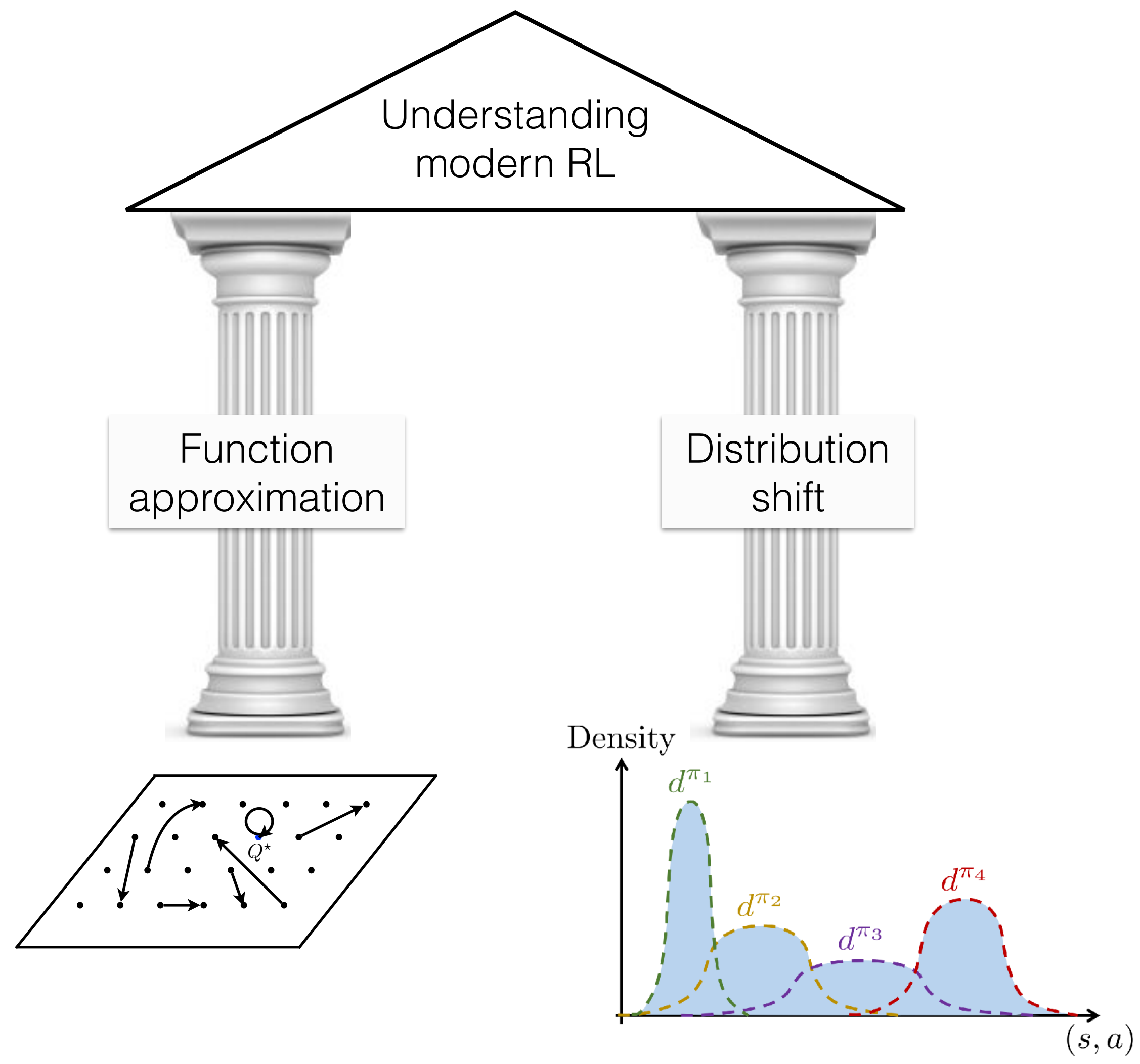
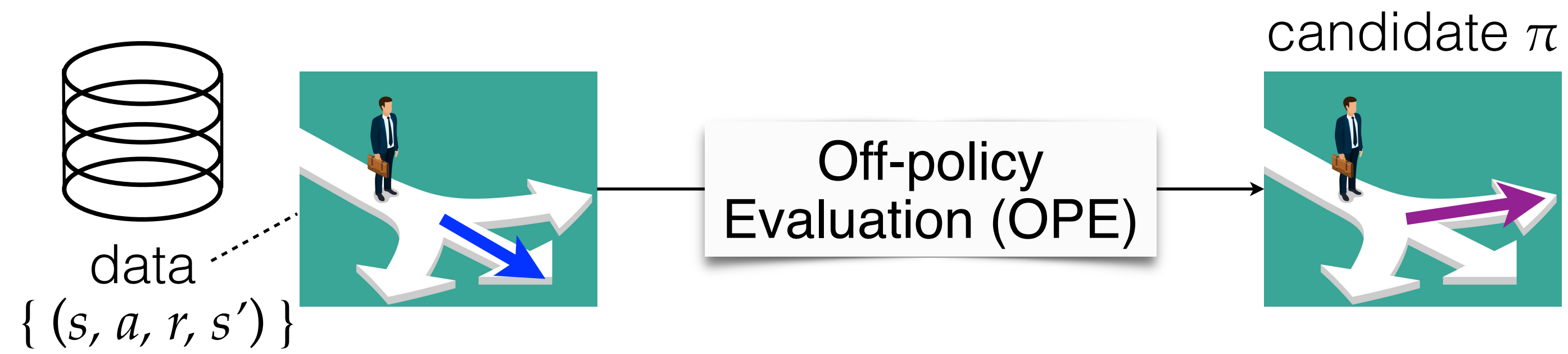


