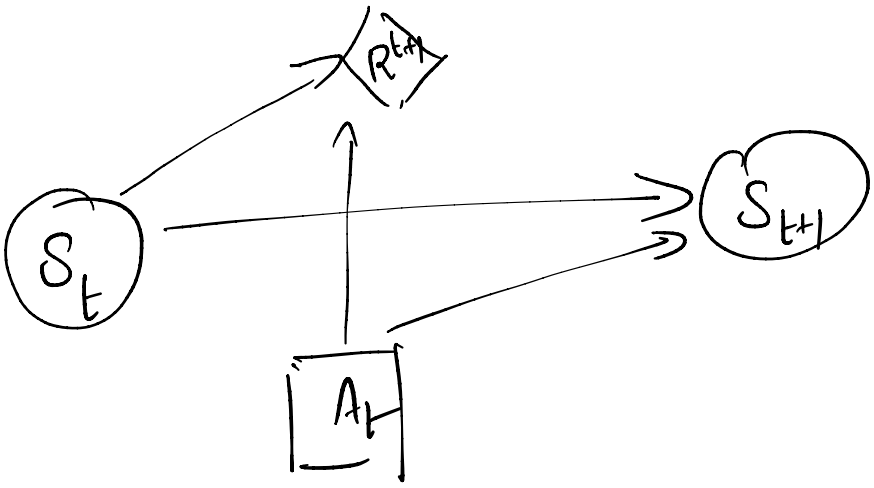
Lecture-12

- Partial observability
- Offline RL

- Partial observability :-

MDP framework :-

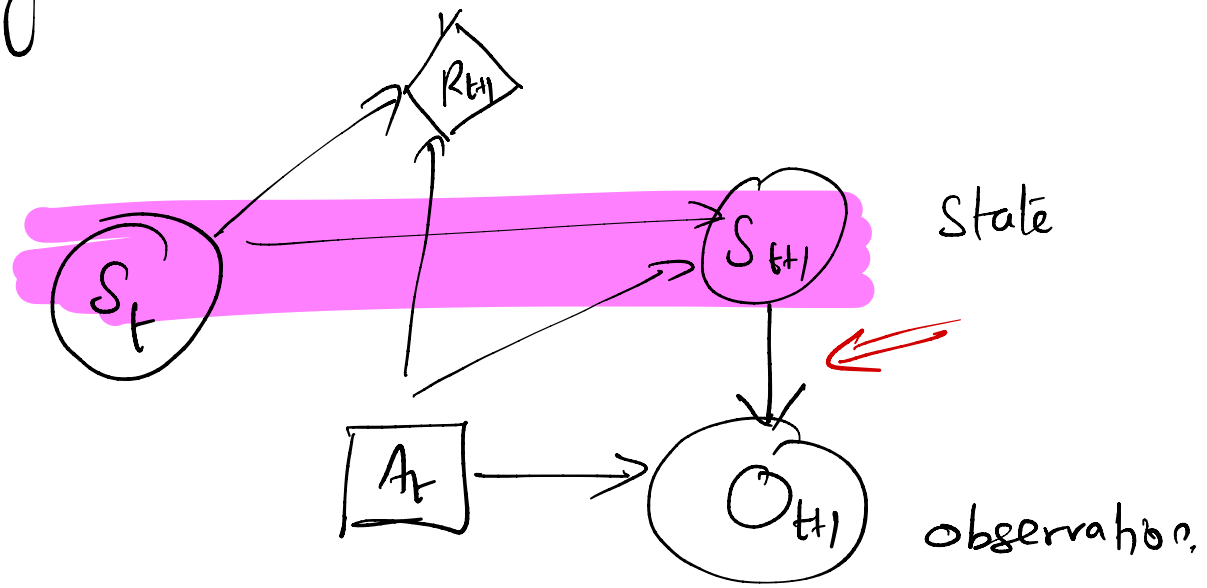
$$Pr(S_{t+1} \mid S_0, S_1 \ldots S_t) = Pr(S_{t+1} \mid S_t)$$



$$P_{ij}^{a} = Pr(S_{t+1} = s_j \mid S_t = s_i, A_t = a)$$

$$T(s, a, s') = Pr(S_{t+1} = s \mid S_t = s, A_t = a)$$

Partially observable MDP:-



$\mathcal{O}$ = be a set of observations.

