CS292F StatRL Lecture 13
OPE in Reinforcement Learning

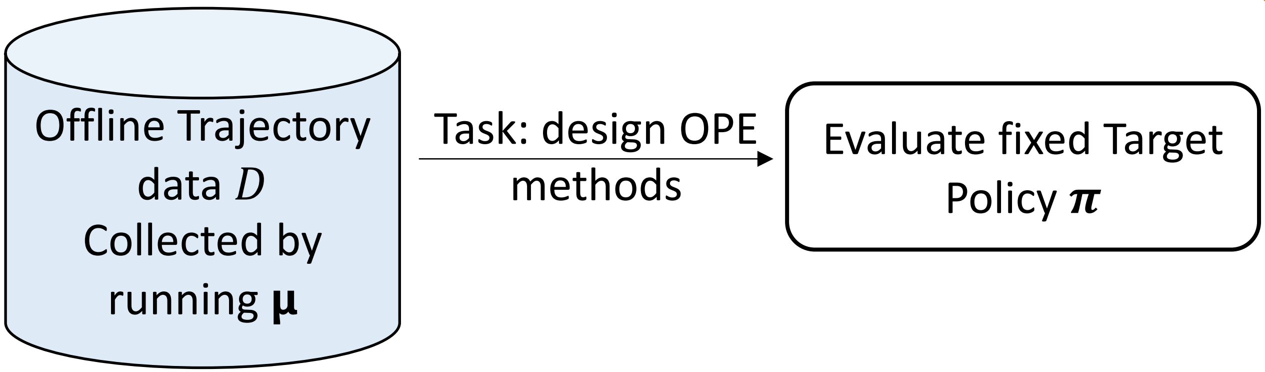
Instructor: Yu-Xiang Wang

Spring 2021

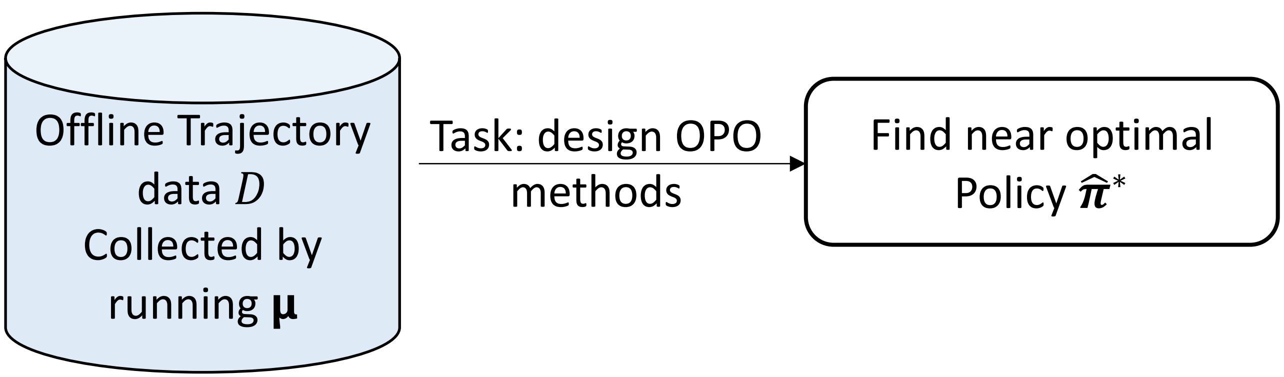
UC Santa Barbara

Recap: Offline Reinforcement Learning, aka. Batch RL

- Task 1: Offline Policy Evaluation. (OPE)



- Task 2: Offline Policy Learning. (OPL)



Recap: Lecture 12

- OPE algorithms in (Contextual) Bandits
  - DM, IS, WIS, DR, SWITCH

- Comparing DM and IS in Multi-armed Bandits:
  - DM is asymptotically more efficient
  - IS is better.

- More generally:
  - DM is asymptotically more efficient if we assume realizability
  - IS cannot be improved when we don’t

Recap: Two standard approaches

- Direct method / regression estimator

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

- Importance sampling / Inverse Propensity Score /

$$\hat{v}^{\pi}_{\mathrm{IPS}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\mu(a_i|x_i)} r_i$$

Recap: Combining DM and IS

- Doubly Robust Estimation
  - Remains unbiased, but limited benefits to the variance
- SWITCH
  - Introduce bias, but drastically reduce variance

This lecture

- Generalizing the bandits OPE idea to RL
- Curse of Horizon
- Marginalized Importance Sampling

OPE in Reinforcement Learning

- Importance sampling on the entire trajectory

- (Per-Step) Importance Sampling

- Exercise:
  - Infinite horizon discounted version?
  - Weighted Importance Sampling Extension?