Off-policy Evaluation via Recursive Reweighting and Moment Matching

Nan Jiang
UIUC

@IMSI Program on Theoretical Advances in RL and Control
Workshop on RL from Offline Data and Human Feedback
Apr. 2026



Markov Decision Process (MDP)

- For $h = 0, 1, 2, \dots$, the agent

- observes **state** $s_h \in S$

layered:
 $\mathcal{S} = \bigcup_h \mathcal{S}_h$

transition dynamics
 $P: S \times A \rightarrow \Delta(S)$

- chooses **action** $a_h \in A$

- receives **reward** $r_h \sim R(s_h, a_h)$

reward function
 $R: S \times A \rightarrow \Delta([0,1])$

- Policy $\pi: S \rightarrow \Delta(A)$

- Wlog: assume (s_0, a_0) is fixed and $r_0 = 0$

$$s_1 \sim P(\cdot | s_0, a_0), a_1 \sim \pi(\cdot | s_1), r_1 \sim R(s_1, a_1), \dots, s_H, a_H, r_H$$

- $\mathbb{E}_\pi[\cdot]$: expectation wrt above distribution
- OPE task: estimate $J(\pi) = \mathbb{E}_\pi[\sum_h r_h]$, where $\sum_h r_h \in [0, V_{\max}]$
using data trajectories sampled from some **behavior** policy

