H_{h}^{n}-1} \lesssim \frac{S_{h}}{f_{e}} \left\| \sum_{f \in \mathcal{F}} (S_{i} \circ_{i}) - \sum_{i=1}^{n} f_{e} + \frac{S_{h}}{f_{e}} \right\|_{H_{h}^{n}-1} \lesssim \frac{S_{h}}{f_{e}} \left\| \sum_{f \in \mathcal{F}} (S_{i} \circ_{i}) - \sum_{i=1}^{n} f_{e} + \frac{S_{h}}{f_{e}} \right\|_{H_{h}^{n}-1} \lesssim \frac{S_{h}}{f_{e}} \left\| \sum_{f \in \mathcal{F}} (S_{i} \circ_{i}) - \sum_{i=1}^{n} f_{e} + \frac{S_{h}}{f_{e}} \right\|_{H_{h}^{n}-1} \lesssim \frac{S_{h}}{f_{e}} \left\| \sum_{f \in \mathcal{F}} (S_{i} \circ_{i}) - \sum_{i=1}^{n} f_{e} + \frac{S_{h}}{f_{e}} \right\|_{H_{h}^{n}-1} \lesssim \frac{S_{h}}{f_{e}} \left\| \sum_{f \in \mathcal{F}} (S_{i} \circ_{i}) - \sum_{i=1}^{n} f_{e} + \frac{S_{h}}{f_{e}} \right\|_{H_{h}^{n}-1} \lesssim \frac{S_{h}}{f_{e}} \left\| \sum_{f \in \mathcal{F}} (S_{i} \circ_{i}) - \sum_{i=1}^{n} f_{e} + \frac{S_{h}}{f_{e}} \right\|_{H_{h}^{n}-1} \lesssim \frac{S_{h}}{f_{e}} \left\| \sum_{f \in \mathcal{F}} (S_{i} \circ_{i}) - \sum_{i=1}^{n} f_{e} + \frac{S_{h}}{f_{e}} \right\|_{H_{h}^{n}-1} \lesssim \frac{S_{h}}{f_{e}} \left\| \sum_{i=1}^{n} (S_{i} \circ_{i}) - \sum_{i=1}^{n} f_{e} + \frac{S_{h}}{f_{e}} \right\|_{H_{h}^{n}-1} \lesssim \frac{S_{h}}{f_{e}} \left\| \sum_{i=1}^{n} (S_{i} \circ_{i}) - \sum_{i=1}^{n} (S_$$
 $\leq \sup_{f} || \mathbb{Z} \varphi_{g} \in \mathbb{Z} [f - \mathbb{Z})||_{r} + \sup_{f} || \mathbb{Z} \varphi_{s} \in \mathbb{Z} [f + \mathbb{Z})||_{r}$ $\frac{2\varepsilon}{\varepsilon_{i}} = \frac{1}{\varepsilon_{i}} = \frac{1}{\varepsilon_{i}} = \frac{1}{\varepsilon_{i}}$ + Sup || 5 \$ \$ 2 T F ||_-1 $\| \geq \phi_i \|_{r^{1-1}}$ 1. apply printuise result for fixed f $\leq \sum_{i} \sqrt{\phi_{i}^{T} \Lambda^{-i} \phi_{i}}$ $|\xi|| \leq 2$ 2. apply chiam bound 7 14-FILSE $N \ge \overline{\phi_i^{-1} \Lambda^{-1} \phi_i}$ 14

Final notes about linear MDPs

- A semi-parametric model
 - The number of parameters to describe the model can be exponentially large: dS describe M^{\star}
 - Efficient algorithm with regret independent to S
- Still very strong assumption on the feature map
 - Interesting open problems:
 - Representation learning ϕ is unknow

- Nonlinear parametric models
- Suboptimal rates when naively applying to the tabular case d=S GUT $Gut \int ST$

Remainder of the lecture

- Introduction to offline reinforcement learning
- Off-policy evaluation in contextual bandits

Recap: RL is among the hottest area of research in ML!





An RL agent learns **interactively** through the **feedbacks** of an environment.



- Learning how the world works (dynamics) and how to maximize the long-term reward (control) at the same time.