Doubly Robust OPE in Reinforcement Learning

- An alternative form for the Per-Step IS

$$V^0_{\text{step-IS}} := 0, \text{ and for } t = 1, \ldots, H,$$
$$V^{H+1-t}_{\text{step-IS}} := \rho_t \left(r_t + \gamma V^{H-t}_{\text{step-IS}} \right).$$

- Given a value function approximator

$$V^0_{\text{DR}} := 0, \text{ and for } t = 1, \ldots, H,$$
$$V^{H+1-t}_{\text{DR}} := \widehat{V}(s_t) + \rho_t \Big(r_t + \gamma V^{H-t}_{\text{DR}} - \widehat{Q}(s_t, a_t) \Big).$$

Jiang, N., & Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In ICML 2016.

Mean and Variance of Doubly Robust OPE in RL

- Doubly Robust OPE in RL is unbiased

- Variance

Theorem 1. *V_{DR} is an unbiased estimator of $v^{\pi_1,H}$, whose variance is given recursively as follows:* $\forall t = 1, \ldots, H$,

$$\mathbb{V}_t\big[V_{DR}^{H+1-t}\big] = \mathbb{V}_t\big[V(s_t)\big] + \mathbb{E}_t\Big[\mathbb{V}_t\big[\rho_t \Delta(s_t, a_t) \mid s_t\big]\Big] + \mathbb{E}_t\Big[\rho_t^2\, \mathbb{V}_{t+1}\big[r_t\big]\Big] + \mathbb{E}_t\Big[\gamma^2 \rho_t^2\, \mathbb{V}_{t+1}\big[V_{DR}^{H-t}\big]\Big], \quad (11)$$

where $\Delta(s_t, a_t) := \widehat{Q}(s_t, a_t) - Q(s_t, a_t)$ *and* $\mathbb{V}_{H+1}\big[V_{DR}^0 \mid s_H, a_H\big] = 0$.

Main challenge of OPE in RL: The curse of Horizon

$$\widehat{v}^{\pi}_{IS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\left[\prod_{t=1}^{h}\frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})}\right] r_h^{(i)}.$$

The curse of horizon. (Liu et al, 2018 NeurIPS)

- The variance is exponential in H!

Example on Curse of Horizon

From Importance Sampling to Marginalized Importance Sampling

$$\widehat{v}^{\pi}_{IS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\left[\prod_{t=1}^{h}\frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})}\right] r_h^{(i)}.$$

$$\widehat{v}^{\pi}_{MIS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\widehat{d}^{\pi}_t(s_t^{(i)})}{\widehat{d}^{\mu}_t(s_t^{(i)})}\widehat{r}^{\pi}_t(s_t^{(i)}).$$

Xie, W., and Ma. (2019): Towards Optimal OPE for RL using Marginalized Importance Sampling. NeurIPS 2019.

What are some ideas for estimating the marginalized importance weight?

$$\widehat{v}^{\pi}_{MIS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\widehat{d}^{\pi}_{t}(s_t^{(i)})}{\widehat{d}^{\mu}_{t}(s_t^{(i)})}\widehat{r}^{\pi}_{t}(s_t^{(i)}).$$

- Idea 1: averaging over multiple visits to the same state.

Idea 2: Recursive estimation

$$d_t^\pi(s_t) = \sum_{s_{t-1}} P_t^\pi(s_t|s_{t-1}) d_{t-1}^\pi(s_{t-1}),$$

$$\widehat{d_t^\pi} = \widehat{P_t^\pi}\, \widehat{d_{t-1}^\pi},$$

$$\text{where } \widehat{P}_t^\pi(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}} \sum_{i=1}^{n} \frac{\pi(a_{t-1}^{(i)}|s_{t-1})}{\mu(a_{t-1}^{(i)}|s_{t-1})} \mathbf{1}((s_{t-1}^{(i)}, s_t^{(i)}) = (s_{t-1}, s_t));$$

$$\widehat{r_t^\pi}(s_t) = \frac{1}{n_{s_t}} \sum_{i=1}^{n} \frac{\pi(a_t^{(i)}|s_t)}{\mu(a_t^{(i)}|s_t)} r_t^{(i)} \mathbf{1}(s_t^{(i)} = s_t),$$

Xie, W., and Ma. (2019): Towards Optimal OPE for RL using Marginalized Importance Sampling. NeurIPS 2019.