$$Z : S \times A \longmapsto \Delta(\mathcal{O})$$

$$Z(s', a, o') = Pr(O_{t+1} = o' \mid S_{t+1} = s', A_t = a)$$

POMDP: $\langle S, A, T, R, \mathcal{O}, Z \rangle$

POMDP

History: $\langle s_0, o_0, a_0 \rangle \langle s_1, o_1, a_1 \rangle \ldots \langle s_t, o_t, a_t \rangle$

$\mathcal{H}$ — set of all complete histories.

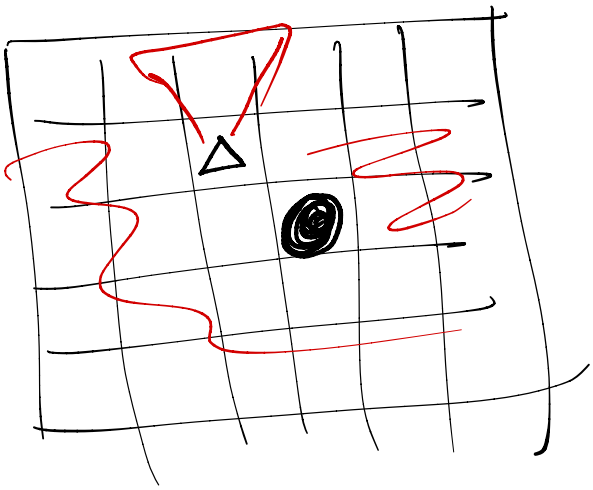
System history $\mathcal{H}_S$ $\qquad V : \mathcal{H}_S \to \mathbb{R}$

$\langle s_0, a_0 \rangle \langle s_1, a_1 \rangle \ldots \langle s_t, a_t \rangle$

Observation history $\mathcal{H}_O$

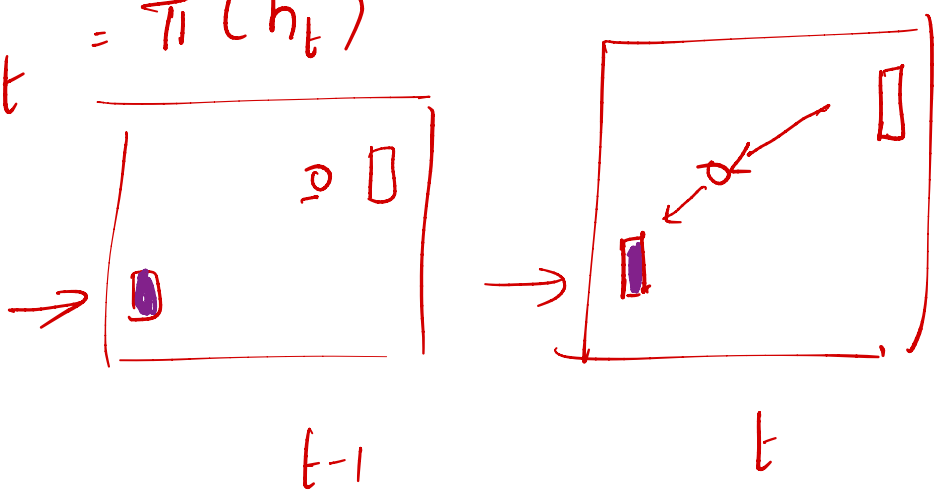
$\langle a_0, o_1 \rangle \langle a_1, o_2 \rangle \ldots \langle a_{t-1}, o_t \rangle$

b_0



$h_t = \langle a_0, o_1 \rangle \langle a_1, o_2 \rangle \cdots \langle a_{t-1}, o_t \rangle$

$a_t = \pi(h_t)$



$t-1$ t

$Z(s', a, o')$

$Z(o' \mid s', a)$

POMDP ⟶ belief state MDP.