Applications of RL in the real life

- RL for robotics.
- RL for dialogue systems.
- RL for personalized medicine.
- RL for self-driving cars.
- RL for new material discovery.
- RL for sustainable energy.
- RL for feature-based dynamic pricing.
- RL for maximizing user satisfaction.
- RL for QoE optimization in networking

• .

Challenges of Reinforcement in the real life

- No access to a simulator
- Every data point is costly.
- Legal, safety issues associated with exploration.
- Large / complex state-space, action space.
- Long horizon
- Limited adaptivity (cannot run too many iterations)

From an Applied ML Scientist point of view, the starting point of a project is often:

۲	00													ľ	An	amia_l	igos.x	ls															\bigcirc
2				D C	1	5	· ମ	Σ.	A	J.	Z.J	E			1009	6 - (?																
New	Open	Save	Print Import	Copy Pas	te Format	Undo	Redo	AutoSur	n Sort.	A-Z S	ort Z-A	Gall	ery To	olbox	Zoo	om H	lelp																
											Sh	eets		Charts	_	Sma	tArt G	aphics	-	Word	dArt												
0	A RESP ID	B C	D	T2 S	F G H	STU F	J K PRO NBV	L M NBI NBPS	NBPB M	O I MB EN	P Q	R :	S T	U R_A	V W	X Y	POI	AA PMAX P	AB /	AC A	D AI BON	E AF	AG NB_c2	AH /		AK L_C	AL A	M Al	AO \$8_10	AP sem10_12	AQ sem12_14	AR sem14_16	AS sem16_1
2	501 3 502 3	888 388 95 395	22/09/11 00:00 23/09/11 00:00	03:16:00 22:44:00	1 1	21 1	1 10	10 1	1 1	4	0 1	3	4	7 14	3 4	4 1 16 1 4 16	49,1	54,5 75	45,4 54	52,3 1 40 23	9,1 5	61	0 5	1	0 0	0	0	0	0 0				1
4	503 3 504 4	19 399 399	24/09/11 00:00	20:16:00	1 1	22 1	1 30	20 50	0 5	1	0 3	9	8	8 19	6 4	4 5 17	88	92 195	61,5	63 45	30 53	68 68	9 3	1	0 1	0	0	1	0 0		1		1
0	505 4	21 421	03/10/11 00:00	22:41:00	1 0	24 1	0 2	18 10	2	3	0 1	6	6	5 17	3 1	1 2 16	60	90 54.4	52 42.2	56 2 43.9 1	1,4	63	4 12	1	0 0	0	0	0	0 0				1
8	507 4	23 423	03/10/11 00:00	23:47:00	1 1	16 1	0 35	100 10	0 2	2	0 2	3	2	4 .	4 1	1 3 16	124	138	108	105	20 1	130	7 7	1	1 0	0	0	0	0 0		1 1	Ľ.	1
IU	509 4	33 433	04/10/11 00:00	22:54:00	1 1	17 0	0 2	5.		3	0 2	1	1	3 15	5 1	1 3 15	2 59	65	59	45 .	50 1	52	5 2	1	0 0	0	0	1	0 0		1.		
12	510 4	36 436 34 434	04/10/11 00:00	23:46:00	1 0	28 1	0 3	30 .	1	6	0 4	6	7	9 21	2 4	4 2 19	48,5	115	46 55	70	18	87	3 16	1	1 0	0	1	0	0 0		1		1
13	512 4 513 4	39 439 41 441	05/10/11 00:00	00:41:00 01:16:00	1 1	17 1 17 1	0 .	3 3	3 3	4	0 5	3	3	4.	9 1	1 3 16 2 3 15	51 52	53 60	44	40 . 39	21	45 52 3	0 3 27 0	1	0 0	0	0	1	0 0			1	1
10	514 4 515 4	44 444	05/10/11 00:00	01:28:00 01:53:00	1 1	17 1 19 1	0 50	7 2	2 2	4	0 4	6	2	8 16 6 13	5 2	2 3 16	63,5	66,6 52	56,6 36	47,6 2	3,3 50 5,6 4	6,2	9 12	1	1 0	1	0	1	0 1		1 1	×	1