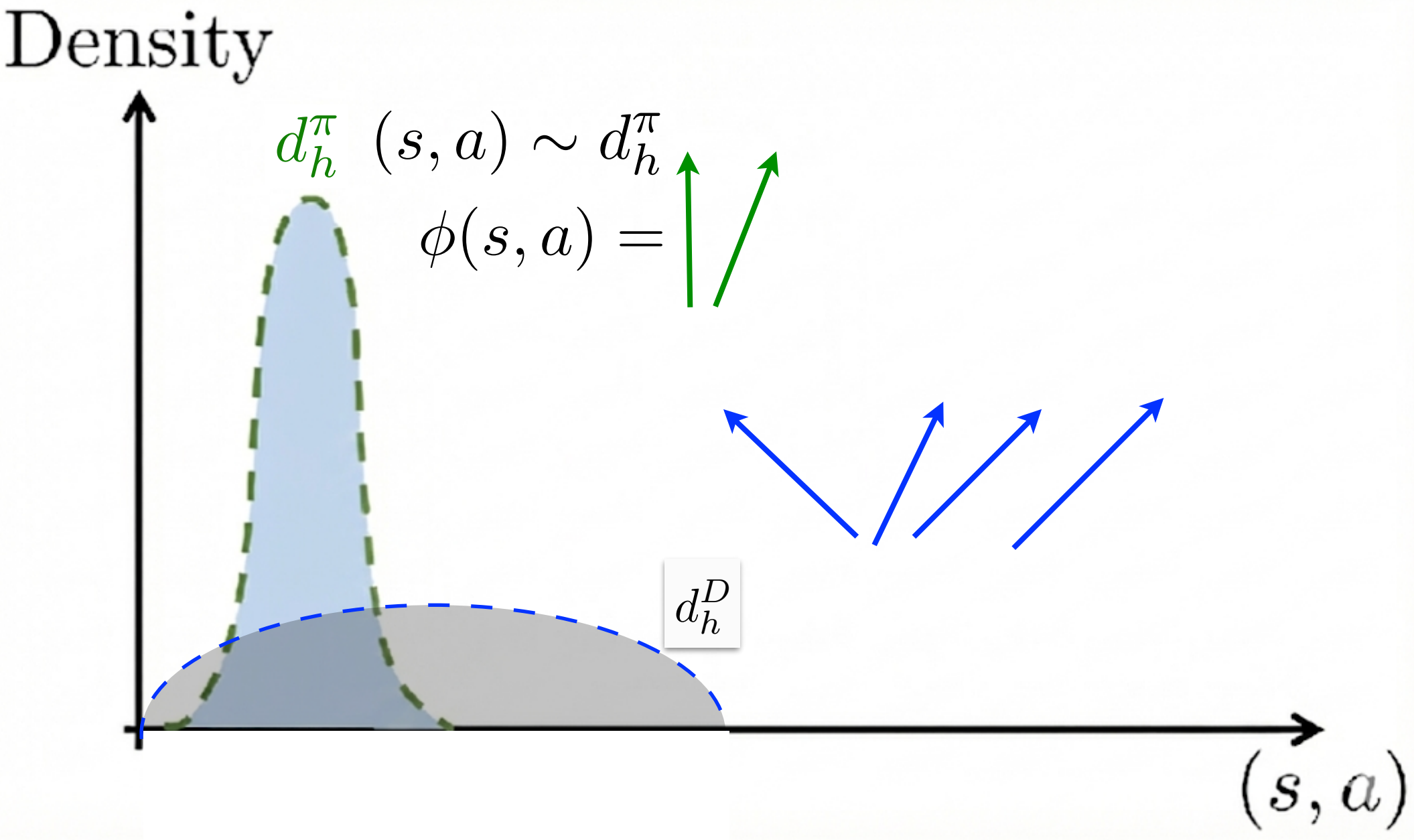
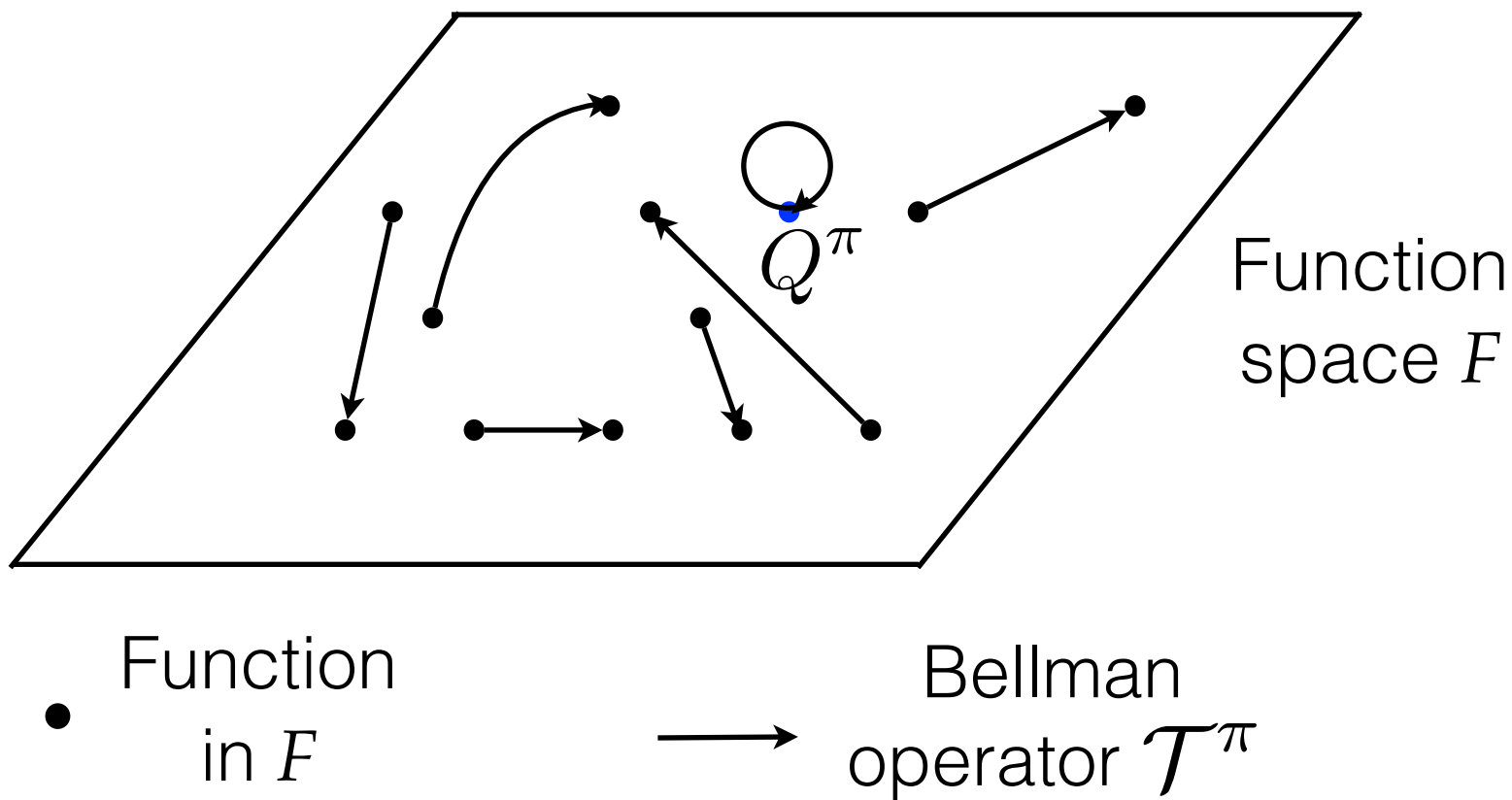
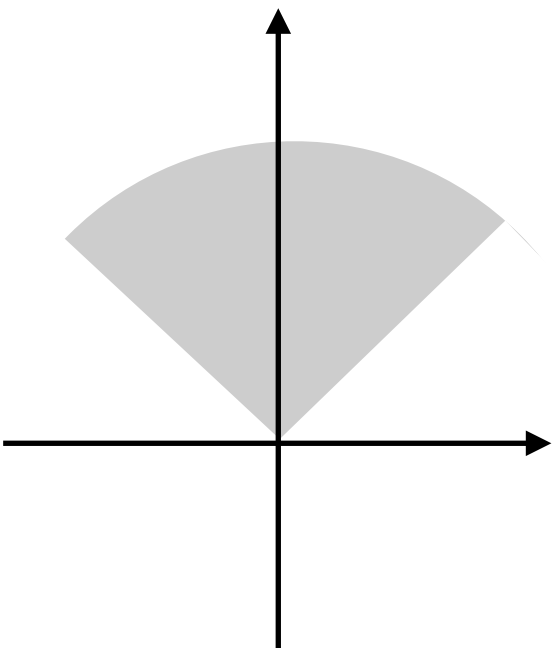
OPE via Value-function Approximation

