Showing Your Offline Reinforcement Learning Work: Online Evaluation Budget Matters

Vladislav Kurenkov, Sergey Kolesnikov

Tinkoff
{v.kurenkov, s.s.kolesnikov}@tinkoff.ai

web-page: tinkoff-ai.github.io/eop/

What is typically reported when presenting new Offline-RL algorithms?

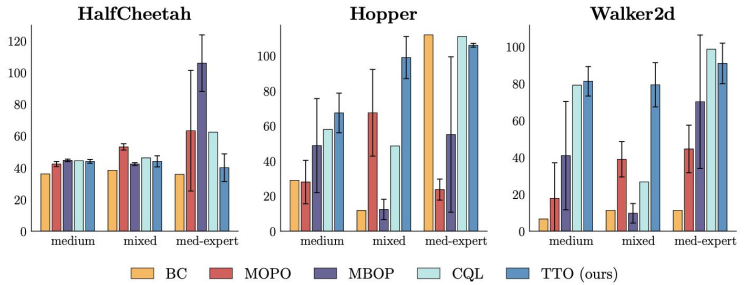
	BC	D4PG	ABM	BCQ	CRR exp	CRR binary	CRR binary max
Cartpole Swingup	386 ± 6	855 ± 13	798 ± 30	444 ± 15	664 ± 22	**860 ± 7**	858 ± 15
Finger Turn Hard	261 ± 39	764 ± 24	566 ± 25	311 ± 38	714 ± 38	755 ± 31	**833 ± 57**
Walker Stand	386 ± 6	**929 ± 46**	689 ± 13	501 ± 5	797 ± 30	881 ± 13	**929 ± 10**
Walker Walk	417 ± 33	939 ± 19	846 ± 15	748 ± 24	901 ± 12	936 ± 3	**951 ± 7**
Cheetah Run	407 ± 56	308 ± 121	304 ± 32	368 ± 129	**577 ± 79**	453 ± 20	415 ± 26
Fish Swim	466 ± 8	281 ± 77	527 ± 19	473 ± 36	517 ± 21	585 ± 23	**596 ± 11**
Manipulator Insert Ball	385 ± 12	154 ± 54	409 ± 4	98 ± 29	625 ± 24	**654 ± 42**	
Manipulator Insert Peg	324 ± 31	71 ± 2	345 ± 12	194 ± 117	**387 ± 36**	365 ± 28	
Humanoid Run	382 ± 2	1 ± 1	302 ± 6	22 ± 3	**586 ± 6**	412 ± 10	

Domain	Task Name	BC	SAC	BEAR	BRAC-p	BRAC-v	CQL($\mathcal{H}$)	CQL(ρ)
AntMaze	antmaze-umaze	65.0	0.0	**73.0**	50.0	70.0	**74.0**	**73.5**
	antmaze-umaze-diverse	55.0	0.0	61.0	40.0	70.0	**84.0**	61.0
	antmaze-medium-play	0.0	0.0	0.0	0.0	0.0	**61.2**	4.6
	antmaze-medium-diverse	0.0	0.0	8.0	0.0	0.0	**53.7**	5.1
	antmaze-large-play	0.0	0.0	0.0	0.0	0.0	**15.8**	3.2
	antmaze-large-diverse	0.0	0.0	0.0	0.0	0.0	**14.9**	2.3
Adroit	pen-human	34.4	6.3	-1.0	8.1	0.6	37.5	**55.8**
	hammer-human	1.5	0.5	0.3	0.3	0.2	**4.4**	2.1
	door-human	0.5	3.9	-0.3	-0.3	-0.3	**9.9**	**9.1**
	relocate-human	0.0	0.0	-0.3	-0.3	-0.3	0.20	**0.35**
	pen-cloned	**56.9**	23.5	26.5	1.6	-2.5	39.2	40.3
	hammer-cloned	0.8	0.2	0.3	0.3	0.3	2.1	**5.7**
	door-cloned	-0.1	0.0	-0.1	-0.1	-0.1	0.4	**3.5**
	relocate-cloned	**-0.1**	-0.2	-0.3	-0.3	-0.3	**-0.1**	**-0.1**
Kitchen	kitchen-complete	33.8	15.0	0.0	0.0	0.0	**43.8**	31.3
	kitchen-partial	33.8	0.0	13.1	0.0	0.0	**49.8**	**50.1**
	kitchen-undirected	47.5	2.5	47.2	0.0	0.0	**51.0**	**52.4**

		BC	BRAC-p	AWAC	CQL	FisherBRC	TD3+BC
Random	HalfCheetah	2.0 ±0.1	23.5	2.2	21.7 ±0.9	32.2 ±2.2	10.2 ±1.3
	Hopper	9.5 ±0.1	11.1	9.6	10.7 ±0.1	11.4 ±0.2	11.0 ±0.1
	Walker2d	1.2 ±0.2	0.8	5.1	2.7 ±1.2	0.6 ±0.6	1.4 ±1.6
Medium	HalfCheetah	36.6 ±0.6	44.0	37.4	37.2 ±0.3	41.3 ±0.5	42.8 ±0.3
	Hopper	30.0 ±0.5	31.2	72.0	44.2 ±10.8	99.4 ±0.4	99.5 ±1.0
	Walker2d	11.4 ±6.3	72.7	30.1	57.5 ±8.3	79.5 ±1.0	79.7 ±1.8
Medium Replay	HalfCheetah	34.7 ±1.8	45.6	-	41.9 ±1.1	43.3 ±0.9	43.3 ±0.5
	Hopper	19.7 ±5.9	0.7	-	28.6 ±0.9	35.6 ±2.5	31.4 ±3.0
	Walker2d	8.3 ±1.5	-0.3	-	15.8 ±2.6	42.6 ±7.0	25.2 ±5.1
Medium Expert	HalfCheetah	67.6 ±13.2	43.8	36.8	27.1 ±3.9	96.1 ±9.5	97.9 ±4.4
	Hopper	89.6 ±27.6	1.1	80.9	111.4 ±1.2	90.6 ±43.3	112.2 ±0.2
	Walker2d	12.0 ±5.8	-0.3	42.7	68.1 ±13.1	103.6 ±4.6	101.1 ±9.3
Expert	HalfCheetah	105.2 ±1.7	3.8	78.5	82.4 ±7.4	106.8 ±3.0	105.7 ±1.9
	Hopper	111.5 ±1.3	6.6	85.2	111.2 ±2.1	112.3 ±0.2	112.2 ±0.2
	Walker2d	56.0 ±24.9	-0.2	57.0	103.8 ±7.6	79.9 ±32.4	105.7 ±2.7
	Total	595.3 ±91.5	284.1	-	764.3 ±61.5	974.6 ±108.3	979.3 ±33.4

