Asymptotically Exact Error Characterization of Offline Policy Evaluation with Misspecified Linear Models

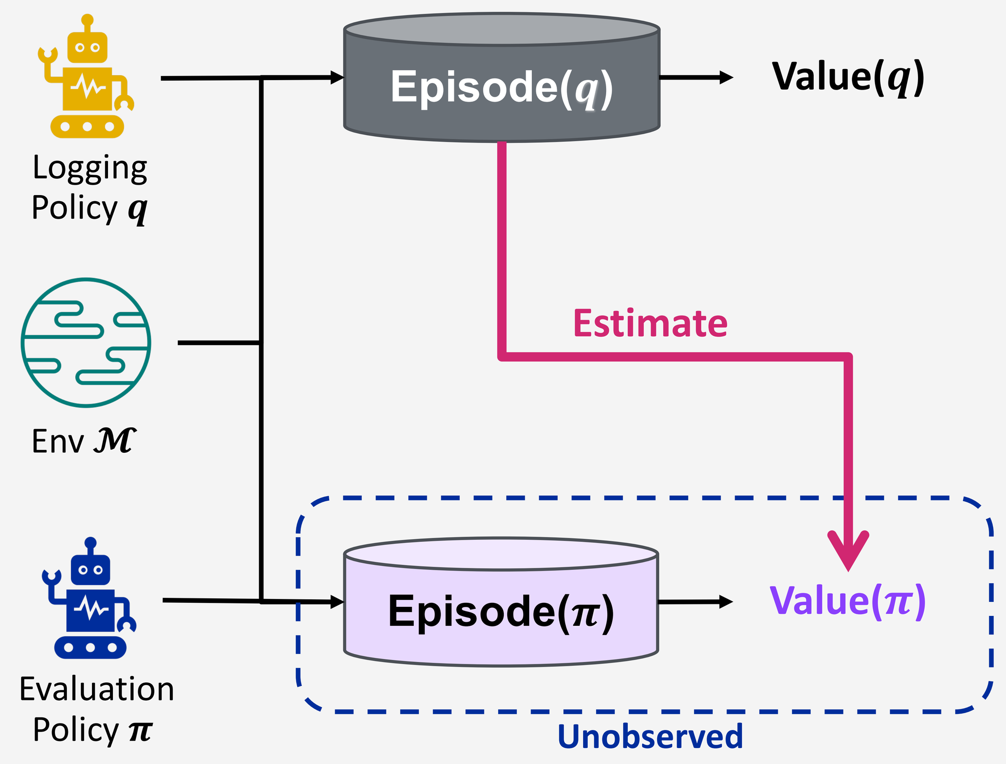
Kohei Miyaguchi, IBM Research
miyaguchi@ibm.com

Message

Linear DM is **doubly robust.**

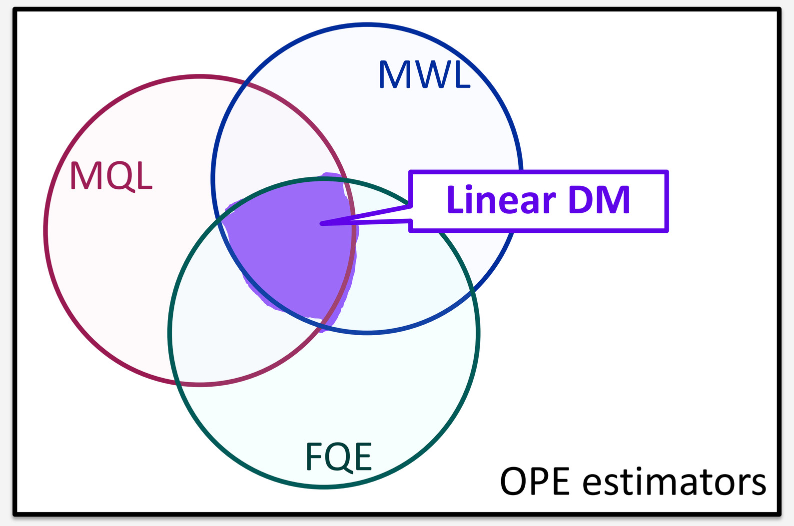
Contribution summary

Offline policy evaluation (OPE)