$\langle s_1, s_2, s_3 \rangle$

b_1 $(0, 0.5, 0.5)$

b_2 $(0.9, 0.05, 0.05)$

⋮

Belief state MDP :-

$B = \Delta(S)$ — Continuous state space.

A = set of actions is same as the original POMDP.

$R^b : B \times A \to \mathbb{R}$

$$R^b(b,a) = \sum_{s \in S} b(s) R(s,a)$$

$T^b : B \times A \to B$

$$T^b(b,a,b') = Pr(b'|b,a)$$

$$= \sum_{o \in O} \underline{Pr(b'|a,b,o)} \underline{P(o|a,b)}$$

$$= \sum_{o \in O} Pr(b'|a,b,o) \sum_{s' \in S} Z(s',a,o) \sum_{s \in S} T(s,a,s') b(s)$$

$$= \sum_{o \in O} \Pr(b' \mid a, b, o) \sum_{s' \in S} \mathcal{I}(s', a, o) \sum_{s \in S} T(s, a, s') b(s)$$

$$\Pr(b' \mid a, b, o) = \begin{cases} 1 & \text{if } b_o^a = b' \\ 0 & \text{otherwise.} \end{cases}$$

$$b_o^a(s') = \frac{\mathcal{I}(s', a, o) \sum_{s \in S} T(s, a, s') b(s)}{\Pr(o \mid a, b)}$$

$$\Pr(o \mid a, b) = \sum_{s' \in S'} \mathcal{I}(s', a, o) \sum_{s \in S} T(s, a, s') b(s)$$

$$Q(b, a) = \sum_{s \in S} b(s) R(s, a) + \gamma \sum_{o \in O} \Pr(o \mid a, b) V^*(b_o^a)$$

$$\pi^*(b) = \operatorname*{argmax}_{a \in A} Q(b, a)$$

Value iteration:

$$V_{t+1}(s) = \max_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') V_t(s') \right]$$

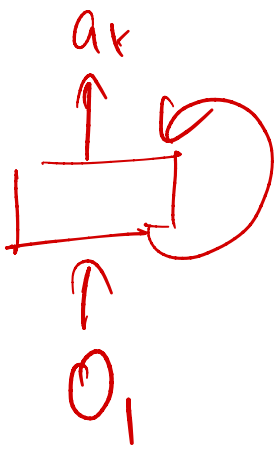
POMDP value iteration.

$$V_{t+1}(b) = \max_{a \in A} \left[R^b(b,a) + \gamma \sum_{b' \in B} T^b(b,a,b') V_t(b') \right]$$

$$\Updownarrow$$

$$V_{t+1}(b) = \max_{a \in A} \left[\sum_{s \in S} b(s) R(s,a) + \gamma \sum_{o \in O} P_r(o|a,b) V_t(b^a_o) \right]$$