	Iterative	One-step			
	Fu et al. [2020]	BC	Easy BCQ	Rev. KL Reg	Exp. Weight
halfcheetah-m	46.3	41.9 ± 0.1	52.6 ± 0.2	**55.2 ± 0.4**	48.4 ± 0.1
walker2d-m	81.1	68.6 ± 6.3	**87.2 ± 1.3**	85.9 ± 1.4	81.8 ± 2.2
hopper-m	58.8	49.9 ± 3.1	74.5 ± 6.2	**83.7 ± 4.5**	59.6 ± 2.5
halfcheetah-m-e	64.7	61.1 ± 2.7	78.2 ± 1.6	**93.8 ± 0.5**	93.4 ± 1.6
walker2d-m-e	111.0	78.5 ± 22.4	112.2 ± 0.3	111.2 ± 0.2	**113.0 ± 0.4**
hopper-m-e	**111.9**	49.1 ± 4.3	85.1 ± 2.2	98.7 ± 7.5	103.3 ± 9.1
etah-m-re	**47.7**	34.6 ± 0.9	38.3 ± 0.3	41.9 ± 0.5	38.1 ± 1.3
2d-m-re	26.7	26.6 ± 3.4	69.1 ± 4.2	**74.9 ± 6.6**	49.5 ± 12.0
-m-re	48.6	23.1 ± 2.7	78.4 ± 7.2	92.3 ± 1.1	**97.5 ± 0.7**
etah-r	**35.4**	2.2 ± 0.0	5.4 ± 0.3	8.8 ± 3.8	3.2 ± 0.1
2d-r	**7.3**	0.9 ± 0.1	3.7 ± 0.1	6.2 ± 0.7	5.6 ± 0.8
-r	**12.2**	2.0 ± 0.1	6.6 ± 0.1	7.9 ± 0.7	7.5 ± 0.4
	56.9	46.9 ± 11.0	**65.9 ± 3.6**	57.4 ± 3.5	60.0 ± 4.1
r-c	2.1	0.4 ± 0.1	**2.9 ± 0.5**	0.2 ± 0.1	2.1 ± 0.7
c	-0.1	-0.1 ± 0.0	**0.3 ± 0.2**	0.2 ± 0.1	0.2 ± 0.1
	0.4	0.0 ± 0.1	**0.6 ± 0.6**	0.2 ± 0.7	0.2 ± 0.3

	BC	BRAC-p	BRAC-v	MBOP	CQL (GitHub)	CQL (Ours)	F-BRC (Ours)
halfcheetah-random	30.5	23.5	28.1	6.3 ± 4.0	27.1 ± 1.3	20.7 ± 0.6	33.3 ± 1.3
hopper-random	11.3	11.1	12.0	10.8 ± 0.3	10.6 ± 0.1	10.4 ± 0.1	11.3 ± 0.2
walker2d-random	4.1	0.8	0.5	8.1 ± 5.5	1.1 ± 2.2	10.0 ± 4.6	1.5 ± 0.7
halfcheetah-medium	36.1	44.0	45.5	44.6 ± 0.8	40.3 ± 0.3	38.9 ± 0.3	41.3 ± 0.3
walker2d-medium	6.6	72.7	81.3	41.0 ± 29.4	77.3 ± 3.8	69.2 ± 8.3	78.8 ± 1.0
hopper-medium	29.0	31.2	32.3	48.8 ± 26.8	42.2 ± 15.5	30.5 ± 0.7	99.4 ± 0.3
halfcheetah-expert	107.0	3.8	−1.1	-	54.4 ± 45.8	103.5 ± 1.3	108.4 ± 0.5
hopper-expert	109.0	6.6	3.7	-	67.7 ± 54.7	112.2 ± 0.2	112.3 ± 0.1
walker2d-expert	125.7	−0.2	−0.0	-	84.7 ± 42.7	107.2 ± 3.8	103.0 ± 5.0
halfcheetah-medium-expert	35.8	43.8	45.3	105.9 ± 17.8	21.7 ± 6.8	58.6 ± 8.7	93.3 ± 10.2
walker2d-medium-expert	11.3	−0.3	0.9	70.2 ± 36.2	104.0 ± 10.1	104.6 ± 10.4	105.2 ± 3.9
hopper-medium-expert	111.9	1.1	0.8	55.1 ± 44.3	111.3 ± 2.1	112.4 ± 0.2	112.4 ± 0.3
halfcheetah-mixed	38.4	45.6	45.9	42.3 ± 0.9	44.9 ± 1.1	42.0 ± 1.1	43.2 ± 1.5
hopper-mixed	11.8	0.7	0.8	12.4 ± 5.8	31.6 ± 3.6	29.0 ± 0.5	35.6 ± 1.0
walker2d-mixed	11.3	−0.3	0.9	9.7 ± 5.3	16.8 ± 3.1	16.5 ± 4.9	41.8 ± 7.9



What is typically reported when presenting new Offline-RL algorithms?