- Occupancy distributions
 - d_h^π : distribution of (s_h, a_h) under π
 - d_h^D : distribution of (s_h, a_h) in data
- Function class $\mathcal{F} = \{\mathcal{F}_h\}_h$ for modeling Q^π : $Q_{h+1}^\pi = \mathcal{T}^\pi Q_h^\pi$
 $(\mathcal{T}^\pi f)(s, a) = \mathbb{E}_{r \sim R(s, a), s' \sim P(\cdot | s, a)}[r + f(s', \pi)]$
- FQE: $\hat{Q}_h^\pi \leftarrow \arg \min_{f_\theta} \hat{\mathbb{E}}_D[(f_\theta(s_h, a_h) - r_h - \hat{Q}_{h+1}^\pi(s_{h+1}, \pi))^2]$

Finite-sample Analyses of FQE

- Completeness ?
- d_h^D “covers” d_h^π ?
- $\log |\mathcal{F}|$?

- Bellman completeness $\mathcal{T}^\pi f \in \mathcal{F}_h, \forall f \in \mathcal{F}_{h+1}$
- (Density) Coverage $\mathbb{E}_{d_h^D} [(d_h^\pi / d_h^D)^2]$
 - Generally $\sup_{f, f' \in \mathcal{F}_h} (\mathbb{E}_{d_h^\pi} [f - f'])^2 / \mathbb{E}_{d_h^D} [(f - f')^2]$
 - F linear in ϕ : $\|\mathbb{E}_{d_h^\pi} [\phi]\|_{\Sigma^{-1}}^2$
- Final bound depends on coverage and $\log |\mathcal{F}|$ (or covering number, etc.)



Linear Bandit (H=1)

- Completeness
- d_h^D “covers” d_h^π
- $\log |\mathcal{F}|$

- Given data $\{(s^{(i)}, a^{(i)}, r^{(i)})\}_{i=1}^n$, and $\mathbb{E}[r|s, a] = \phi(s, a)^\top \theta^*$

- Estimate $J(\pi) = \mathbb{E}_\pi[r] = (\mathbb{E}_\pi[\phi(s, a)])^\top \theta^*$

- Given $\{(x_i, y_i)\}_{i=1}^n$, where $x_i = \phi(s^{(i)}, a^{(i)})$, $y_i = r^{(i)}$

$$\hat{\Sigma} = \frac{1}{n} \sum_i x_i x_i^\top$$

- Predict y for $x_{\text{new}} = \mathbb{E}_\pi[\phi(s, a)]$

- FQE $\rightarrow$ linear regression: error $\leq \|x_{\text{new}}\|_{\hat{\Sigma}^{-1}} \sqrt{\mathbf{d} + \log(1/\delta)}$

- Can drop $\mathbf{d}$!

- LR prediction: $\hat{y}_{\text{new}} = \frac{1}{n} \sum_i \underbrace{(x_{\text{new}}^\top \hat{\Sigma}^{-1} x_i)}_{\hat{w}^{(i)}} y_i$, where $\hat{w}^{(i)}$ is the solution of

$$\arg \min_{w^{[n]}} \left\| \frac{1}{n} \sum_i w^{(i)} x_i - x_{\text{new}} \right\|_2$$

$$= \arg \min_{w^{[n]}} \sup_{f \in \mathcal{F}_{\text{lin}}} \left| \frac{1}{n} \sum_i w^{(i)} f(x_i) - f(x_{\text{new}}) \right|$$

linear class

(minimizer is not unique)

0 mean
conditioned on x_i

- $\hat{w}^{[n]}$ gets 0 loss (when $\hat{\Sigma}$ invertible)

$$\hat{y}_{\text{new}} - x_{\text{new}}^\top \theta^* = \frac{1}{n} \sum_{i=1} \hat{w}^{(i)} \underbrace{(y_i - x_i^\top \theta^*)}_{\text{0 mean conditioned on } x_i}$$

$$\text{Error} \propto \sqrt{\frac{1}{n} \sum_i (\hat{w}^{(i)})^2} = \|x_{\text{new}}\|_{\hat{\Sigma}^{-1}}$$

Linear FQE

- Completeness $\Rightarrow \|B_h^\pi\|_{\text{op}} \leq 1$
- d_h^D “covers” ~~d_h^π~~ $\Rightarrow \psi_h$
- $\log |\mathcal{F}| \Rightarrow$ not in leading term

• FQE: $\hat{Q}_h^\pi \leftarrow \arg \min_{f_\theta} \hat{\mathbb{E}}_D[(f_\theta(s_h, a_h) - r_h - \hat{Q}_{h+1}^\pi(s_{h+1}, \pi))^2]$

• Dim-free result is known for linear FQE [Duan et al'20]

• Only realizability ($Q^\pi \in \mathcal{F}$); no completeness needed [Perdomo et al'23, Amortila et al'26]

• FQE implicitly learns: $\phi(s_{h+1}, \pi) \approx B_h^\pi \phi(s_h, a_h)$

• A linear dyn sys: $\psi_{h+1} = B_h^\pi \psi_h$

• “Intention”: $\psi_h \approx \mathbb{E}_{d_h^\pi}[\phi]$

$$B_h^\pi = \Sigma_h^{-1} \Sigma_{h,h+1}^{\text{cr}}$$

lin approx
of reward

• $\hat{J}(\pi) = \sum_h \psi_h^\top \theta_R \approx \sum_h \mathbb{E}_{d_h^\pi}[\phi^\top \theta_R]$