POMDPs using Recurrent Neural Networks :-

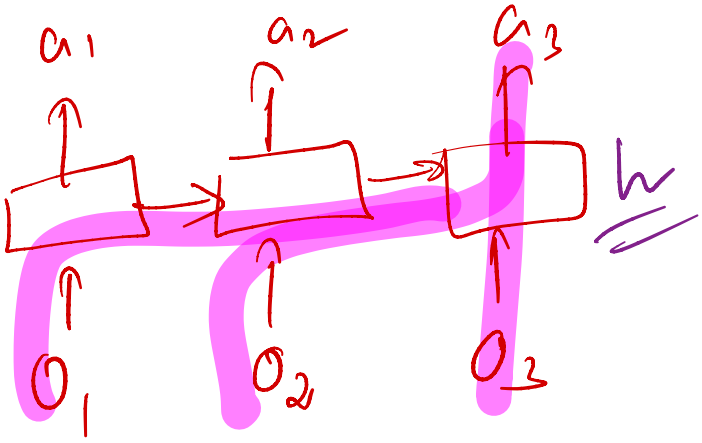


$$h_t = f(O_t) = \sigma(wO_t + b)$$

$$a_t = f(h_t) = vh_t + c$$

$$h_t = f(O_t, h_{t-1}) = \sigma(wO_t + Ph_{t-1} + b)$$

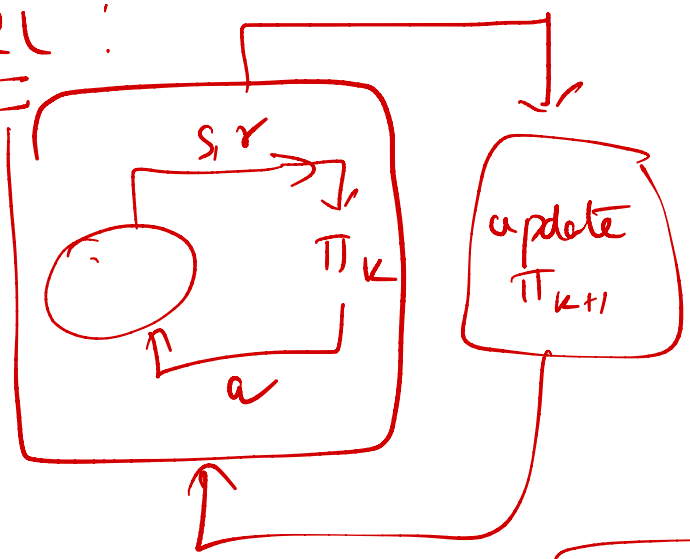
$$a_t = f(h_t) = vh_t + c$$



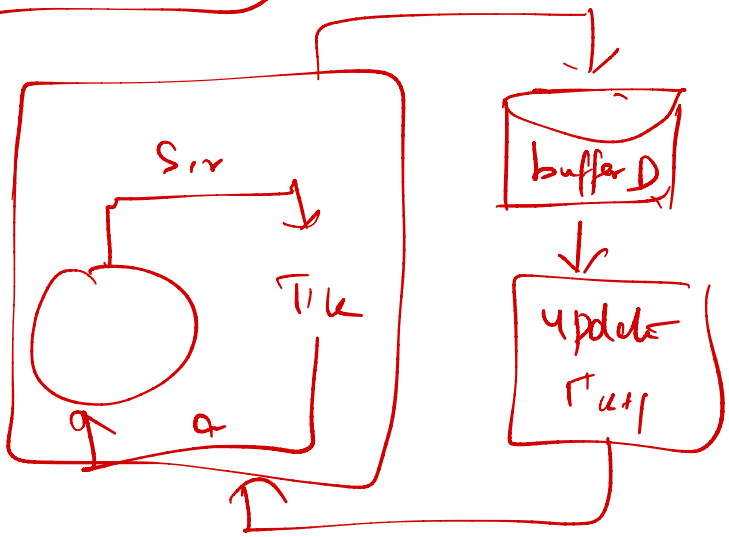
→ offline RL

→ learning through interaction
- → on-policy
- → off-Policy.