	BC	D4PG	ABM	BCQ	CRR exp	CRR binary	CRR binary max
Cartpole Swingup	386 ± 6	855 ± 13	798 ± 30	444 ± 15	664 ± 22	**860 ± 7**	858 ± 15
Finger Turn Hard	261 ± 39	764 ± 24	566 ± 25	311 ± 38	714 ± 38	755 ± 31	**833 ± 57**
Walker Stand	386 ± 6	**929 ± 46**	689 ± 13	501 ± 5	797 ± 30	881 ± 13	**929 ± 10**
Humanoid Run	382 ± 2	1 ± 1	302 ± 6	22 ± 3	**586 ± 6**	412 ± 10	226

Maximum Performance after Online Policy Selection

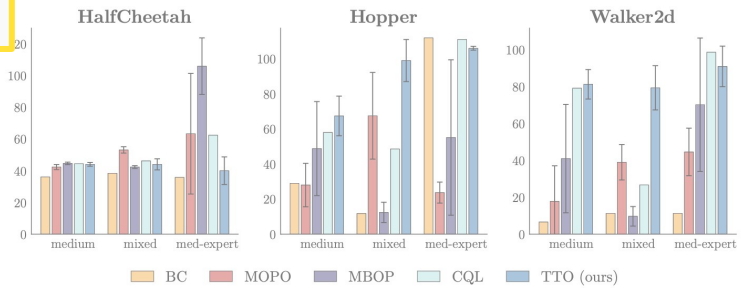
Domain	Task Name	BC	SAC	BEAR	BRAC-p	BRAC-v	CQL($\mathcal{H}$)	CQL(ρ)
AntMaze	antmaze-umaze	65.0	0.0	**73.0**	50.0	70.0	**74.0**	**73.5**
	antmaze-umaze-diverse	55.0	0.0	61.0	40.0	70.0	**84.0**	61.0
	antmaze-medium-play	0.0	0.0	0.0	0.0	0.0	**61.2**	4.6
	antmaze-medium-diverse	0.0	0.0	8.0	0.0	0.0	**53.7**	5.1
	antmaze-large-play	0.0	0.0	0.0	0.0	0.0	**15.8**	3.2
	antmaze-large-diverse	0.0	0.0	0.0	0.0	0.0	**14.9**	2.3
Adroit	pen-human	34.4	6.3	-1.0	8.1	0.6	37.5	**55.8**
	hammer-human	1.5	0.5	0.3	0.3	0.2	**4.4**	2.1
	door-human	0.5	3.9	-0.3	-0.3	-0.3	**9.9**	**9.1**
	relocate-human	0.0	0.0	-0.3	-0.3	-0.3	0.20	**0.35**
	pen-cloned	**56.9**	23.5	26.5	1.6	-2.5	39.2	40.3
	hammer-cloned	0.8	0.2	0.3	0.3	0.3	2.1	**5.7**
	door-cloned	-0.1	0.0	-0.1	-0.1	-0.1	0.4	**3.5**
Kitchen								

Online Budget = #Hyperparameter Assignments

	BC	BRAC-p	AWAC	CQL	FisherB
Random	HalfCheetah 2.0 ±0.1	23.5	2.2	21.7 ±0.9	32.2 ±
	Hopper 9.5 ±0.1	11.1	9.6	10.7 ±0.1	11.4 ±
	Walker2d 1.2 ±0.2	0.8	5.1	2.7 ±1.2	0.6 ±
Medium	HalfCheetah 36.6 ±0.6	44.0	37.4	37.2 ±0.3	41.3 ±
	Hopper 30.0 ±0.5	31.2	72.0	44.2 ±10.8	99.4 ±
	Walker2d 11.4 ±6.3	72.7	30.1	57.5 ±8.3	79.5 ±
Medium Replay	HalfCheetah 34.7 ±1.8	45.6	-	41.9 ±1.1	43.3 ±
	Hopper 19.7 ±5.9	0.7	-	28.6 ±0.9	35.6 ±
	Walker2d 8.3 ±1.5	-0.3	-	15.8 ±2.6	42.6 ±
Medium Expert	HalfCheetah 67.6 ±13.2	43.8	36.8	27.1 ±3.9	96.1 ±9.5	97.9 ±4.4
	Hopper 89.6 ±27.6	1.1	80.9	111.4 ±1.2	90.6 ±43.3	112.2 ±0.2
	Walker2d 12.0 ±5.8	-0.3	42.7	68.1 ±13.1	103.6 ±4.6	101.1 ±9.3
Expert	HalfCheetah 105.2 ±1.7	3.8	78.5	82.4 ±7.4	106.8 ±3.0	105.7 ±1.9
	Hopper 111.5 ±1.3	6.6	85.2	111.2 ±2.1	112.3 ±0.2	112.2 ±0.2
	Walker2d 56.0 ±24.9	-0.2	57.0	103.8 ±7.6	79.9 ±32.4	105.7 ±2.7
	Total 595.3 ±91.5	284.1	-	764.3 ±61.5	974.6 ±108.3	979.3 ±33.4

	Iterative	One-step			
	Fu et al. [2020]	BC	Easy BCQ	Rev. KL Reg	Exp. Weight
halfcheetah-m	46.3	41.9 ± 0.1	52.6 ± 0.2	**55.2 ± 0.4**	48.4 ± 0.1
walker2d-m	81.1	68.6 ± 6.3	**87.2 ± 1.3**	85.9 ± 1.4	81.8 ± 2.2
hopper-m	58.8	49.9 ± 3.1	74.5 ± 6.2	**83.7 ± 4.5**	59.6 ± 2.5
halfcheetah-m-e	64.7	61.1 ± 2.7	78.2 ± 1.6	**93.8 ± 0.5**	93.4 ± 1.6
walker2d-m-e	111.0	78.5 ± 22.4	112.2 ± 0.3	111.2 ± 0.2	**113.0 ± 0.4**
hopper-m-e	**111.9**	49.1 ± 4.3	85.1 ± 2.2	98.7 ± 7.5	103.3 ± 9.1
halfcheetah-m-re	**47.7**	34.6 ± 0.9	38.3 ± 0.3	41.9 ± 0.5	38.1 ± 1.3
walker2d-m-re	26.7	26.6 ± 3.4	69.1 ± 4.2	**74.9 ± 6.6**	49.5 ± 12.0
hopper-m-re	48.6	23.1 ± 2.7	78.4 ± 7.2	92.3 ± 1.1	**97.5 ± 0.7**
halfcheetah-r	**35.4**	2.2 ± 0.0	5.4 ± 0.3	8.8 ± 3.8	3.2 ± 0.1
walker2d-r	**7.3**	0.9 ± 0.1	3.7 ± 0.1	6.2 ± 0.7	5.6 ± 0.8
hopper-r	**12.2**	2.0 ± 0.1	6.6 ± 0.1	7.9 ± 0.7	7.5 ± 0.4
pen-c	56.9	46.9 ± 11.0	**65.9 ± 3.6**	57.4 ± 3.5	60.0 ± 4.1
hammer-c	2.1	0.4 ± 0.1	**2.9 ± 0.5**	0.2 ± 0.1	2.1 ± 0.7
relocate-c	-0.1	-0.1 ± 0.0	**0.3 ± 0.2**	0.2 ± 0.1	0.2 ± 0.1
door-c	0.4	0.0 ± 0.1	**0.6 ± 0.6**	0.2 ± 0.7	0.2 ± 0.3