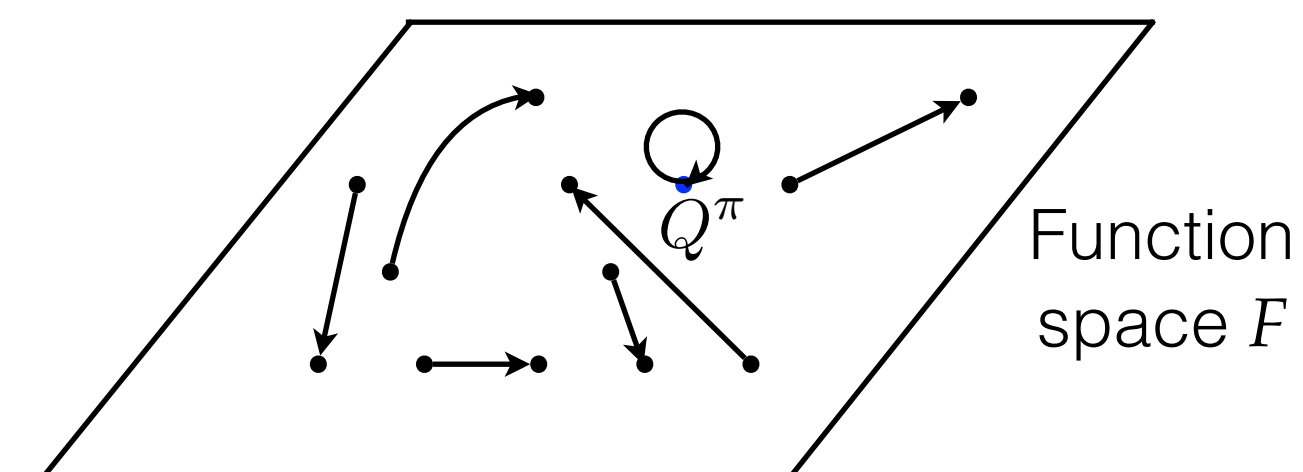
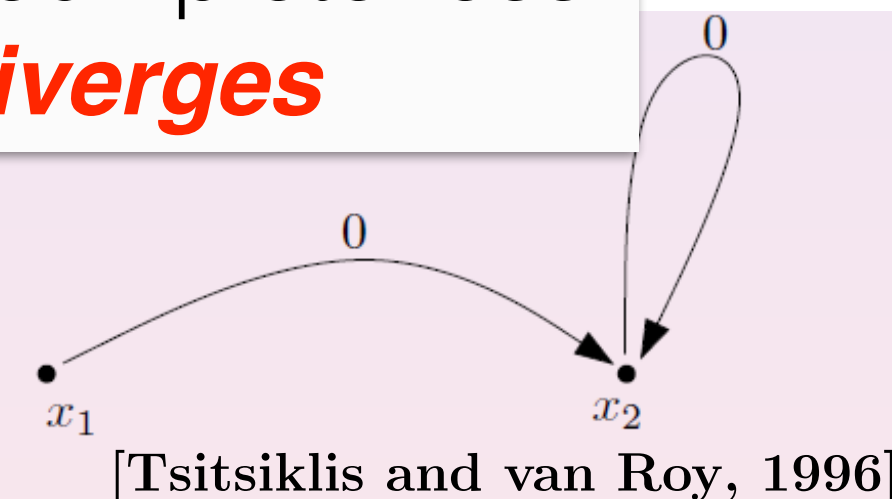
• Guarantee depends on coverage parameter: $\|\psi_h\|_{\Sigma_h^{-1}}$

• ψ_h may be wrong & can blow-up *exponentially* **||**

• Completeness: both $\approx$ become exact $\Rightarrow \|\mathbb{E}_{d_h^\pi}[\phi]\|_{\Sigma_h^{-1}}$

TD w/o completeness

diverges



• Function in $F \rightarrow$ Bellman operator $\mathcal{T}^\pi$

Duan et al. ICML-20. Minimax-optimal off-policy evaluation with linear function approximation.

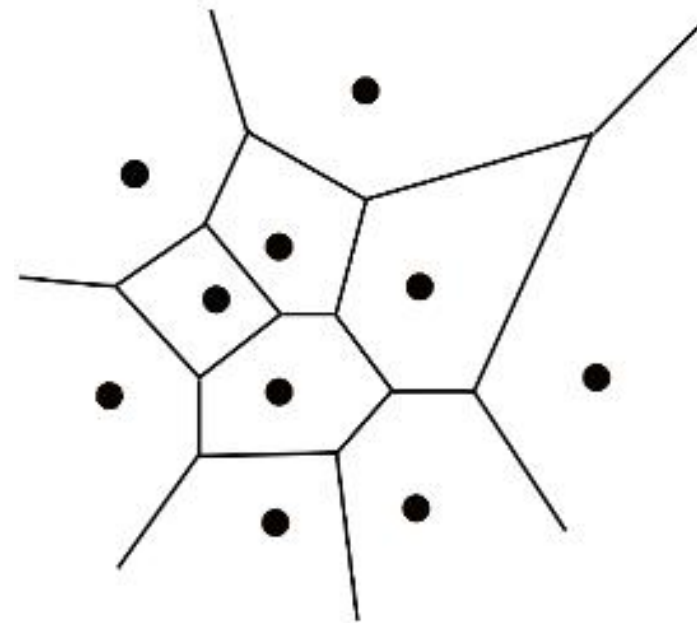
Perdomo et al. JMLR-23. A Complete Characterization of Linear Estimators for Offline Policy Evaluation.

