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 ?
 - 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...

(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:

 $r_h(s,a) = \theta_h^\star \cdot \phi(s,a), \quad P_h(\cdot|s,a) = \mu_h^\star \phi(s,a), \forall h$

(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

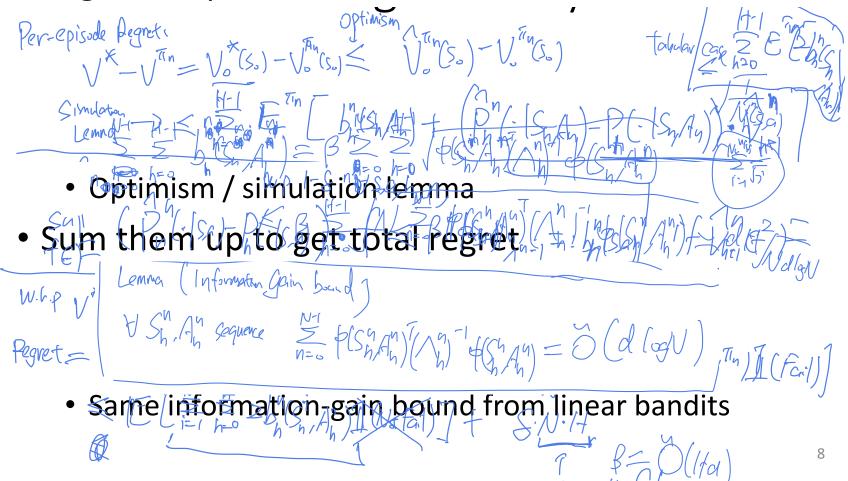
• Choose
$$\beta = Hd \left(\sqrt{\ln \frac{H}{\delta}} + \sqrt{\ln(W+H)} + \sqrt{\ln B} + \sqrt{\ln d} + \sqrt{\ln N} \right)$$
$$\lambda = 1$$

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

• Regret $\tilde{O}\left(H^2\sqrt{d^3N}\right)$

Recap: Regret analysis

• Regret of episode t



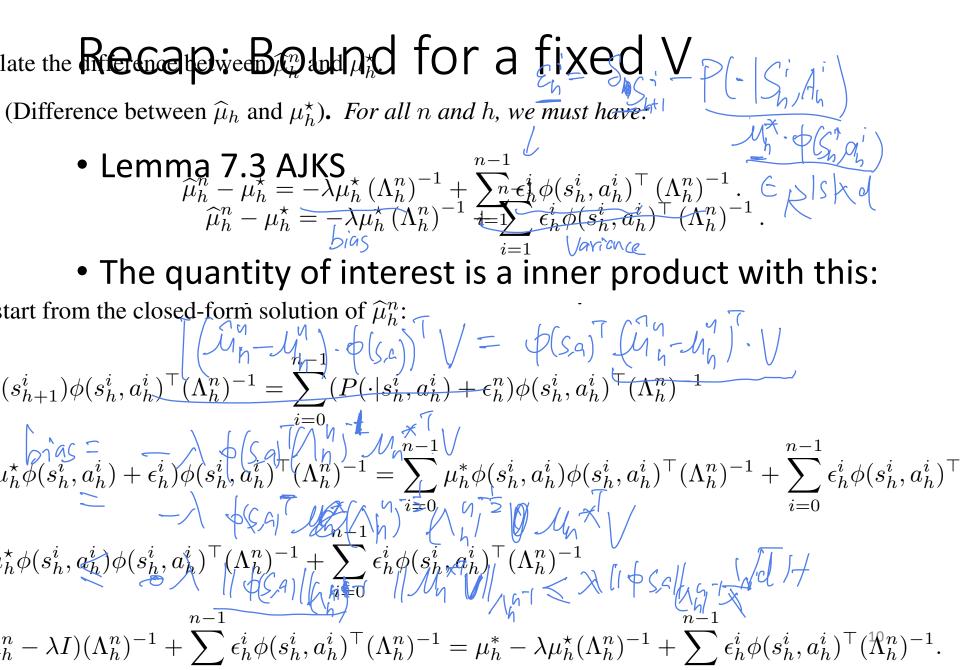
Recap: It remains to prove

- 1. Uniform convergence bound
- 2. "Optimism"

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 dependent



Challenge: we cannot use union bound because we have an infinite number of value functions

• A covering number argument.