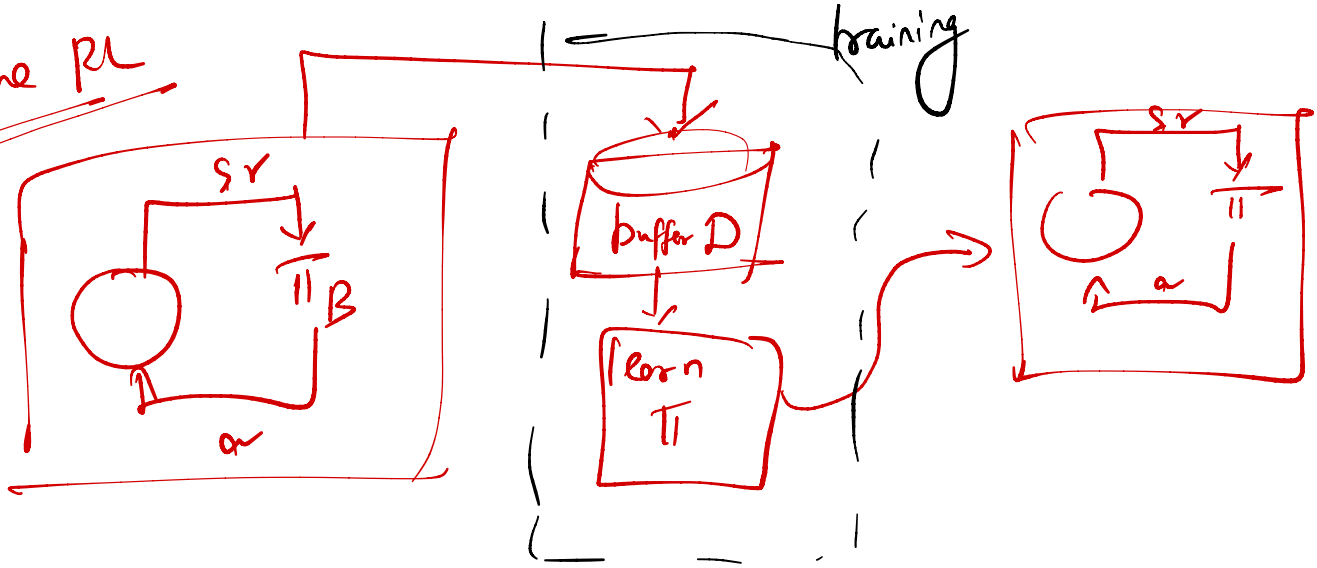
Online RL :



off-Policy RL



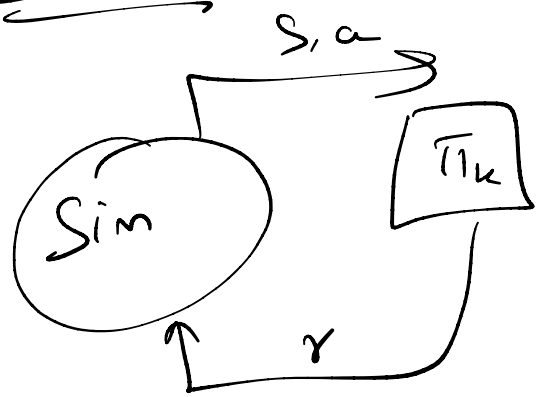
Offline RL



why offline RL ?

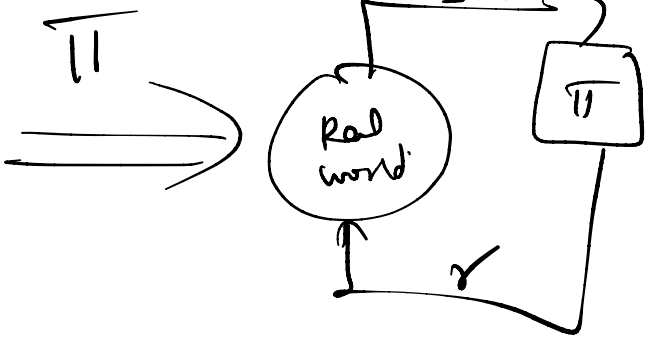
- Safe RL
- offline RL → pretraining for online RL.
- data efficiency
- multi-tasking.