halfcheetah-random	30.5	23.5	28.1	6.3 ± 4.0	27.1 ± 1.3	20.7 ± 0.6	33.3 ± 1.3
hopper-random	11.3	11.1	12.0	10.8 ± 0.3	10.6 ± 0.1	10.4 ± 0.1	11.3 ± 0.2
walker2d-random	4.1	0.8	0.5	8.1 ± 5.5	1.1 ± 2.2	10.0 ± 4.6	1.5 ± 0.7
halfcheetah-medium	36.1	44.0	45.5	44.6 ± 0.8	40.3 ± 0.3	38.9 ± 0.3	41.3 ± 0.3
walker2d-medium	6.6	72.7	81.3	41.0 ± 29.4	77.3 ± 3.8	69.2 ± 8.3	78.8 ± 1.0
hopper-medium	29.0	31.2	32.3	48.8 ± 26.8	42.2 ± 15.5	30.5 ± 0.7	99.4 ± 0.3
halfcheetah-expert	107.0	3.8	−1.1	-	54.4 ± 45.8	103.5 ± 1.3	108.4 ± 0.5
hopper-expert	109.0	6.6	3.7	-	67.7 ± 54.7	112.2 ± 0.2	112.3 ± 0.1
walker2d-expert	125.7	−0.2	−0.0	-	84.7 ± 42.7	107.2 ± 3.8	103.0 ± 5.0
halfcheetah-medium-expert	35.8	43.8	45.3	105.9 ± 17.8	21.7 ± 6.8	58.6 ± 8.7	93.3 ± 10.2
walker2d-medium-expert	11.3	−0.3	0.9	70.2 ± 36.2	104.0 ± 10.1	104.6 ± 10.4	105.2 ± 3.9
hopper-medium-expert	111.9	1.1	0.8	55.1 ± 44.3	111.3 ± 2.1	112.4 ± 0.2	112.4 ± 0.3
halfcheetah-mixed	38.4	45.6	45.9	42.3 ± 0.9	44.9 ± 1.1	42.0 ± 1.1	43.2 ± 1.5
hopper-mixed	11.8	0.7	0.8	12.4 ± 5.8	31.6 ± 3.6	29.0 ± 0.5	35.6 ± 1.0
walker2d-mixed	11.3	−0.3	0.9	9.7 ± 5.3	16.8 ± 3.1	16.5 ± 4.9	41.8 ± 7.9



What is typically reported when comparing new Offline-RL algorithms?

	BC	D4PG	ABM	BCQ	CRR exp	CRR binary	CRR binary max
Cartpole Swingup	386 ± 6	855 ± 13	798 ± 30	444 ± 15	664 ± 22	**860 ± 7**	858 ± 15
Finger Turn Hard	261 ± 39	764 ± 24	566 ± 25	311 ± 38	714 ± 38	755 ± 31	**833 ± 57**
Walker Stand	386 ± 6	**929 ± 46**	689 ± 13	501 ± 5	797 ± 30	881 ± 13	**929 ± 10**
Humanoid Run	382 ± 2	1 ± 1	302 ± 6	22 ± 3	**586 ± 6**	412 ± 10	226

Maximum Performance after Online Policy Selection

	Iterative	One-step			
	Fu et al. [2020]	BC	Easy BCQ	Rev. KL Reg	Exp. Weight
halfcheetah-m	46.3	41.9 ± 0.1	52.6 ± 0.2	**55.2 ± 0.4**	48.4 ± 0.1
walker2d-m	81.1	68.6 ± 6.3	**87.2 ± 1.3**	85.9 ± 1.4	81.8 ± 2.2
hopper-m	58.8	49.9 ± 3.1	74.5 ± 6.2	**83.7 ± 4.5**	59.6 ± 2.5
halfcheetah-m-e	64.7	61.1 ± 2.7	78.2 ± 1.6	**93.8 ± 0.5**	93.4 ± 1.6
halfcheetah-r	**35.4**	2.2 ± 0.0	5.4 ± 0.3	8.8 ± 3.8	3.2 ± 0.1
walker2d-r	**7.3**	0.9 ± 0.1	3.7 ± 0.1	6.2 ± 0.7	5.6 ± 0.8
hopper-r	**12.2**	2.1 ± 0.1	6.6 ± 0.1	7.9 ± 0.7	7.5 ± 0.4
pen-c	56.9	46.9 ± 11.0	**65.9 ± 3.6**	57.4 ± 3.5	60.0 ± 4.1
hammer-c	2.1	0.4 ± 0.1	**2.9 ± 0.5**	0.2 ± 0.1	2.1 ± 0.7
relocate-c	-0.1	-0.1 ± 0.0	**0.3 ± 0.2**	0.2 ± 0.1	0.2 ± 0.1
door-c	0.4	0.0 ± 0.1	**0.6 ± 0.6**	0.2 ± 0.7	0.2 ± 0.3

Maximum Performance after Offline Policy Selection

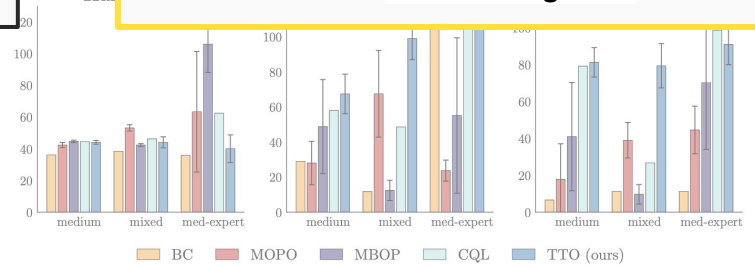
Domain	Task Name	BC	SAC	BEAR	BRAC-p	BRAC-v	CQL($\mathcal{H}$)	CQL(ρ)
AntMaze	antmaze-umaze	65.0	0.0	**73.0**	50.0	70.0	**74.0**	**73.5**
	antmaze-umaze-diverse	55.0	0.0	61.0	40.0	70.0	**84.0**	61.0
	antmaze-medium-play	0.0	0.0	0.0	0.0	0.0	**61.2**	4.6
	antmaze-medium-diverse	0.0	0.0	8.0	0.0	0.0	**53.7**	5.1
	antmaze-large-play	0.0	0.0	0.0	0.0	0.0	**15.8**	3.2
	antmaze-large-diverse	0.0	0.0	0.0	0.0	0.0	**14.9**	2.3
Adroit	pen-human	34.4	6.3	-1.0	8.1	0.6	37.5	**55.8**
	hammer-human	1.5	0.5	0.3	0.3	0.2	**4.4**	2.1
	door-human	0.5	3.9	-0.3	-0.3	-0.3	**9.9**	**9.1**
	relocate-human	0.0	0.0	-0.3	-0.3	-0.3	0.20	**0.35**
	pen-cloned	**56.9**	23.5	26.5	1.6	-2.5	39.2	40.3
	hammer-cloned	0.8	0.2	0.3	0.3	0.3	2.1	**5.7**
	door-cloned	-0.1	0.0	-0.1	-0.1	-0.1	0.4	**3.5**
Kitchen								