17	516 4 517 4	48 448 50 450	05/10/11 00:00	02:15:00 03:34:00	1 1	18 1 24 0	0 1	1 1	1 0	2	0 2	4	4	7 17 7 21	4 1	1 2 16 2 3 17	60 65	64 67	52 54	40 54 2	23 2,4	50 65	0 10 9 5	1	1 0	0	0	0	0 0		1 1	-	-
19	518 4	57 457	05/10/11 00:00	03:54:00	1 1	21 1	0 6	10 2	2 1	1	0 2	4	5	8 18	4 2	2 2 16	52	66 57	51	45 2	0,4	52	5 4	1	0 0	0	1	0	1 0		1		a.,
21	520 4	61 461	05/10/11 00:00	12:27:00	1 1	18 1	0 15	5 8	3 2	3	0 1	1	1	2.	5 1	1 1 14	25	30	23	24 .	26 4	33	6 5	1	0 0	0	0	0	0 0			3	
23	522 4	465 465	05/10/11 00:00	14:41:00	1 1	19 1	0 3	70 6	3 3	6	0 3	7	11	9.	6	1 9 16	77	101	44	31 3	4,8	46	3 3	1	1 0	o	0	1	0 0		1		1
24	524 4	75 475	05/10/11 00:00	19:55:00	1 1	17 1	0 9	3 1	2 2	4	0 1	3	3	5 16	7	1 2 15	47,6	63	43,1	40 2	9,2 2,4	40	4 2	1	0 0	0	0	0	0 0		1 1	1	
20	527 4 528 4	479 479 483 483	05/10/11 00:00	22:13:00	1 1	32 0 18 1	1 8	5 4	1 1	4	0 2	3	5	5 16	6 2	2 6 16	2 55 63	64 65	48	46 2 50	23	54 56	8 12 6 2	1	0 0	0	0	0	0 0		1	1	10
28	529 4 530 4	485 485 480 480	05/10/11 00:00	22:41:00 22:55:00	1 0	22 1 18 0	0 4	8 3	3 1	9	0 3	5	4	9 15 8 14	8	1 3 17. 1 7 17	e 65 77,3	75 82,7	60 43	50 . 45 2	4,4	70 68	6 5 6 3	1	1 0	0	1	0	0 0		1.		
30	531 4 532 4	184 484 194 494	05/10/11 00:00 06/10/11 00:00	23:41:00 04:09:00	1 1	16 1 21 1	1 300	700 10	3 2	2	0 1	5	6	9. 7 18	6 1	1 3 15 3 5 15	49 6 60	52 66	44 39	40 1	9,4	54 3	2 16	1	0 0	0	0	0	1 (1	
52	533 4	95 495 00 500	06/10/11 00:00	05:02:00	1 1	17 1	1 3	7 !	5 3	2	0 4	2	6	8. 5. 18	5 2	2 2 15	49,5	54	46	40 2	1,3	52 50	5 5	1	1 0	1	0	1	0 0			7	1
34	535 4	99 499	06/10/11 00:00	12:55:00	1 1	19 1	1 8	20 5	5 5	3	0 4	1	1	6 16	4 1	1 2 17	65	81	60	58 .	57	73	4 4	1	0 1	1	1	0	0 0				
30	537 5	03 503	06/10/11 00:00	16:34:00	1 1	26 1	0 2	2 !	5 3	1	0 1	6	1	8 16	2 2	2 1 16	44,4	50,8	38	44,4 .	50	0,8	2 12	1	0 0	0	0	0	0 0		. 1		1
30	538 5	06 506 09 509	06/10/11 00:00	16:40:00	1 1	26 1	0 30	3 10	0 5	1	0 3	2	2	8.14	2 1	2 3 17	44	97 54	37	40 1	3,5 5,8	52 3	4 2	1	1 0	0	0	1	0 0		-	1	
39 40	541 5 542 5	i12 512 i14 514	06/10/11 00:00 06/10/11 00:00	21:24:00 22:35:00	1 0	24 0	1 50	15 20 800 5	0 10 5 3	4	0 2	3	5	6 11 5 16	6 1	1 9 .	40	90 64	30 47	46 1	45 9,5	26 19	9 4 32 12	1	0 1	0	0	0	0 0			1	
41 42	543 5 544 5	18 518 21 521	07/10/11 00:00 07/10/11 00:00	01:01:00 09:35:00	1 1	17 1 27 0	1 40 0 yes	250 8 /es yes	3 6 yes	4	0 1	4	4	5.5	6 1	1 2 16	56,7 68	58,96 74	53,1 : 39	52,16 2 43 3	1,9 57, 0,3 no	,15	2 12 4 5	1	0 0	0	0	0	1 0				1
43	545 5	522 522	07/10/11 00:00	10:52:00	1 1	42 0	1 3	20 5	5 3	1	0 2	2	3	6.	5 2	2 3 16		67	43	47 .		55	4 8	1	0 1	0	0	0	0 1			a.	
45	547 5	511 511	07/10/11 00:00	13:37:00	1 1	27 0	1 20	10 5	5 2	2	0 2	5	4	6.	5 2	2 4 17	67	70	64	60 .		60	2 2	1	0 1	0	0	0	0 0		1 1		1
