OptiDICE: Offline Policy Optimization via Stationary Distribution Correction Estimation

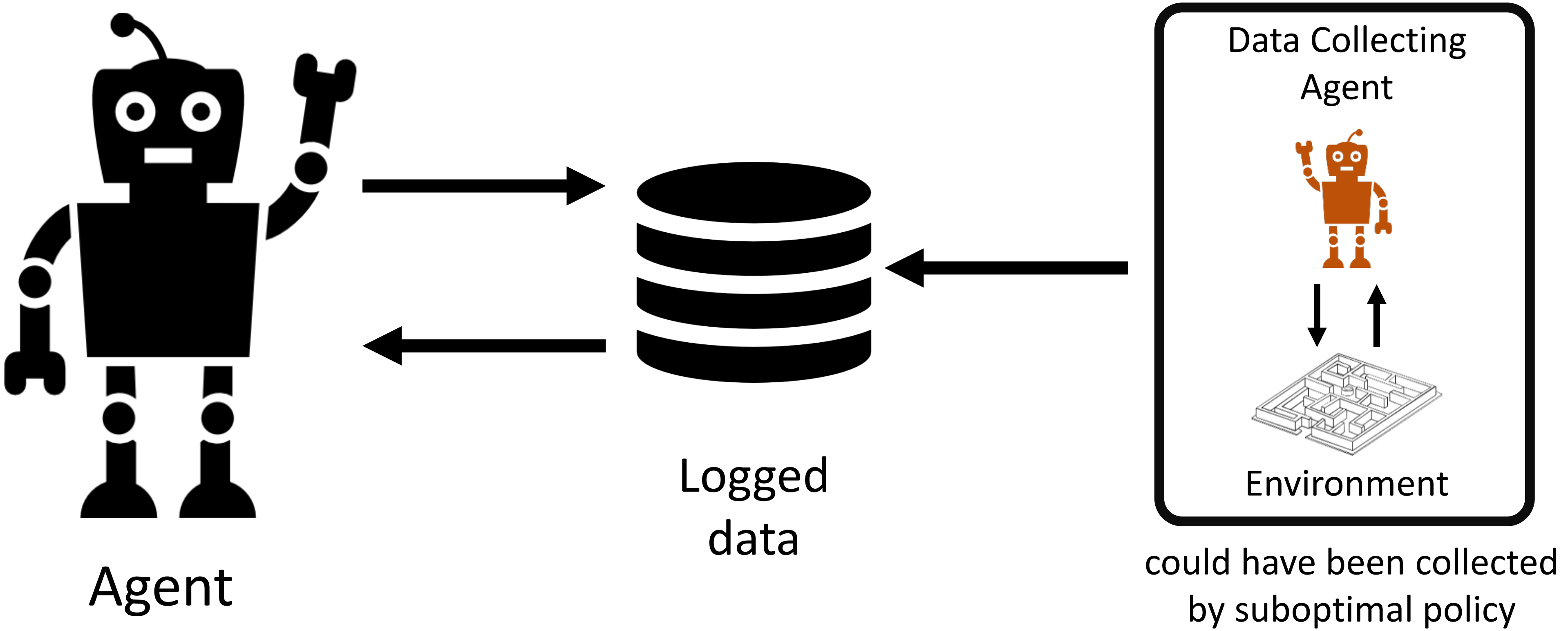
ICML 2021

Jongmin Lee*[1], **Wonseok Jeon***[2,3], Byung-Jun Lee[1,4], Joelle Pineau[2,3,5], Kee-Eung Kim[1]
(*Equal contribution)

[1]KAIST, [2]Mila, [3]McGill University, [4]Gauss Labs Inc., [5]Facebook AI Research

Offline Reinforcement Learning (RL)

- Goal: Compute a **policy** that performs better than the data-collecting policy **without further environment interaction.**



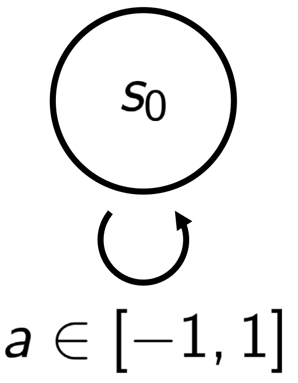
Existing Offline RL Algorithms (1/2)