 Covering number: the number of balls with radius ε that is needed to cover all points in a set.

Family of value functions we consider

$$f_{w,\beta,\Lambda}(s) = \min\left\{\max_{a} \left(w^{\top}\phi(s,a) + \beta\sqrt{\phi(s,a)^{\top}\Lambda^{-1}\phi(s,a)}\right), H\right\}, \forall s \in \mathcal{S}.$$

$$\mathcal{F} = \{ f_{w,\beta,\Lambda} : \|w\|_2 \le L, \beta \in [0,B], \sigma_{\min}(\Lambda) \ge \lambda \}.$$

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?

Covering number calculations

From covering number to a uniform convergence bound

Final notes about linear MDPs

- A semi-parametric model
 - The number of parameters to describe the model can be exponentially large: d S
 - Efficient algorithm with regret independent to S
- Still very strong assumption on the feature map
 - Interesting open problems:
 - Representation learning
 - Nonlinear parametric models
 - Suboptimal rates when naively applying to the tabular case

Remainder of the lecture

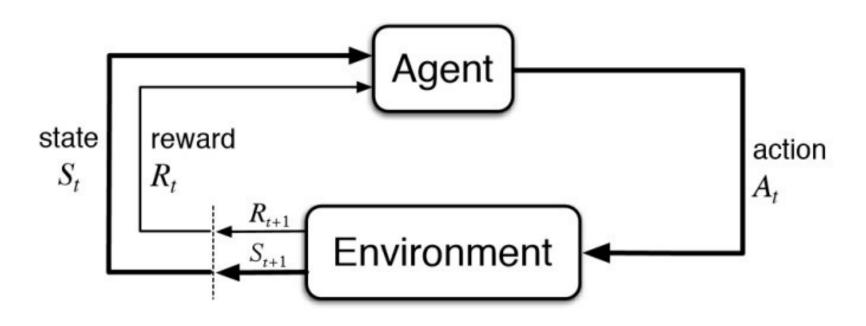
- Introduction to offline reinforcement learning
- Off-policy evaluation in contextual bandits

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

• .

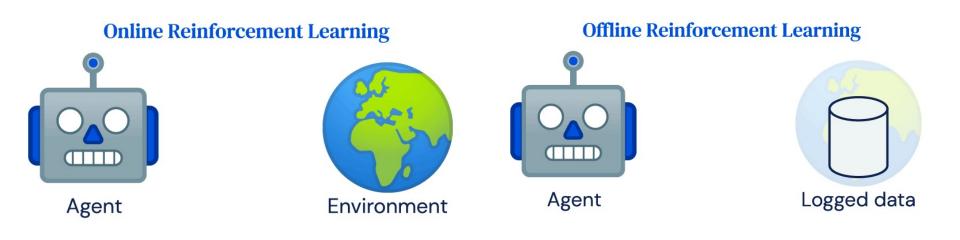
Challenges of Reinforcement in the real life

- 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:

00	Anamia_Egos.xls															\bigcirc																		
9 R			h È		3	0 - 0	M +	2	-	A.	Z				1	00%																		
		int Import																																
new ope	IT DAVE IT	int import :	copy rase	ic ron	inat : O	nuo i	teuo :	Auto.		1.4.2		heets	IICTY	Cha					phics		WordA	et				_								
♦ A	BC	D	E F	G	H		K	L	M N	0	PO	R	S	T	UV	wx	Y	ZA	AAB	AC	AD	AE	AF	AG	AH A		AK	AL A	MA	N AO	AP	AQ	AR	AS
1 RESP	ID MOD LA 388 388	STMODIF 22/09/11 00:00	T2 ST	SEX 1	AGE S	TU PRO	NBV	NBI NI 10	BPS NBPB	MB E	ENF L_T	w_t w	0_t IN	TR	A D	C 0	TAL PC	DI PM	AX PMIN 54,5 45,4	PJUS	IMC	BONP 57,2	NB_c1	NB_c2	UH L	E L_W	L_C	LPL	S \$6_1	8 s8_10	sem10_12	sem12_14	sem14_16	sem16_1
3 502 4 503	395 395	23/09/11 00:00 24/09/11 00:00	22:44:00	1 1	24	1	1 10	30	30 . 50 .	2	0	4 6	6	6	14 E	5 1 4 5 4 5	165	64	75 54	4 4	0 23,62	2 61	7	9	1	0	1 0	1	1	0	0.			1
505 504 505	419 419	03/10/11 00:00	19:02:00 22:41:00	1 1	27	0	0 4	5	3.	5	0	3 3	5	9	18 5	9 1 7	175	154	90 5	4 4	5 53	3 68	9	3	1		0 1	1	0	1	0.	1		
/ 506	422 422	03/10/11 00:00	23:04:00	1 1	24	1	0 1	18 20	3 3	3 2	0	1 4	6	6	19 6	3 1 2 5 1 3	165	50,4 5	54,4 42,2	2 43,9	9 18,5	5 50,67	7	12	1	0	0 0	0 0	0	0	0.	1		1
8 507 9 508	426 426	03/10/11 00:00 04/10/11 00:00	14:10:00	1 1	16 37	0		100	10 2	2 2	1	2 3	2	4 . 5	17 8	8 2 5		106	118 63	3 5		3 125	1	4	1	0	0 0	0 0	1	0	0.	1		1
1U 509 11 510	436 436	04/10/11 00:00 04/10/11 00:00	22:54:00 23:46:00	1 1	17 21	0	0 2	5.	2	3	0	2 1 4 8	1	3	15 5	5 1 3 6 1 4	152 162 195	59 48,5	65 59 50 46	6 3	5.	52		2	1	0	0 0	0 0	1	0	0.		-	
12 511 13 512	434 434 439 439	04/10/11 00:00 05/10/11 00:00		1 0	28	1		30 .	3 3	6	0	3 6	7	9	21 2	2 4 2	195	65 51	115 55 53 44			8 87			1	1	0 0	1	0	0	0.			1
14 513 15 514	441 441	05/10/11 00:00	01:16:00	1 1	17	1	0 15	2	10 .	3	0	2 3	3	4 .	4	4 2 3	157	52	60 48	8 3	9 21	1 52	27	0	1	0	0 0	0	1	1	0.			1.
10 515	447 447	05/10/11 00:00		1 1	19	1	0 50	99	20	1	0	3 6	2	6	13 4	4 1 1	160	40	52 36	6 3	4 15,6	47,3	30	14			0 1	0	0	0	1	1		
17 516 18 517	450 450	05/10/11 00:00 05/10/11 00:00	03:34:00	1 1	24	0		25	7 3	3 3	0	2 4 3 6	4	7	21 4	4 1 2 4 2 3	170	65	64 53 67 54	4 5	4 22,4	65	5 9	5	1	0	0 0	0	0	0	1	1		1
19 518 20 519	457 457	05/10/11 00:00 05/10/11 00:00	03:54:00 08:19:00	1 1	21	1	1 45	10 25	2 2	1		2 4	5	8 5 .	4	4 2 2	158		66 51 57 46,5	5 4		57	12		1	0	1 1	0 1	0	0	0.	1	-	
∠1 520 ∠∠ 521	461 461	05/10/11 00:00	12:27:00 12:46:00	1 1	18 20	1	0 15	5	8 3	2 3		1 1		2.	19 5	5 1 1		25 61	30 23			33		5	1	0	0 0	0 0	0	0	0.			1