Sim2Real



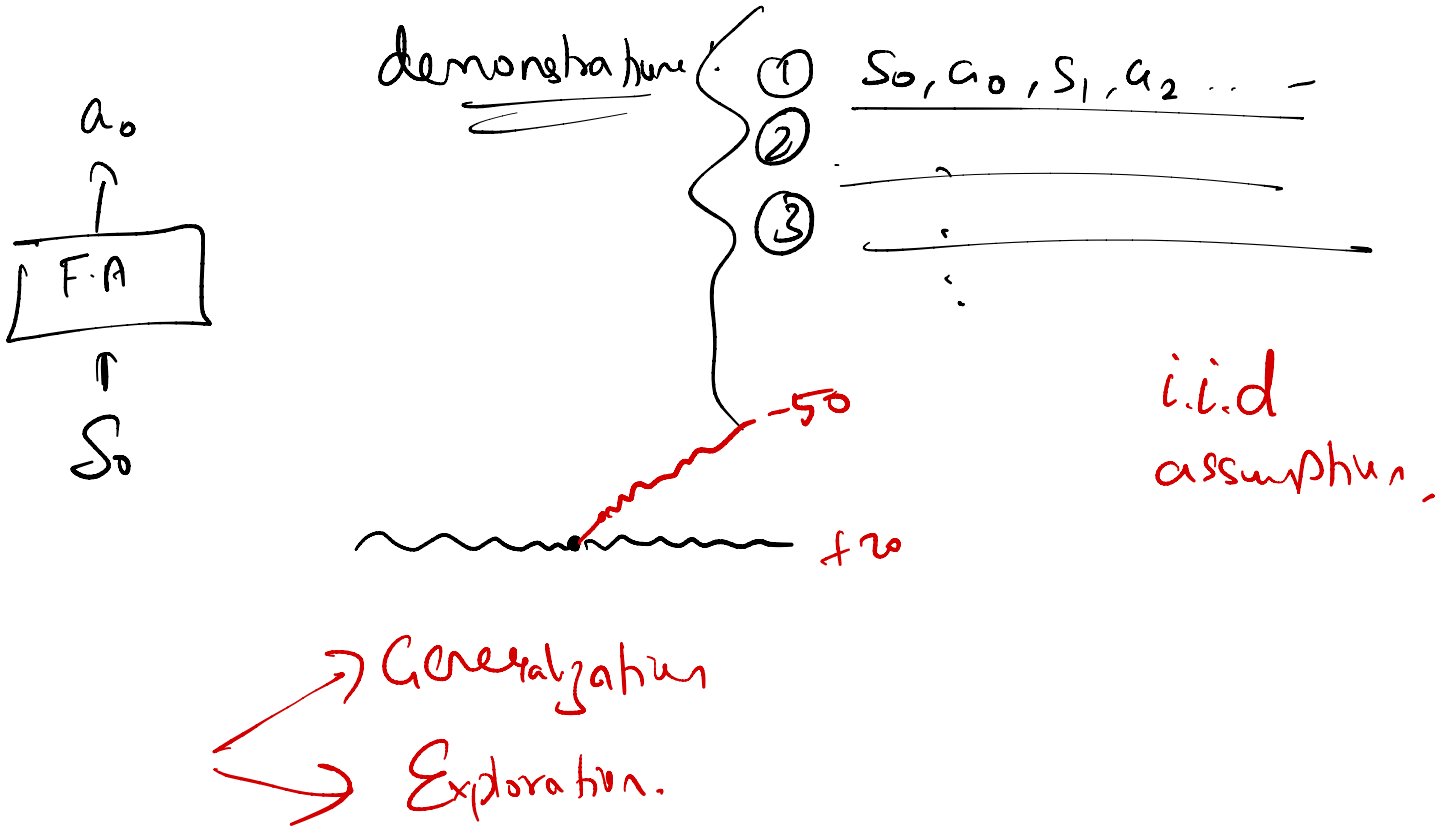
Simulator.

CARLA



Behavior cloning:

$S, A, P(s'|s,a)$



→ Generalization

→ Exploration.

DAGGER: Dataset Aggregation.