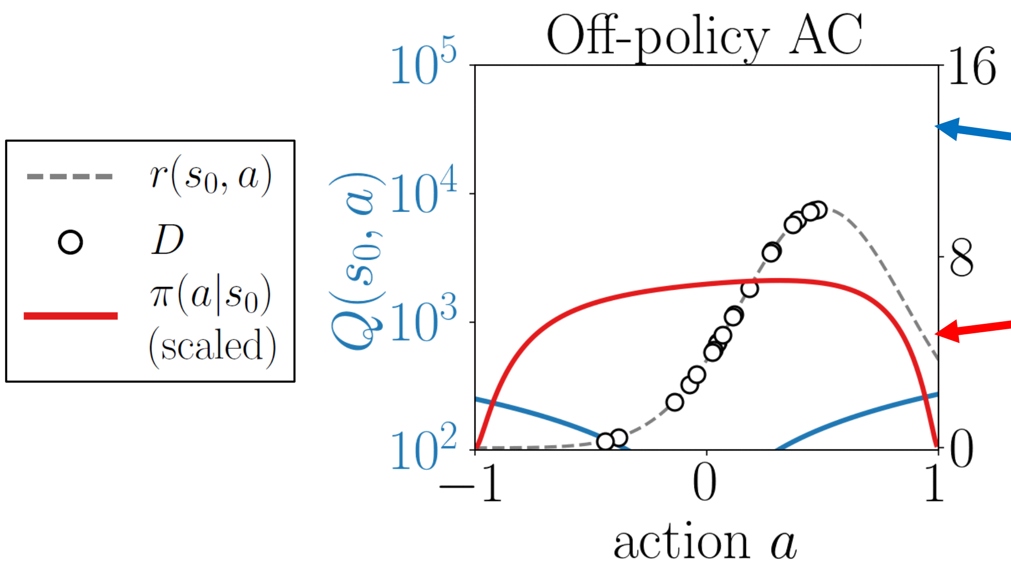
- Off-policy actor-critic

$$\min_{Q} \mathbb{E}_{(s,a,s')\sim d^D, a'\sim\pi(s')} \left[\left(Q(s,a) - (r(s,a) + \gamma \bar{Q}(s', a')) \right)^2 \right]$$

$$\max_{\pi} \mathbb{E}_{s\sim d^D, a\sim\pi(s)} \left[Q(s,a) \right]$$

  - Overestimation of Q due to bootstrapping with out-of-distribution (OOD) action a'.





Exploding value due to overestimation

Policy wrongly converges.

Existing Offline RL Algorithms (2/2)

- Off-policy actor-critic **+ conservatism**

$$\min_{Q} \mathbb{E}_{(s,a,s')\sim d^D, a'\sim\pi(s')}\left[\left(Q(s,a) - (r(s,a) + \gamma \bar{Q}(s', a'))\right)^2\right] + \mathcal{R}_1(Q,\pi)$$

$$\max_{\pi} \mathbb{E}_{s\sim d^D, a\sim\pi(s)}\left[Q(s,a)\right] - \mathcal{R}_2(Q,\pi)$$

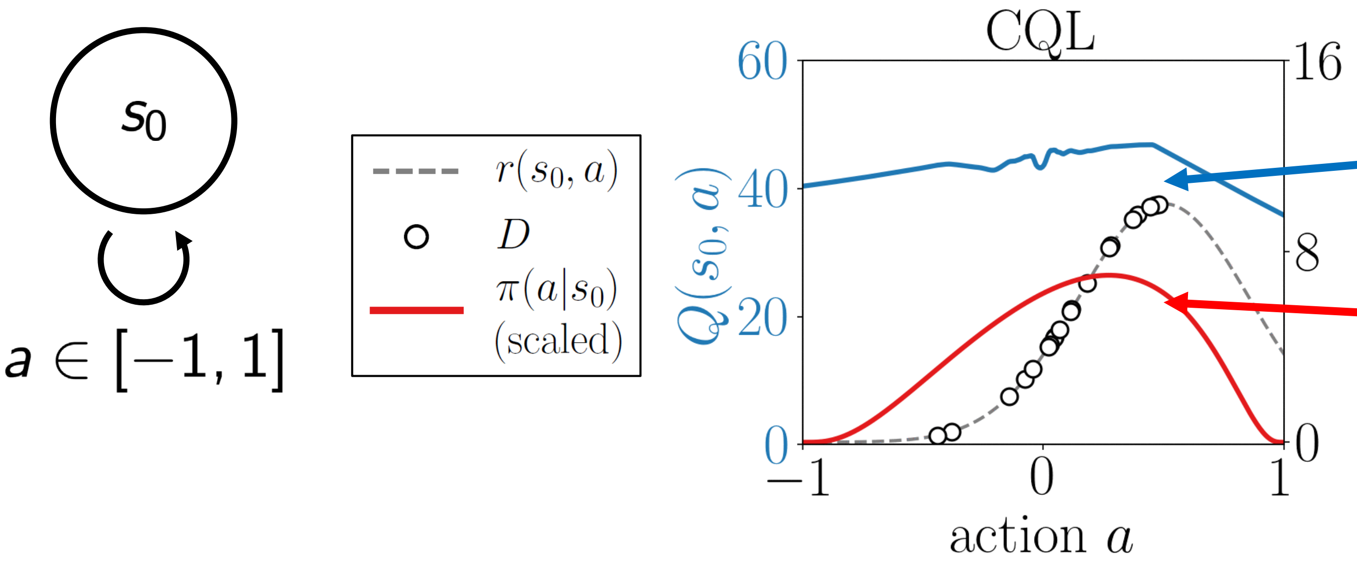
  - **The regularization terms** are to
    - underestimate Q
    - prevent deviating too much from data-collecting policy.