4/	549 5	27 527	08/10/11 00:00	07:11:00	1 1	23 0	0 2	1 1	1,	2	0 1	6	4	7 15	5 3	3 5 16	65	80	47	52 .	0,1	52	7 8	1	0 0	0	0	0	0 0				1
48	550 5 551 5	i29 529 i30 530	08/10/11 00:00 08/10/11 00:00	09:10:00	1 1	20 1 26 0	0 10 8	8 2	2 1	3	0 2	4	5	8. 6 18	6 2	2 5 16	5 70,3 5 49	95,2 62	68,2	57 2 40 1	5,4 5,9	68 60	7 19	1	0 0	0	0	1	0 0		1		1
20	552 5 553 5	i41 541 i42 542	09/10/11 00:00	16:33:00 17:45:00	1 1	18 1 18 1	1 7	5 5	5 3	4 4	0 1	1	2	3 13 5 .	4 1	1 2 16	55,8	65,77 87	51,3 -	45,35 1 50 2	9,9 45, 7,9	60	5 4 6 4	1	0 0	0	0	0	0 0		1. 1.	1	1
52	554 5 555 5	i45 545	10/10/11 00:00 10/10/11 00:00	02:08:00 12:13:00	1 1	17 1	1 7	5 1	1.	5	0 1	4	6	5. 6 18	6 1	1 3 16	52	57,2	50 96	45 1 96	8,5	58 50	6 12	1	0 0	0	0	0	0 0				1.
54	556 5	i54 554	10/10/11 00:00	21:42:00	1 0	36 0	1 20	40 10	0 5	2	0 2	4	2	5 16	4 3	3 3 19	70	75	62	65 37 1	20	80 48	2 14	1	0 1	0	0	0	0 1		1 1		1.
50	558 5	63 563	12/10/11 00:00	17:42:00	1 1	16 1	0 10	5 4		3	0 1	4	4	8 .	4 2	2 3 17	69	70	64	55 2	2,5	65	0 1	1	0 0	0	0	0	0 0				1
58	560 5	69 569	17/10/11 00:00	16:19:00	1 1	17 1	0 yes	/es 4	4 3	6	0 1	4	7	9 16	7	1 4 15	5 50,3	63,5	49	30	21 1	120	5 10	1	0 0	0	0	0	1 0		1		1
29	561 5 562 5	573 573 576 576	20/10/11 00:00 22/10/11 00:00	23:21:00	1 1	16 1 16 1	0 1	20 5	5 1	6	0 2	3	2	5 15 9 15	6 7	1 4 16	94 63,6	94 70,5	76 59,1	45 3	3,6 53-68 24 56	8 6,8 :	7 5	1	1 0	0	0	0	0 0			1	
01	563 5 564 5	677 577 686 586	22/10/11 00:00 22/10/11 00:00	17:33:00 19:41:00	1 1	18 1 21 1	0 3 1 Yes	6 t res No	5 3 No	4	0 2	6	10	6. 8 16	3 2	2 1 16	45	50 116	42	45 1 49 2	7,7 4,3	48 70	6 3 2 5	1	1 0	0	0	0	0 0				1 .
03	565 5 566 5	689 589 192 592	23/10/11 00:00 23/10/11 00:00	00:44:00	1 1	24 1 28 0	1 20	50 5	5 3	1	0 5	8	9	7 16	6 2	1 2 16	50	54	46	44 1	9,3	51 60	2 21	1	1 1	0	1	1	0 1		1 1		1
05	567 5	96 596 00 600	23/10/11 00:00	18:52:00	1 1	17 1	0 30	40 10	5 3	3	0 4	4	6	5 15	4 1	1 2 16	49,9	61,7	43,1	45 1	8,3 5	4,4	6 6	1	1 0	1	1	0	0 0		1.	1	
0/	569 5	125 525	24/10/11 00:00	01:13:00	1 1	19 0	0 3	4 5	5 6		0 1	1	2	4 .	8 2	2 4 15	60	80	50	40	23	50	1 1	1	0 0	0	0	0	0 0			-	
69	571 6	602 602	24/10/11 00:00	01:39:00	1 1	16 1	0 2	1 2	2 1		0 1		1	3.	9 4	4 4 14	100	120	90	40 n/a	20 ?		5 3	1	0 0	0	0	0	0 0			1	
/1	572 6	08 608 10 610	24/10/11 00:00 24/10/11 00:00	15:04:00 15:50:00	1 1	35 0 22 1	1 yes 1	10 t	yes 5 1	4	0 4	4	4	o 14 7 16	5 3	3 5 16 2 4 17	77	91 85	61 58	64 58 2	5,9	72	3 5 4 5	1	0 1	0	1	0	0		1 21 1	1	
12	574 6	611 611	24/10/11 00:00	16:19:00	1 1 Eeuil3	19 1	0 4	2 2	2 1		0 4	4	1	6 16	5 2	2 9 15	50	50	40	40 .		43	8 2	1	1 (0	1	1	0 0				1
Trees.						4																											