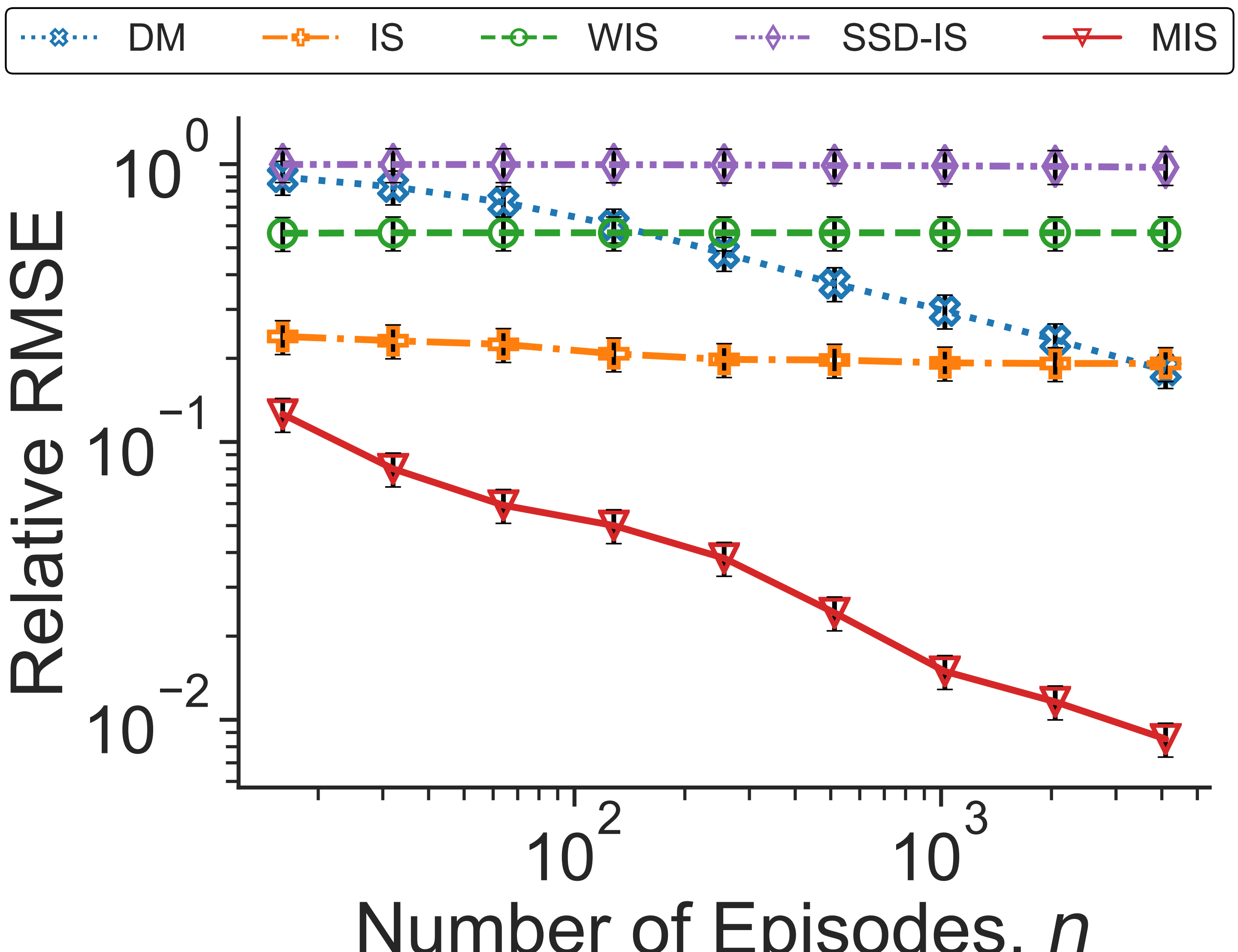
Results: OPE error bound of MIS

- The MSE of MIS estimator obeys:

$$\frac{1}{n}\sum_{t=1}^{H}\mathbb{E}_\mu\left[\frac{d_t^\pi(s_t)^2}{d_t^\mu(s_t)^2}\mathrm{Var}_\mu\left[\frac{\pi_t(a_t|s_t)}{\mu_t(a_t|s_t)}\left(V_{t+1}^\pi(s_{t+1})+r_t\right)\middle|s_t\right]\right]+\tilde{O}(n^{-1.5})$$

Xie, W., and Ma. (2019): Towards Optimal OPE for RL using Marginalized Importance Sampling. NeurIPS 2019.