Amortila et al. ICLR-26. A Unifying View of Coverage in Linear Off-Policy Evaluation.

More on $\psi_h \neq \mathbb{E}_{d_h^\pi}[\phi]$

- Consider 1-hot $\phi = [0, \dots, 1, \dots, 0]^T$
 - realizable $\Rightarrow$ aggregating (s, a) pairs w/ equal Q^π
- $\mathbb{E}[\phi(s_{h+1}, \pi) | s_h, a_h] \approx B_h^\pi \phi(s_h, a_h)$
 - B_h^π : tabular MRP under aggregation
 - $\psi_{h+1} = B_h^\pi \psi_h$: π 's occupancy in B_h^π

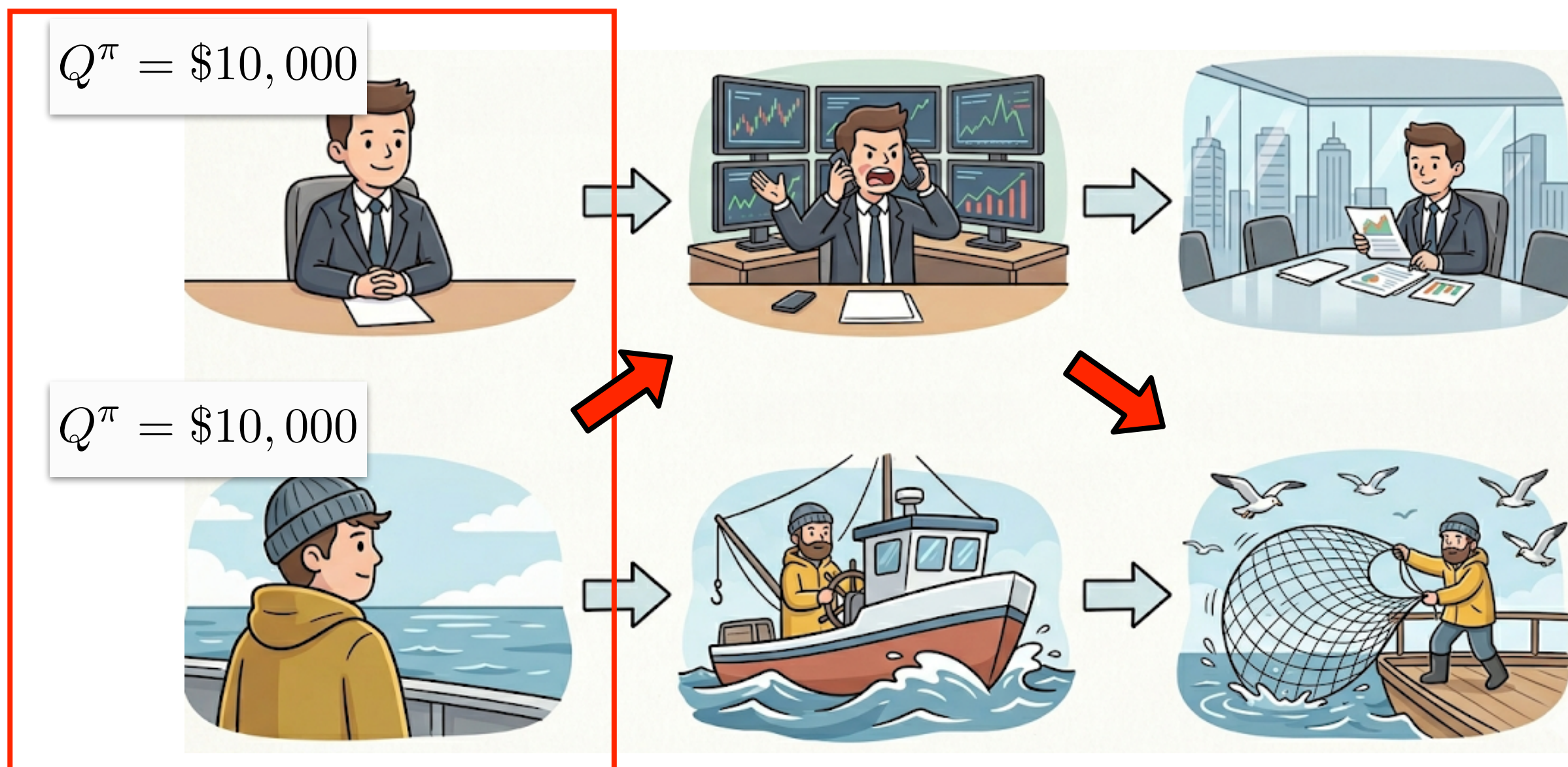
Proof sketch



preserves Q^π

- processed data $(\phi(s, a), r, \phi(s', \pi))$ can be viewed as gen. from B_h^π

- As if learning entirely occurs in B_h^π



$\|\mathbb{E}_{d_h^\pi}[\phi]\|_{\Sigma_h^{-1}}$ **can't** be the right concept

- $\|d_h^\pi / d_h^D\|_\infty$ is stronger
- yet lower bound exists under realizability [Foster et al'22]

General function approximation?