Online Budget = #Hyperparameter Assignments

	BC	BRAC
Random	Hopper	9.5 ±0.1	11.1	9.6	10.7 ±0.1	11.4 ±
	Walker2d	1.2 ±0.2	0.8	5.1	2.7 ±1.2	0.6 ±
Medium	HalfCheetah	36.6 ±0.6	44.0	37.4	37.2 ±0.3	41.3 ±
	Hopper	30.0 ±0.5	31.2	72.0	44.2 ±10.8	99.4 ±
	Walker2d	11.4 ±6.3	72.7	30.1	57.5 ±8.3	79.5 ±
Medium Replay	HalfCheetah	34.7 ±1.8	45.6	-	41.9 ±1.1	43.3 ±
	Hopper	19.7 ±5.9	0.7	-	28.6 ±0.9	35.6 ±
	Walker2d	8.3 ±1.5	-0.3	-	15.8 ±2.6	42.6 ±
Medium Expert	HalfCheetah	67.6 ±13.2	43.8	36.8	27.1 ±3.9	96.1 ±9.5	97.9 ±4.4
	Hopper	89.6 ±27.6	1.1	80.9	111.4 ±1.2	90.6 ±43.3	112.2 ±0.2
	Walker2d	12.0 ±5.8	-0.3	42.7	68.1 ±13.1	103.6 ±4.6	101.1 ±9.3
Expert	HalfCheetah	105.2 ±1.7	3.8	78.5	82.4 ±7.4	106.8 ±3.0	105.7 ±1.9
	Hopper	111.5 ±1.3	6.6	85.2	111.2 ±2.1	112.3 ±0.2	112.2 ±0.2
	Walker2d	56.0 ±24.9	-0.2	57.0	103.8 ±7.6	79.9 ±32.4	105.7 ±2.7
	Total	595.3 ±91.5	284.1	-	764.3 ±61.5	974.6 ±108.3	979.3 ±33.4

Online Budget = 1

halfcheetah-random	30.5	23.5	28.1	6.3 ± 4.0	27.1 ± 1.3	20.7 ± 0.6	33.3 ± 1.3
hopper-random	11.3	11.1	12.0	10.8 ± 0.3	10.6 ± 0.1	10.4 ± 0.1	11.3 ± 0.2
walker2d-random	4.1	0.8	0.5	8.1 ± 5.5	1.1 ± 2.2	10.0 ± 4.6	1.5 ± 0.7
halfcheetah-medium	36.1	44.0	45.5	44.6 ± 0.8	40.3 ± 0.3	38.9 ± 0.3	41.3 ± 0.3
walker2d-medium	6.6	72.7	81.3	41.0 ± 29.4	77.3 ± 3.8	69.2 ± 8.3	78.8 ± 1.0
hopper-medium	29.0	31.2	32.3	48.8 ± 26.8	42.2 ± 15.5	30.5 ± 0.7	99.4 ± 0.3
halfcheetah-expert	107.0	3.8	−1.1	-	54.4 ± 45.8	103.5 ± 1.3	108.4 ± 0.5
hopper-expert	109.0	6.6	3.7	-	67.7 ± 54.7	112.2 ± 0.2	112.3 ± 0.1
walker2d-expert	125.7	−0.2	−0.0	-	84.7 ± 42.7	107.2 ± 3.8	103.0 ± 5.0
halfcheetah-medium-expert	35.8	43.8	45.3	105.9 ± 17.8	21.7 ± 6.8	58.6 ± 8.7	93.3 ± 10.2
walker2d-medium-expert	11.3	−0.3	0.9	70.2 ± 36.2	104.0 ± 10.1	104.6 ± 10.4	105.2 ± 3.9
hopper-medium-expert	111.9	1.1	0.8	55.1 ± 44.3	111.3 ± 2.1	112.4 ± 0.2	112.4 ± 0.3
halfcheetah-mixed	38.4	45.6	45.9	42.3 ± 0.9	44.9 ± 1.1	42.0 ± 1.1	43.2 ± 1.5
hopper-mixed	11.8	0.7	0.8	12.4 ± 5.8	31.6 ± 3.6	29.0 ± 0.5	35.6 ± 1.0
walker2d-mixed	11.3	−0.3	0.9	9.7 ± 5.3	16.8 ± 3.1	16.5 ± 4.9	41.8 ± 7.9



😢 **In some problems, it is not feasible to deploy <u>all</u> policies online**

😢 **In some problems, it is not feasible to deploy <u>all</u> policies online**

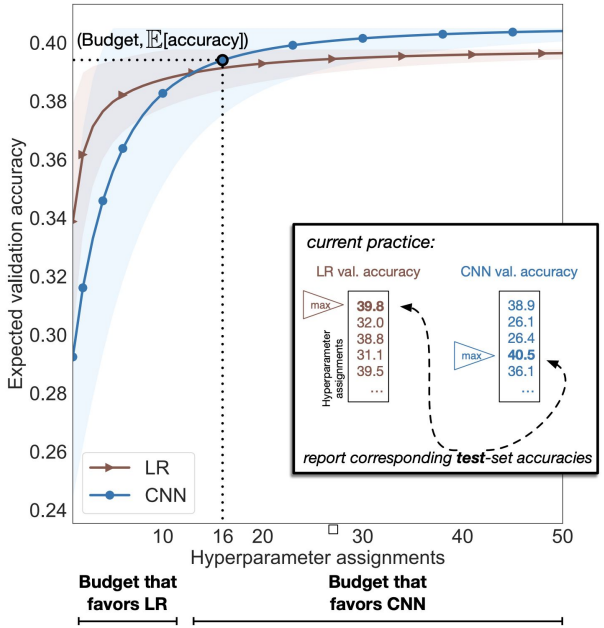
😼 **But it is feasible to deploy more than <u>one</u> policy!**