23 522 24 524	465 465	05/10/11 00:00	14:41:00	1 1	19	1		70	6	6	0	3 7	11	9 .	e	5 1 9 4 1 2	161	77 47,6 5	101 44	4 3	1 34,8	3 46	3	3	1	1	0 0		1	0	0	1		1
25 526	475 475	05/10/11 00:00	19:55:00	1 1	17	1	0 9	3	1.	4	0	1 3	3	7.	10 7	7 1 4	157	51	63 5	1 4	0 22,4	40	10	2	1	0	0 0	0 0	0	0	0.			
20 527 27 528	483 483	05/10/11 00:00 05/10/11 00:00	22:23:00	1 1	32	1	0 4	5	2	1	0	2 3	5	5 4 .	6	5 2 6	163	55 63	64 48 65 53	2 5	0 23	3 56	6	12	1	0	0 0	0 0	0	0	0.	1	1	10
28 529 29 530	480 480	05/10/11 00:00 05/10/11 00:00	22:41:00 22:55:00	1 0	22	1		8	3 .	9	0	3 5	4	9	15 E	6 1 3 8 1 7	172	65 77,3 8	75 60	3 4		70	16		1	1	0 0	0 1	0	0	1.	1		
3U 531 31 532	484 484	05/10/11 00:00	23:41:00	1 1	16	1	1 300		10 3	3 2	0	1 5	7	7	18 5	8 1 7 6 1 3 5 3 5	159	49	52 44 66 35	4 4	0 19,4	54	31				0 0	0 0	0	1	0.		1	
52 533 55 534	495 495	06/10/11 00:00 06/10/11 00:00	05:02:00	1 1	17	1	1 3	7	5 3	2	0	4 2	6	8.	1	5 2 2 8 3 9	152	49,5 63	54 46 66 63	6 4	0 21,3	3 52	5	i 5	1	1	0 1	0	1	0	0.		-	
54 535	499 499	06/10/11 00:00	12:55:00	1 1	19	1	1 8	20	5 1	5 3	0	4 1	1	6	16 4	4 1 2	175	65	81 60	0 5	8 .	73	4	4	1	0	1 1	1	0	0	0.	1000 1000		1
35 536 30 537	503 503	06/10/11 00:00	16:34:00	1 1	23 26	1	0 2	yes ye 2	s yes	3 1	0	2 4	1	8	16 6	5 2 4 2 2 1	165 164 175	57,2 11 44,4 5	50,8 38	8 44,4	4 .	57,15 50,8	12	12	1	0	0 0	0 0	0	0	0.	. 1		1
57 238 28 239	509 509	06/10/11 00:00 06/10/11 00:00	16:40:00 18:37:00	1 1	26	1	0 30	3	2 2	2 5	0	3 2	2	8.	14 2	6 2 3 2 1 1	175	72	97 57 54 37		6 23,5 0 15,8				1	1	0 0	0 0	1	0	0.			14 A
39 541 40 542		06/10/11 00:00 06/10/11 00:00	21:24:00 22:35:00	1 0	24	0	1 50 1 300	15	20 10	4	0	2 3	1	6	11 4 16 6	4 1 9 6 1 2			90 30 64 47						1	0	1 0	0 0	0	0	0.		1	
41 543 4∠ 544	518 518	07/10/11 00:00 07/10/11 00:00	01:01:00 09:35:00	1 1	17	1	1 40	250 ves ve	8 (3 4	0	1 4	4	5.	2	2 1 2	160	56,7 58	8,96 53,1 74 39	1 52,1		57,15		12	1	0	0 0	0	0	1	0.	· · · · · ·		
43 545	522 522	07/10/11 00:00	10:52:00	1 1	42	0	1 3	20	5 3		0	2 2	3	6.		5 2 3	160 .		67 43	3 4	7.	55	6 4	8	1	0	1 0	0 0	0	0	1.			
44 546 45 547	511 511	07/10/11 00:00 07/10/11 00:00	13:21:00 13:37:00	1 1	22 27	0	1 20	25 10	5 3	3 4 2 2	0	1 1 2 5	4	3 6 .	1	2 1 2 5 2 4	170	67	80 50 70 64	4 6	0.	50	2	2	1	0	0 0	0 0	0	0	0.	1		1
40 548 47 549	527 527 528 528	08/10/11 00:00 08/10/11 00:00	03:33:00 07:11:00	1 1	24	1		15	4 3	2 1	0	5 8	7	9		5 2 4 5 3 5	163	65 65	78 60 80 47			55		5	1	1	1 0	0 1	1	0	1 .	1		1