Q-MMR for MDPs

- OPE: estimate $J(\pi)$ from data $\left\{ (s_0^{(i)}, a_0^{(i)}, r_0^{(i)}, \dots, s_H^{(i)}, a_H^{(i)}, r_H^{(i)}) \right\}_{i=1}^n$
- (Wrong but useful) intuition: learn $\hat{w}_h^{(i)} \approx d_h^\pi(s_h^{(i)}, a_h^{(i)}) / d_h^D(s_h^{(i)}, a_h^{(i)})$
 - so $\hat{J}(\pi) := \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^H \hat{w}_h^{(i)} r_h^{(i)}$
- inductively, $\hat{w}_h^{(i)}$ -weighted $(s_h^{(i)}, a_h^{(i)})$ is a particle rep of d_h^π
 - then, $\hat{w}_h^{(i)}$ -weighted $(s_{h+1}^{(i)}, \pi) \Rightarrow d_{h+1}^\pi$
 - “distill” to $\hat{w}_{h+1}^{(i)}$ -weighted $(s_{h+1}^{(i)}, a_{h+1}^{(i)})$

Algorithm 1 Q-MMR (Q-function-based Moment-Matched Reweighting)

- 1: Initialize $\hat{w}_0^{(i)} \equiv 1, \forall i$.
- 2: **for** $h = 0$ to $H - 1$ **do**
- 3: Choose $\hat{w}_{h+1}^{[n]}$ as the $w_{h+1}^{[n]}$ that (approximately) minimizes the loss

$$\hat{\mathcal{L}}_{h+1}(w_{h+1}^{[n]}; \hat{w}_h^{[n]}) := \sup_{f \in \mathcal{F}_{h+1}} \left| \frac{1}{n} \sum_i w_{h+1}^{(i)} f(s_{h+1}^{(i)}, a_{h+1}^{(i)}) - \frac{1}{n} \sum_i \hat{w}_h^{(i)} f(s_{h+1}^{(i)}, a_{h+1}^{(i)}) \right|$$

- (Optional) Among $\hat{w}_{h+1}^{[n]}$ with the same loss, choose the one that minimizes $\|\hat{w}_{h+1}^{[n]}\|_{[n]} := \sqrt{\frac{1}{n} \sum_i (\hat{w}_{h+1}^{(i)})^2}$
- 4: **end for**

Analysis of Q-MMR?

- Goal: $\hat{J}(\pi) := \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^H \hat{w}_h^{(i)} r_h^{(i)} \approx J(\pi)$
- Wrong intuition: $\hat{w}_h^{(i)} \approx d_h^\pi(s_h^{(i)}, a_h^{(i)}) / d_h^D(s_h^{(i)}, a_h^{(i)})$
 - Bounding their diff requires very strong assumption
 - Even embedding $\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \hat{w}_h^{(i)} f - \mathbb{E}_{d_h^\pi}[f] \right|$ requires completeness
- How to deal with #params = #data points?

Algorithm 1 Q-MMR (Q-function-based Moment-Matched Reweighting)

- 1: Initialize $\hat{w}_0^{(i)} \equiv 1, \forall i$.
- 2: **for** $h = 0$ to $H - 1$ **do**
- 3: Choose $\hat{w}_{h+1}^{[n]}$ as the $w_{h+1}^{[n]}$ that (approximately) minimizes the loss

$$\hat{\mathcal{L}}_{h+1}(w_{h+1}^{[n]}; \hat{w}_h^{[n]}) := \sup_{f \in \mathcal{F}_{h+1}} \left| \frac{1}{n} \sum_i w_{h+1}^{(i)} f(s_{h+1}^{(i)}, a_{h+1}^{(i)}) - \frac{1}{n} \sum_i \hat{w}_h^{(i)} f(s_{h+1}^{(i)}, \pi) \right|.$$

- (Optional) Among $\hat{w}_{h+1}^{[n]}$ with the same loss, choose the one that minimizes $\|\hat{w}_{h+1}^{[n]}\|_{[n]} := \sqrt{\frac{1}{n} \sum_i (\hat{w}_{h+1}^{(i)})^2}$
- 4: **end for**

$$\begin{aligned}
& \left| \sum_{h=0}^H \frac{1}{n} \sum_{i=1}^n \hat{w}_h^{(i)} \boxed{r_h^{(i)}} - J(\pi) \right| \quad \begin{array}{l} 0 \text{ mean, i.i.d.} \\ \text{conditioned on } (s_h^{(i)}, a_h^{(i)}) \end{array} \\
& \leq \left| \sum_{h=0}^H \frac{1}{n} \sum_{i=1}^n \underline{\hat{w}_h^{(i)}} \left(\boxed{Q^\pi(s_h^{(i)}, a_h^{(i)})} - \underline{Q^\pi(s_{h+1}^{(i)}, \pi)} \right) - J(\pi) \right| + \\
& \leq \left| \sum_{h=0}^H \frac{1}{n} \sum_{i=1}^n \left(\hat{w}_h^{(i)} Q^\pi(s_h^{(i)}, a_h^{(i)}) - \underline{\hat{w}_{h+1}^{(i)} Q^\pi(s_{h+1}^{(i)}, a_{h+1}^{(i)})} \right) - J(\pi) \right| + \\
& \qquad \qquad \qquad + \sum_{h=0}^H \epsilon_h^{\text{stat}} \\
& = \sum_{h=0}^{H-1} \hat{\mathcal{L}}_{h+1}(\hat{w}_{h+1}^{[n]}; \hat{w}_h^{[n]}) + \sum_{h=0}^H \epsilon_h^{\text{stat}}
\end{aligned}$$

$$\propto \sqrt{\frac{1}{n} \sum_i (\hat{w}_h^{(i)})^2}$$

Linear case

- $\hat{\mathcal{L}} = 0$ if $\hat{\Sigma}_h$ invertible
- $\sqrt{\frac{1}{n} \sum_i (\hat{w}_h^{(i)})^2} = \|\hat{\psi}_h\|_{\hat{\Sigma}_h^{-1}}$