😢 **In some problems, it is not feasible to deploy <u>all</u> policies online**

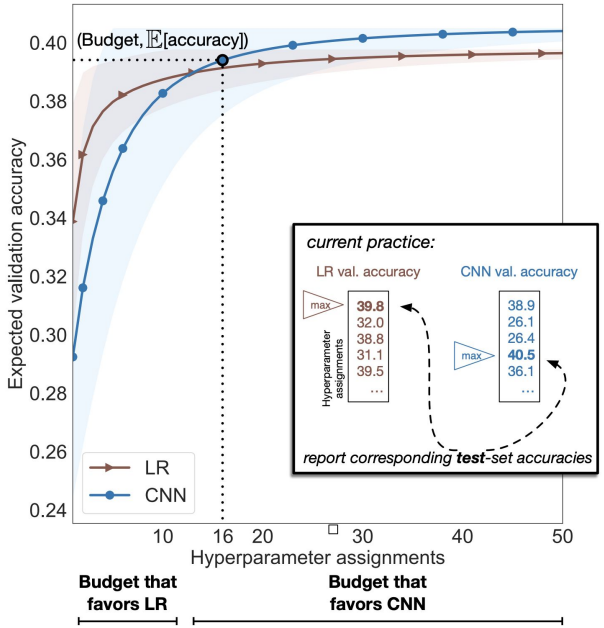
😼 **But it is feasible to deploy more than <u>one</u> policy!**

🤔 **So which algorithm should we prefer, if we are restricted to evaluate no more than N policies online?**

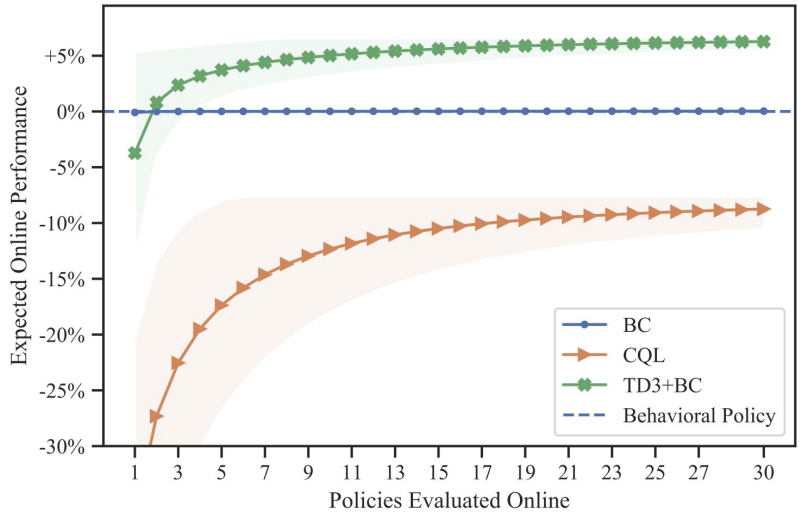
From Expected Validation Performance to Expected Online Performance



From Expected Validation Performance to Expected Online Performance



NLP to Offline-RL



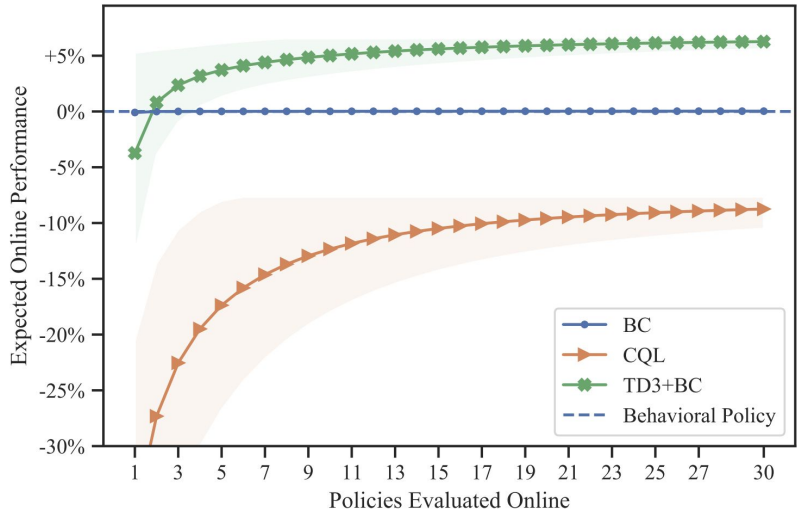
(b) FinRL

Expected Online Performance

1. **Target Performance**
   (Performance Relative to the Best Behavioral Policy)