Experiment on mountain car



Challenges of the analysis

1. **Dependent data:** The data within each trajectory are not independent

2. **An annoying bias:** there is a non-zero probability that some states are not visited at all. And it affects all future estimates

3. **Error propagation** from recursive estimation

Addressing Challenge 1: Define an appropriate martingale

- Consider the data collection in parallel
- Group all data for time h together
- Conditioning on the number of times states are visited

Addressing Challenge 2: Fictitious estimator technique

$$\widetilde{v}^{\pi} := \sum_{t=1}^{H} \sum_{s_t} \widetilde{d}_t^{\pi}(s_t) \widetilde{r}_t^{\pi}(s_t) \quad \text{where} \quad \widetilde{d}_t^{\pi} = \widetilde{\mathbb{P}}_{t,t-1}^{\pi} \widetilde{d}_{t-1}^{\pi}$$

$$\widetilde{r}_t^{\pi}(s_t) = \begin{cases} \widehat{r}_t^{\pi}(s_t) & \text{if } n_{s_t} \geq n d_t^{\mu}(s_t)(1-\delta) \\ r_t^{\pi}(s_t) & \text{otherwise;} \end{cases}$$

$$\widetilde{\mathbb{P}}_{t,t-1}^{\pi}(\cdot|s_{t-1}) = \begin{cases} \widehat{\mathbb{P}}_{t,t-1}^{\pi} & \text{if } n_{s_{t-1}} \geq n d_t^{\mu}(s_{t-1})(1-\delta) \\ \mathbb{P}_{t,t-1}^{\pi} & \text{otherwise.} \end{cases}$$

Multiplicative Chernoff Bound

Lemma A.1 (Multiplicative Chernoff bound [Chernoff et al., 1952]).
Let X be a Binomial random variable with parameter p, n.
For any $\delta > 0$, we have that

$$\mathbb{P}[X < (1-\delta)pn] < e^{-\frac{\delta^2 pn}{2}}$$

- Apply to our problem

Address Challenge 3: Empirical / Offline version of Bellman equation of variance

$$\mathrm{Var}[\widetilde{v}^{\pi}] = \sum_{h=0}^{H} \sum_{s_h} \mathbb{E}\left[\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h}} \mathbf{1}\left(n_{s_h} \geq \frac{n d_h^{\mu}(s_h)}{(1-\delta)^{-1}}\right)\right] \mathrm{Var}_{\mu}\left[\frac{\pi(a_h^{(1)}|s_h)}{\mu(a_h^{(1)}|s_h)}(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)}) \middle| s_h^{(1)} = s_h\right].$$

Bounding error propagation

$$\mathbb{E}\left[\frac{\widetilde{d}_h^\pi(s_h)^2}{n_{s_h}}\mathbf{1}\left(n_{s_h} \geq \frac{n d_h^\mu(s_h)}{(1-\delta)^{-1}}\right)\right] \leq \frac{(1-\delta)^{-1}}{n}\left(\frac{d_h^\pi(s_h)^2}{d_h^\mu(s_h)} + \mathrm{Var}\left[\widetilde{d}_h^\pi(s_h)\right]\right),$$

- Bounding the variance of is somewhat tedious
  - Requires use to bound the covariance

$$\mathrm{Var}[\widetilde{d}_h^\pi(s_h)] \leq \frac{2(1-\delta)^{-1} h d_h^\pi(s_h)}{n}.$$

Is MIS optimal for OPE in RL?

- It depends on the settings.

- For finite-state / infinite action space, we conjecture that it is.

(Still an open problem now.)

- For fully tabular setting, it is not optimal, at least asymptotically.

Tabular MIS

$$\widehat{v}^{\pi}_{\mathrm{MIS}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\widehat{d}^{\pi}_{t}(s_t^{(i)})}{\widehat{d}^{\mu}_{t}(s_t^{(i)})}\widehat{r}^{\pi}_{t}(s^{(i)}).$$

- With a minor change to the following recursive estimation

$$\widehat{P}^{\pi}_{t}(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}}\sum_{i=1}^{n}\frac{\pi(a_{t-1}^{(i)}|s_{t-1})}{\mu(a_{t-1}^{(i)}|s_{t-1})}\cdot \mathbf{1}((s_{t-1}^{(i)}, s_t^{(i)}, a_t^{(i)}) = (s_{t-1}, s_t, a_t));$$

$$\widehat{r}^{\pi}_{t}(s_t) = \frac{1}{n_{s_t}}\sum_{i=1}^{n}\frac{\pi(a_t^{(i)}|s_t)}{\mu(a_t^{(i)}|s_t)}r_t^{(i)}\cdot \mathbf{1}(s_t^{(i)} = s_t).$$

A short detour: How shall we do DM in RL?