48 550 49 551	529 529	08/10/11 00:00 08/10/11 00:00		1 1	20	1	0 10	8	2	3	0	2 4	4	8 .	E	8 2 5	165	70,3 9	95,2 68,2 62 4	2 5	7 25,4	68	15		1	-	0 0		1	-	0.	1		
DU 552	541 541	09/10/11 00:00	16:33:00	1 1	26	1		20 5	10 4	1		2 5	5	3	18 2	2 1 2	1/5	55,8 65	5,77 51,3	3 45,3	5 19,9	45,35	5 5	19	1	0	0 0	0 0	0	0	0.	- 1		1
51 553 52 554	545 545	09/10/11 00:00 10/10/11 00:00	17:45:00 02:08:00	1 1	18	1		4	4	4		4 4	4	5.	4	4 1 3 6 1 3	167	52 5	87 72 57,2 50	0 4	5 18,5	5 58	16	4	1	0	0 1	0	0	0	0.	10. 10	1	1 .
33 555 34 556	546 546	10/10/11 00:00 10/10/11 00:00	12:13:00 21:42:00	1 1	24	1	0 5	40	3 10	3		2 5	4	6 5	18 7	7 1 2	149 190	102 70	129 96 75 62					0	1	1	0 0	0 0	0	0	1	- 1		1 .
557 50 558	555 555	10/10/11 00:00	22:30:00	1 1	16 16	1	0 50	10	3	2	0	2 5	9	7.	4	4 1 2	158	44	50 38,5 70 64	5 3	7 17,7	48	18	3 1	1	1	0 0	0 0	0	0	0			1
50 559 57 559 58 560	561 561	12/10/11 00:00	18:57:00	1 1	22	1	1 7	100	4 3	2 2	0	3 5	4	6	17 4	4 1 3	165	67	70 57	7 5	1 24,5	5 57	11	6	1	0	1 0	0	1	0	1	÷		1
59 561	573 573	20/10/11 00:00	23:21:00	1 1	16	1	0 yes 0 1	10 .	4 3	5	0	2 1	5	9 5	15 6	8 1 4	155 168	94	94 78	6 4	5 33,6	53-68	7	5	1	0	0 0	0 0	1	0	0.	4 		
6U 562 61 563	577 577	22/10/11 00:00 22/10/11 00:00	17:28:00 17:33:00	1 1	16 18	1	0 5	20	5 5	6	0	2 3	2	9	15 7	3 2 1	161	45	70,5 59,1 50 43	2 4	5 17,7	48	6	4	1	1	0 0	0 0	0	0	0.		-	
62 564 55 565	586 586	22/10/11 00:00 23/10/11 00:00	19:41:00 00:44:00	1 1	21	1	1 Yes 1 20	Yes No 50	0 No	5	1	2 6	10	8		5 1 3	173	72	116 57 54 46			3 70	2	21	1	0	0 0	0 0	1	0	0.	1		1.
04 566 05 567	592 592	23/10/11 00:00 23/10/11 00:00	10:50:00	1 1	28	0	1 6	4	4 10	1	0	1 2	3	7.	6	8 2 6	151	70	77 .			60	4	4	1	0	0 0	0	0	0	0.			1
DD 568	600 600	24/10/11 00:00	00:23:00	1 1	19	1	0 9	2	5 3	3 4	0	3 1	1	4	13 5	5 1 3	141	54	55 50	0 4	0.	50	5		1		0 1	0	1	0	0.		7	
67 569 570	601 601	24/10/11 00:00 24/10/11 00:00	01:13:00 01:22:00	1 1	19 23	1	0 8	4	5 6	2 2	0	1 1 2 1	2	5	18 7	8 2 4 7 2 1	151	60	80 50 55 50	0 4	0 23	3 50	1	1	1	1	0 0	0 0	0	0	0.	4 1	-	1
09 571 70 572		24/10/11 00:00 24/10/11 00:00	15:04:00	1 1	16 35	1		1 yes ye	2 Yes	. 4		1 1	1	3 . 6	14 5	9 4 4 5 3 5	163	100 77	120 90 91 61	0 4		?	5		1	0	0 0		0	0	0.			
11 573 12 574	610 610	24/10/11 00:00 24/10/11 00:00	15:50:00	1 1	22	1	1 4	10	5	6	0	4 5	4	7	16 5	5 2 4	170	76	85 58 50 40	8 5		43		5	1	0	1 1		0	0	1	21 1	1	1
		► Feuil		_		_		~	-	1.	*			~				~			-1.					- 1				-1	-1-			