$D \leftarrow \emptyset$

initialize $\hat{\pi}_1$ to any policy in Π.

for $i = 1$ to N:

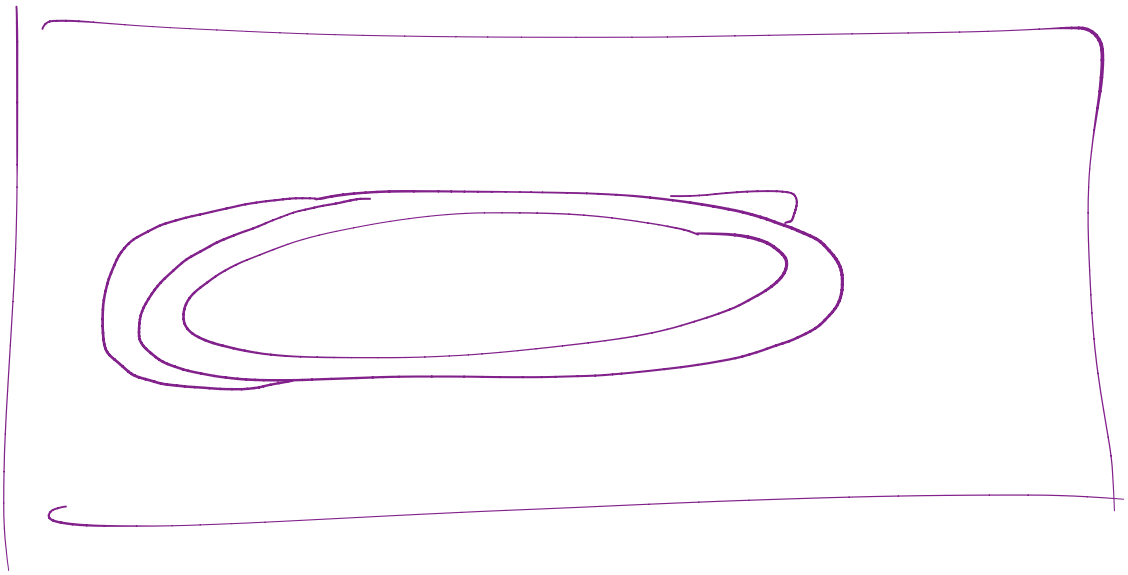
Let $\pi_i = \beta_i \pi^* + (1-\beta_i)\hat{\pi}_i$

Sample T-step traj. using π_i

→ $D_i = \{(s, \pi^*(s))\}$ of visited states and actions given by expert.

Aggr. $D \leftarrow D \cup D_i$

Train $\hat{\pi}_{i+1}$ on D.



- Pure offline RL

Policy constraint methods

$$\pi(a'|s') \sim \pi_\beta(a'|s')$$

→ Policy constraint

→ Policy penalty.

Policy constraint:

$$\hat{Q}^{\pi}_{k+1} \leftarrow \operatorname{argmin}_{Q} \mathbb{E}_{(s,a,s')\sim\mathcal{D}} \left[\left(Q(s,a) - \left(r(s,a) + \gamma \mathbb{E}_{a'\sim\pi_k(a'|s')} \left[\hat{Q}^{\pi}_{k}(s',a') \right] \right) \right)^2 \right]$$

$$\pi_{k+1} \leftarrow \operatorname{argmax}_{\pi} \mathbb{E}_{s\sim\mathcal{D}} \left[\mathbb{E}_{a\sim\pi(a|s)} \left[\hat{Q}^{\pi}_{k+1}(s+1) \right] \right]$$

$$\text{s.t. } D(\pi, \pi_\beta) \leq \epsilon$$

Policy penalty methods :-

$$\hat{Q}^{\pi}_{k+1} \leftarrow \arg\min_{Q} \mathbb{E}_{(s,a,s') \sim D} \left[\left(Q(s,a) - \left(\tilde{r}(s,a) + \gamma \mathbb{E}_{a'} \left[Q^{\pi}_{k}(s',a') \right] \right) \right)^2 \right]$$

$$\tilde{r}(s,a) = r(s,a) - \alpha D(\pi(\cdot|s), \pi_{\beta}(\cdot|s))$$

$$\pi_{k+1} \leftarrow \arg\max_{\pi} \mathbb{E}_{s \sim D} \left[\mathbb{E}_{a} \left[Q^{\pi}_{k+1}(s,a) - \alpha D[\ ] \right] \right)$$