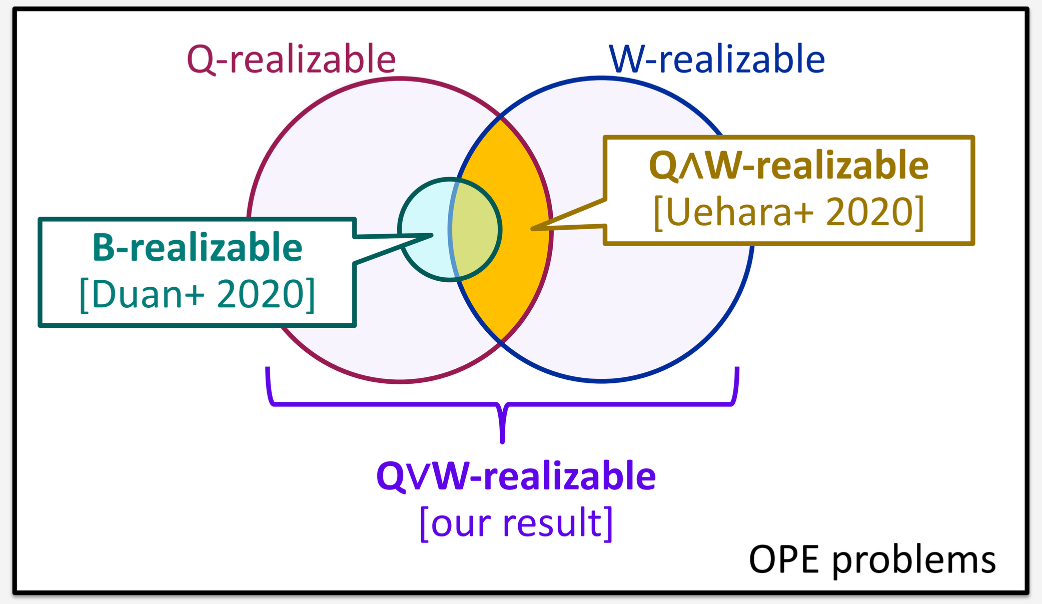
- OPE = estimating policy value with other policies' episodes.
- Useful for pre-deployment performance analysis & hyperparameter tuning.

Previous studies & our contribution



- ← **Linear direct method (DM)** is a fundamental OPE method with linear function approximation.
- However, it has been not well understood **when and how exactly linear DM works**.

- Our contribution is the exact asymptotic analysis of linear DM.
- → As a result, linear DM is shown to be **more robust than previously known**: *Either* Q- *or* W-realizability is sufficient to solve OPE.



Main result

Algorithm: Linear DM

Input: data D_n, policy π, initial density p_0, discount factor γ, feature ϕ.
Output: policy value estimate $\hat{J}(\pi)$

1. $\hat{Q} \coloneqq \mathrm{LSTDQ}(D_n, \pi, \gamma, \phi)$
2. $\hat{J}(\pi) \coloneqq \hat{Q}(p_0, \pi)$

Vector-valued function $(s,a) \mapsto \phi(s,a) \in \mathbb{R}^K$

Theorem (asymptotic error)

- If ϕ is compatible (and standard assumptions of OPE hold), then:

$$\underbrace{\hat{J}(\pi) - J(\pi)}_{\text{Error of linear DM}} = -\frac{1}{1-\gamma}\mathbb{E}^q\left[\underbrace{\mathcal{R}_W(s,a)}_{\text{W-residual}}\underbrace{\mathcal{R}_Q(s,a)}_{\text{Q-residual}}\right] + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

Definition (residual functions)

- W-residual is given by **the projection residual of the marginal density ratio** of Episode(π) to Episode(q).

$$\mathcal{R}_W(s,a) \coloneqq \underbrace{w^\pi(s,a)}_{\text{Marginal density ratio}} - \underbrace{w^\#(s,a)}_{\text{Projection of } w^\pi \text{ onto } \phi}$$

- Q-residual is given by **the projection residual of the Q-function** of π.

$$\mathcal{R}_Q(s,a) \coloneqq \underbrace{B^\pi}_{\text{Bellman operator}} \underbrace{Q^\#(s,a)}_{\text{Projection of } Q^\pi \text{ onto } \phi} - \underbrace{Q^\#(s,a)}_{\text{Projection of } Q^\pi \text{ onto } \phi}$$

Implication (double robustness)

- The residual functions being zero are equivalent to W- and Q-realizabilities.

$$\mathcal{R}_W(\cdot,\cdot) \equiv 0 \Leftrightarrow w^\pi \in \mathrm{span}(\phi)$$
$$\mathcal{R}_Q(\cdot,\cdot) \equiv 0 \Leftrightarrow Q^\pi \in \mathrm{span}(\phi)$$

- Linear DM is doubly robust and **could be consistent even if LSTDQ is not**, i.e., $Q^\pi \notin \mathrm{span}(\phi)$.

$$w^\pi \in \mathrm{span}(\phi) \vee Q^\pi \in \mathrm{span}(\phi)$$
$$\Rightarrow \hat{J}(\pi) - J(\pi) = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