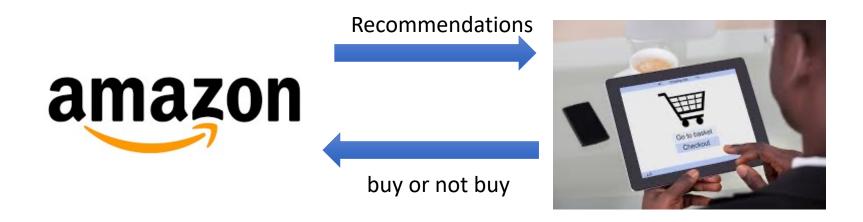
Online RL vs Offline RL



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**?

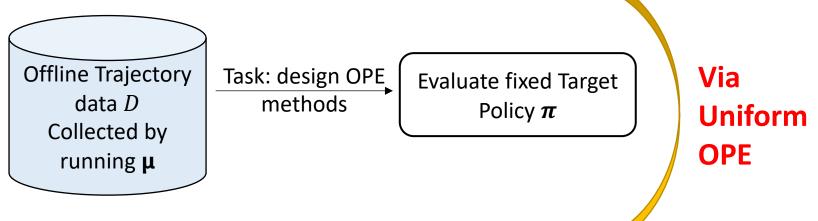
Off-Policy learning: an example



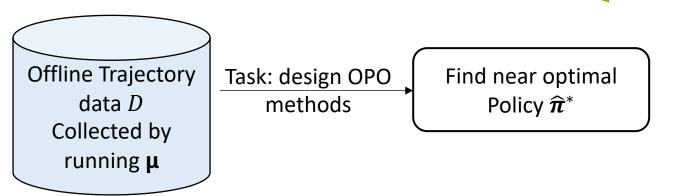
- How to evaluate a new algorithm without actually running it live?
- How to learn a better system than the one that is deployed.

Offline Reinforcement Learning, aka. Batch RL

• Task 1: Offline Policy Evaluation. (OPE)



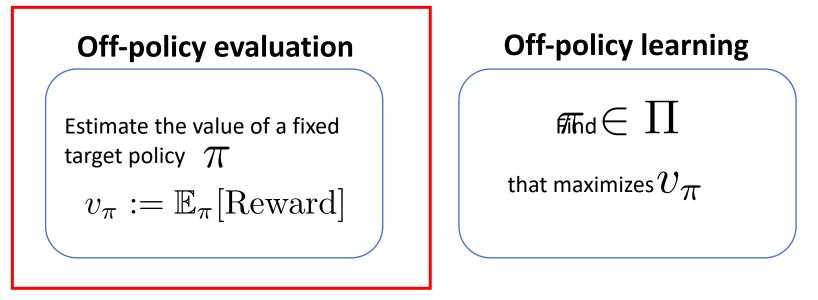
• Task 2: Offline Policy Learning. (OPL)



Contextual bandits model

- Contexts:
 - * $x_1,...,x_n\sim\lambda$ drawn iid, possibly infinite domain
- Actions:
 - * $a_i \sim \mu(a|x_i)$ Taken by a randomized "Logging" policy
- Reward:
 - $r_i \sim D(r|x_i,a_i)$ 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 $(x_i, a_i, r_i)_{i=1}^n$
- 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

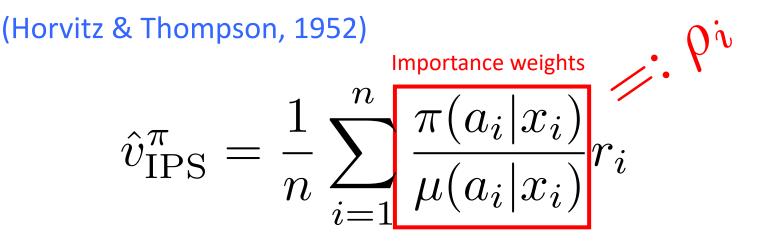
• Then for any target policy

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



Pros:

- No assumption on rewards
- Unbiased
- Computationally efficient

Cons:

 High variance when the weight is large

Next lecture: OPE for reinforcement learning

Importance sampling

• Marginalized importance sampling