Online RL vs Offline RL



Exploration is often **expensive**, **unsafe**, **unethical** or **illegal** in practice, e.g., in self-driving cars, or in medical applications.

Can we learn a policy from already **logged interaction data**?

Off-Policy learning: an example



- How to evaluate a new algorithm without actually running it live?
- How to learn a better system than the one that is deployed.

Offline Reinforcement Learning, aka. Batch RL

• Task 1: Offline Policy Evaluation. (OPE)



• Task 2: Offline Policy Learning. (OPL)



Contextual bandits model

- Contexts: State \checkmark drawn iid, possibly infinite domain $x_1,...,x_n\sim\lambda$
- Actions:

* $a_i \sim \mu(a|x_i)$ Taken by a randomized "Logging" policy

• Reward:

$$\cdot r_i \sim D(r|x_i, a_i)$$

Revealed only for the action taken

- Value: • $v^{\mu} = \mathbb{E}_{x \sim \lambda} \mathbb{E}_{a \sim \mu(\cdot|x)} \mathbb{E}_D[r|x, a]$
- We collect data $(x_i, a_i, r_i)_{i=1}^n$ by the above processes.

Off-policy Evaluation and Learning



- Using data $(x_i, a_i, r_i)_{i=1}^n$ $(x_i, a_i, r_i)_{i=1}^n$ $\mathcal{M}(a_i, x)$ often the policy μ or logged propensities $(\mu_i)_{i=1}^n$

ATE estimation is a special case of off-policy evaluation

- a: Action \Leftrightarrow T: Treatment {0,1}
- r: Reward \Leftrightarrow Y: Response variable
- x: Contexts \Leftrightarrow X: covariates

Direct Method / Regression-estimator

• Fit a regression model of the reward

$$\hat{r}(x,a) pprox \mathbb{E}(r|x,a)$$
 using the data

• Then for any target policy

$$\hat{v}_{\text{DM}}^{\pi} = \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \hat{r}(x_i, a) \pi(a | x_i)$$
Pros: Cons:

- Low-variance.
- Can evaluate on unseen contexts

- Often high bias
- The model can be wrong/hard to learn

Inverse propensity score / Importance sampling (Horvitz & Thompson, 1952) Importance weights

Pros:

- No assumption on rewards
- Unbiased
- Computationally efficient

Cons:

 High variance when the weight is large

Analyzing the performance of importance sampling estimator

Importance Sampling and Direct Method are surprisingly similar in some cases

• Consider the MAB case

Next lecture: OPE for reinforcement learning

Importance sampling

• Marginalized importance sampling