Algorithm 1 Q-MMR (Q-function-based Moment-Matched Reweighting)

- 1: Initialize $\hat{w}_0^{(i)} \equiv 1, \forall i$.
- 2: **for** $h = 0$ to $H - 1$ **do**
- 3: Choose $\hat{w}_{h+1}^{[n]}$ as the $w_{h+1}^{[n]}$ that (approximately) minimizes the loss

$$\hat{\mathcal{L}}_{h+1}(w_{h+1}^{[n]}; \hat{w}_h^{[n]}) := \sup_{f \in \mathcal{F}_{h+1}} \left| \frac{1}{n} \sum_i \boxed{w_{h+1}^{(i)} f(s_{h+1}^{(i)}, a_{h+1}^{(i)})} - \frac{1}{n} \sum_i \boxed{\hat{w}_h^{(i)} f(s_{h+1}^{(i)}, \pi)} \right|.$$

- (Optional) Among $\hat{w}_{h+1}^{[n]}$ with the same loss, choose the one that minimizes $\|\hat{w}_{h+1}^{[n]}\|_{[n]} := \sqrt{\frac{1}{n} \sum_i (\hat{w}_{h+1}^{(i)})^2}$
- 4: **end for**

Data-independent (“random design”) bound

- Define $w_h^* \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ as the population counterpart of $\hat{w}_h^{[n]} \in \mathbb{R}^n$
- Replace $\sqrt{\frac{1}{n} \sum_i (\hat{w}_h^{(i)})^2}$ with $\sqrt{\mathbb{E}_{d_h^D} [(w^*)^2]}$
 - $\|\hat{w}_h^{[n]}\|_{[n]} \sqrt{1/n} = \|w^*\|_{2, d_h^D} \sqrt{1/n} + \boxed{(\|\hat{w}_h^{[n]}\|_{[n]} - \|w^*\|_{2, d_h^D}) \sqrt{1/n}}$
 - as long as $\hat{w}_h^{[n]}$ tracks w_h^* , produces a higher-order term
- Linear: need bounded leverage $\sup_{s_h, a_h} \|\Sigma_h^{-1/2} \phi(s_h, a_h)\|_2$
- general: $\sup_{f \in \text{span}(\mathcal{F}_h), s_h, a_h} |f(s, a)| / \sqrt{\mathbb{E}_{d_h^D} [f^2]}$

Algorithm 1 Q-MMR (Q-function-based Moment-Matched Reweighting)

- 1: Initialize $\hat{w}_0^{(i)} \equiv 1, \forall i$.
- 2: **for** $h = 0$ to $H - 1$ **do**
- 3: Choose $\hat{w}_{h+1}^{[n]}$ as the $w_{h+1}^{[n]}$ that (approximately) minimizes the loss

$$\hat{\mathcal{L}}_{h+1}(w_{h+1}^{[n]}; \hat{w}_h^{[n]}) := \sup_{f \in \mathcal{F}_{h+1}} \left| \frac{1}{n} \sum_i w_{h+1}^{(i)} f(s_{h+1}^{(i)}, a_{h+1}^{(i)}) - \frac{1}{n} \sum_i \hat{w}_h^{(i)} f(s_{h+1}^{(i)}, \pi) \right|.$$

- (Optional) Among $\hat{w}_{h+1}^{[n]}$ with the same loss, choose the one that minimizes $\|\hat{w}_{h+1}^{[n]}\|_{[n]}$.
- 4: **end for**

	General + realizability		General + completeness		Linear + realizability		Linear + completeness	
	Coverage	Dim-free?	Coverage	Dim-free?	Coverage	Dim-free?	Coverage	Dim-free?
FQE	N/A		$\sup_{f, f' \in \mathcal{F}_h} \frac{ \mathbb{E}_{d_h^\pi}[f - f'] }{\sqrt{\mathbb{E}_{d_h^D}[(f - f')^2]}}$	✗	N/A		$\ \mathbb{E}_{d_h^\pi}[\phi]\ _{\Sigma_h^{-1}}$	✗
Linear FQE			N/A					
Ours			$\sup_{f \in \text{span}(\mathcal{F}_h)} \frac{ \mathbb{E}_{d_h^\pi}[f] }{\sqrt{\mathbb{E}_{d_h^D}[f^2]}}$	✓	$\ \psi_h\ _{\Sigma_h^{-1}}$	✓	$\ \mathbb{E}_{d_h^\pi}[\phi]\ _{\Sigma_h^{-1}}$	✓



Philip Amortila



Audrey Huang



Akshay
Krishnamurthy



Xiang Li

Amortila et al. ICLR-26. A Unifying View of Coverage in Linear Off-Policy Evaluation.

Li & Jiang. Q-MMR: Off-policy Evaluation via Recursive Reweighting and Moment Matching.