Key assumption

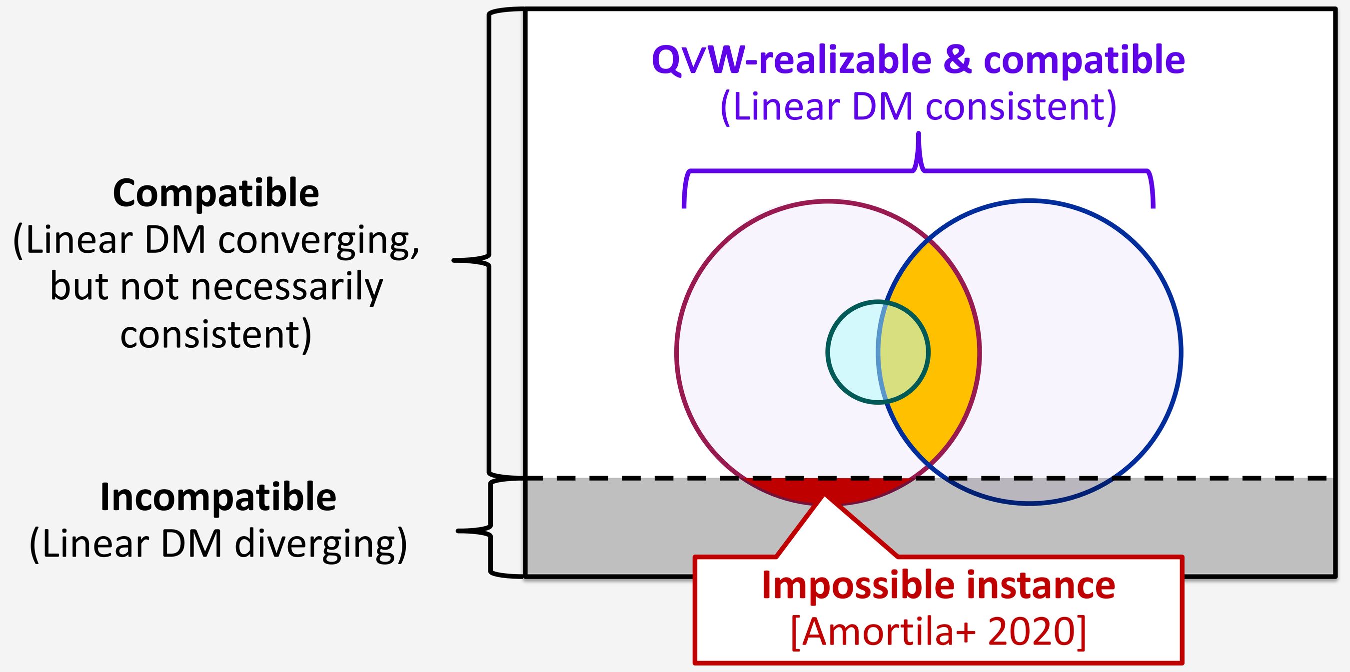
Definition (compatibility)

- ϕ is compatible iff $\mathrm{LSTDQ}(D_n, \pi, \gamma, \phi)$ converges as $n \to \infty$, i.e.,

$$\det \mathbb{E}^q[\phi(\phi - \gamma P^\pi \phi)^\top] \neq 0$$

Properties

- Compatibility is **always satisfied** with onehot features ϕ.
- Compatibility is **statistically testable** via concentration of determinant.
- Compatibility is **non-vacuous**: It eliminates the hard instance of [Amortila+2020].



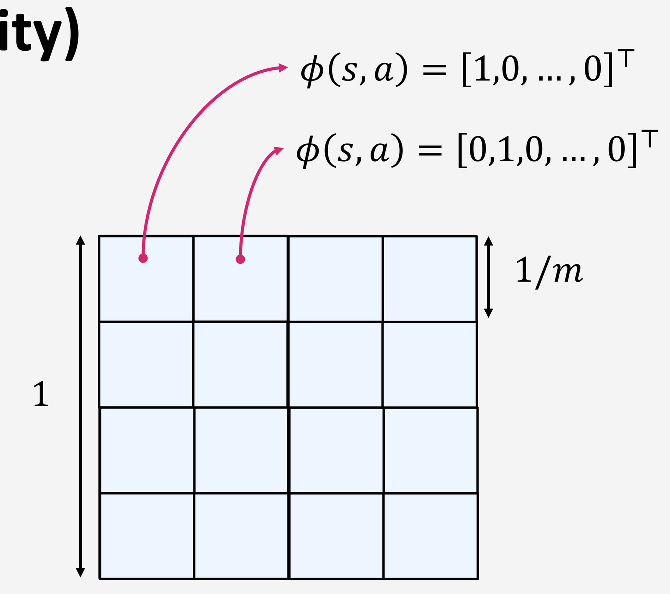
Extended result

Proposition (consistency *without* realizability)

- Linear DM with the homogeneous tile-coding feature ϕ is consistent if

1. Episode(q) is sufficiently strongly mixing and near-uniform.
2. The number of tiles grows steadily, e.g.,

$$K = \Theta(\sqrt{n})$$



Homogeneous $K(=m^d)$-tile coding of state-action space

Open questions

- Double robustness of nonlinear estimators (e.g., MQL, MWL, FQE)
- Nonparametric consistency with non-onehot features (e.g., NTK)