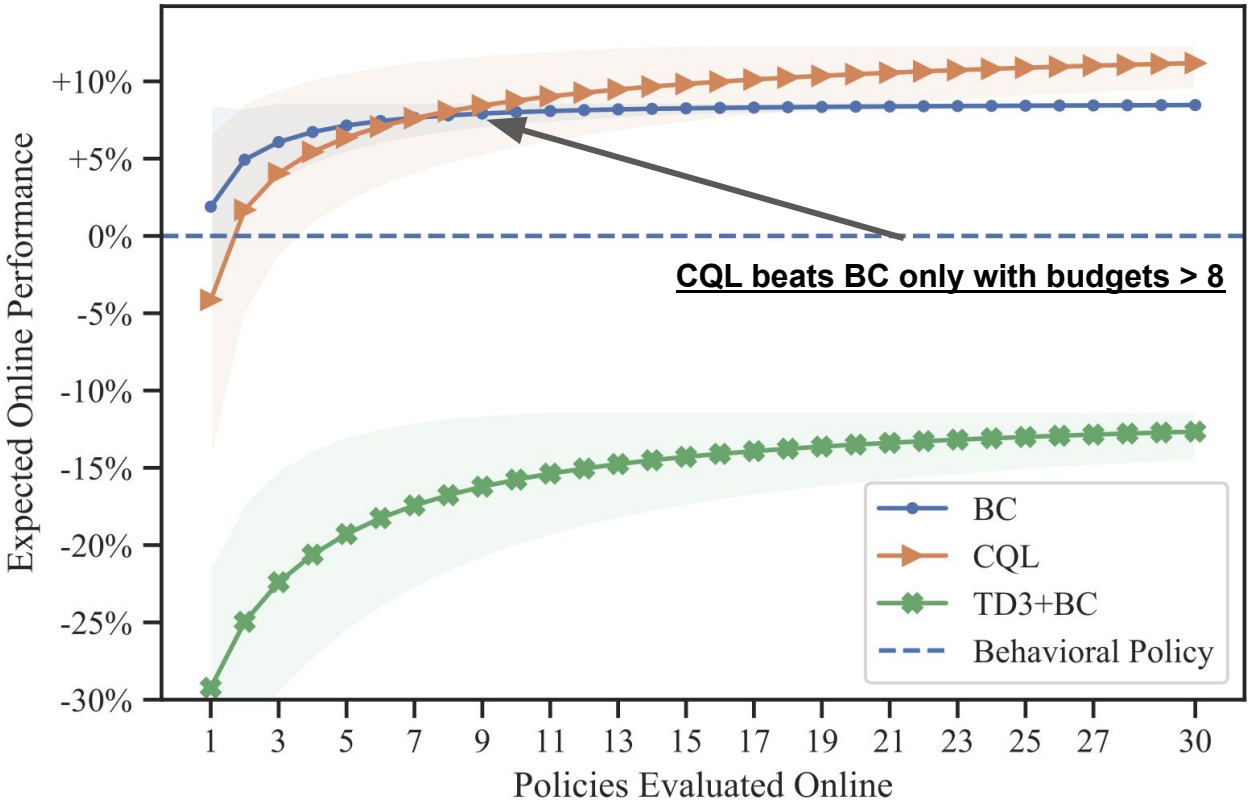
2. **Online Budget**
   (Number of policies evaluated online)

3. **Offline Policy Selection strategy**
   (Uniform Selection for an Efficient Estimator)