- How would you do DM in this case?
  1. Estimate MDP
  2. Plug-in the target policy

TMIS is equivalent to DM --- a model-based approach

MSE of the TMIS / model-based OPE estimator

- Theorem 3.1 (Yin and W., 2020)

$$\mathbb{E}[(\widehat{v}^{\pi}_{\text{TMIS}} - v^{\pi})^2]$$

$$\leq \frac{1}{n} \sum_{h=0}^{H} \sum_{s_h, a_h} \frac{d_h^{\pi}(s_h)^2}{d_h^{\mu}(s_h)} \frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)} \text{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)}) \middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h \right]$$

$$+ O(n^{-1.5})$$

Yin & W. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *AISTATS-2020*

TMIS vs on-policy evaluation

Lemma 3.4. *For any policy* π *and any MDP.*

$$\mathrm{Var}_\pi\left[\sum_{t=1}^{H} r_t^{(1)}\right] = \sum_{t=1}^{H}\Bigg(\mathbb{E}_\pi\left[\mathrm{Var}\left[r_t^{(1)} + V_{t+1}^{\pi}(s_{t+1}^{(1)})\middle| s_t^{(1)}, a_t^{(1)}\right]\right]$$
$$+ \mathbb{E}_\pi\left[\mathrm{Var}\left[\mathbb{E}[r_t^{(1)} + V_{t+1}^{\pi}(s_{t+1}^{(1)}) | s_t^{(1)}, a_t^{(1)}]\middle| s_t^{(1)}\right]\right]\Bigg).$$

Combined with the previous observation:

1. TMIS has an error that is linear in H.

2. TMIS is better than MC even when we are doing on-policy evaluation

Fitted Q Iterations

- Recall Bellman Optimality equation and the Bellman operator

$$\mathcal{T}f(s,a) := r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} f(s',a').$$

- Given offline transition data and a function class

$$\text{FQI:} \quad f_t \in \mathrm{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} \left(f(s_i', a_i) - r_i - \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s_i, a_i) \right)^2 .$$

Iteratively from some initialization.

- For the finite horizon episodic case:

Fitted Q iterations for OPE

- Recall Bellman equation for a fixed policy

$$Q_{h-1}^{\pi}(s,a) = r(s,a) + \mathbb{E}\left[V_h^{\pi}(s') \,\middle|\, s,a\right]$$

- Given offline transition data and a function class

$\widehat{Q}_{H+1}^{\pi} := 0$ and for $h = H, H-1, \ldots, 0$,

$$\hat{Q}_h^{\pi} = \arg\min_{f_h \in \mathcal{F}} \sum_{i=1}^{n} \left(f_h(s_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - \sum_{a' \in \mathcal{A}} \pi(a' | s_{h+1}^{(i)}) f_{h+1}(s_{h+1}^{(i)}, a') \right)^2$$

FQI in the tabular case

$$\hat{Q}_h^\pi = \arg\min_{f_h \in \mathcal{F}} \sum_{i=1}^{n} \left(f_h(s_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - \sum_{a' \in \mathcal{A}} \pi(a' | s_{h+1}^{(i)}) f_{h+1}(s_{h+1}^{(i)}, a') \right)^2$$

- Let's work out the optimal solution!

In conclusion, in the tabular MDP case, they are all equivalent.

- TMIS

$$\widehat{v}^{\pi}_{\text{MIS}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\widehat{d}^{\pi}_t(s_t^{(i)})}{\widehat{d}^{\mu}_t(s_t^{(i)})}\widehat{r}^{\pi}_t(s^{(i)}).$$

- Model-based Plugin

$$\hat{v}^{\pi}_{\text{DM}} = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\hat{d}^{\pi}_h(s)\hat{r}^{\pi}_h(s)$$

- Fitted Q Iteration

$$\hat{v}^{\pi}_{\text{FQI}} = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\hat{d}_1(s)\pi(a|s)\hat{Q}_1(s,a)$$