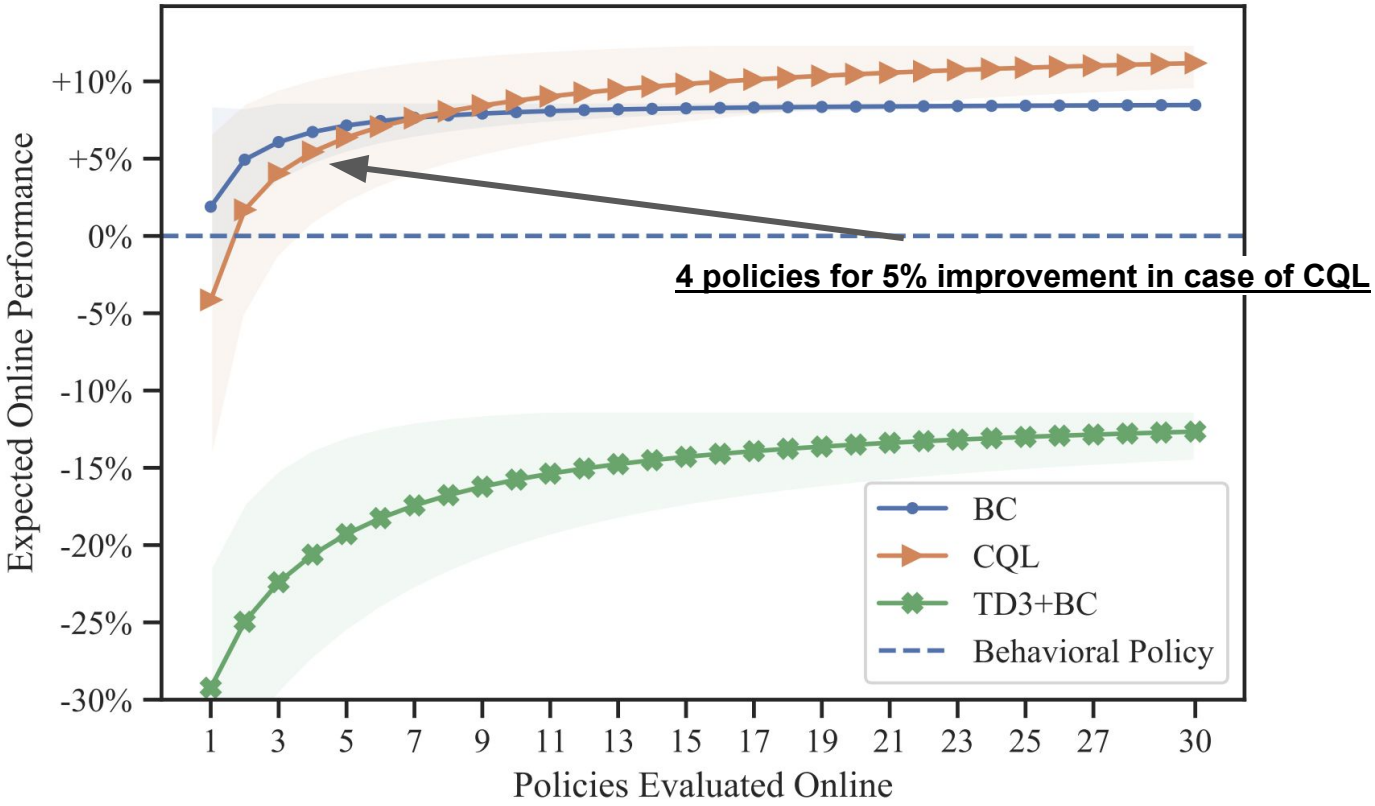


(b) FinRL

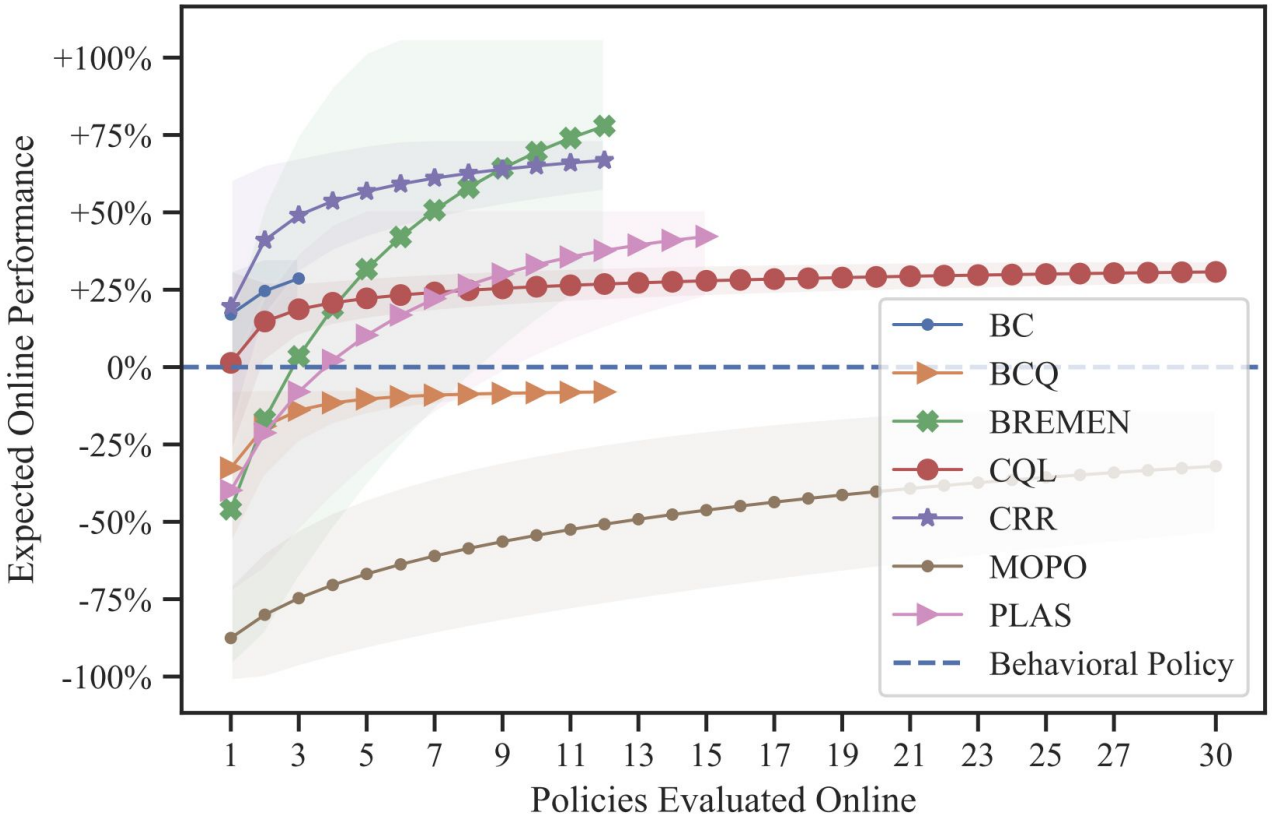
EOP: Different algorithms are preferred under varied budgets



EOP: How many policies to deploy for a satisfactory performance?



Can we apply EOP to already existing benchmarks?



Walker-2d
Qin et al., 2021,
NeoRL: A Near Real-World Benchmark for Offline Reinforcement Learning

EOP: Compare Offline Policy Selection methods

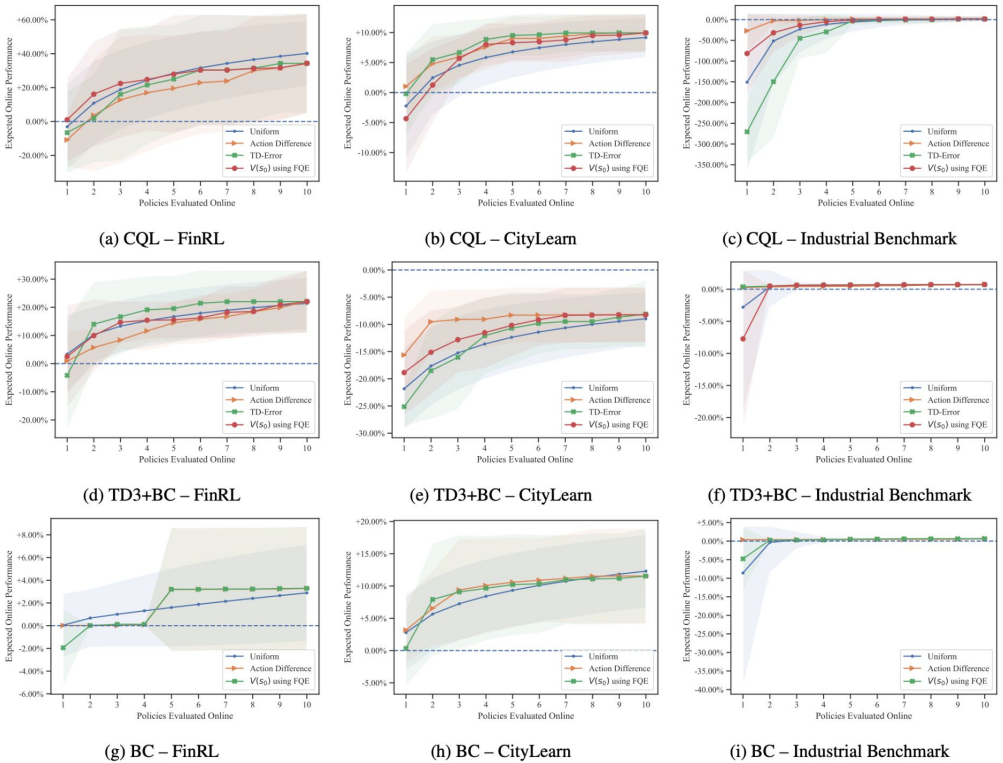


Figure 2: Expected Online Performance under different offline policy selection methods. In most cases, the resulting curves are hardly distinguishable, suggesting that uniform selection should not be overlooked in research reports and practitioner toolsets. The shaded area represents one standard deviation.

Future Work

- Adapt for risk-sensitive scenarios
- Estimator for OPS methods beyond uniform strategy

web-page: tinkoff-ai.github.io/eop/