Existing Offline RL Algorithms (2/2)

- Conservative Q-Learning (CQL) [Kumar et al. 2020]

$$\min_{Q} \max_{\mu} \mathbb{E}_{(s,a,s')\sim\mathcal{D}} \mathbb{E}_{a'\sim\pi(\cdot|s')}[(r(s,a) + \gamma \bar{Q}(s',a') - Q(s,a))^2]$$
$$+ \alpha(\mathbb{E}_{s\sim\mathcal{D},a\sim\mu(\cdot|s)} Q(s,a) - \mathbb{E}_{(s,a)\sim\mathcal{D}} Q(s,a))$$

decreases overestimated Q value *increases Q value for in-distribution actions*



Q value of out-of-distribution actions are lowered.

Policy correctly converges.

Our Contribution

- Existing offline RL algorithms
  - Without proper hyperparameters, overestimation still can occur due to **bootstrapping with OOD action values.**

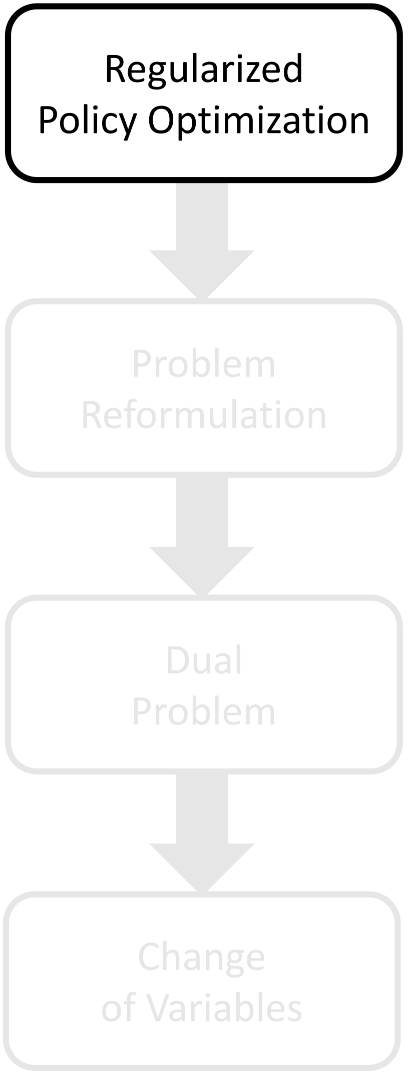
- OptiDICE (Offline Policy ***Opti***mization via Stationary ***DI***stribution ***C***orrection ***E***stimation)
  - Directly optimize **stationary distribution correction** $\boldsymbol{w(s,a) := \frac{d^{\pi}(s,a)}{d^{D}(s,a)}}$.
  - ***No*** alternation between policy evaluation and policy improvement.
  - Free from error due to OOD actions since a’ is not used.

OptiDICE: Objective Function (1/4)

1. Policy optimization with f-divergence regularization:

$$\max_{\pi} \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] - \alpha \underbrace{\mathbb{E}_{(s,a)\sim d^D}\left[f\left(\frac{d^\pi(s,a)}{d^D(s,a)}\right)\right]}_{:= D_f(d^\pi(s,a)||d^D(s,a))\ (\text{f is convex.})}$$

- Encourage visiting state-action pairs in data distribution.



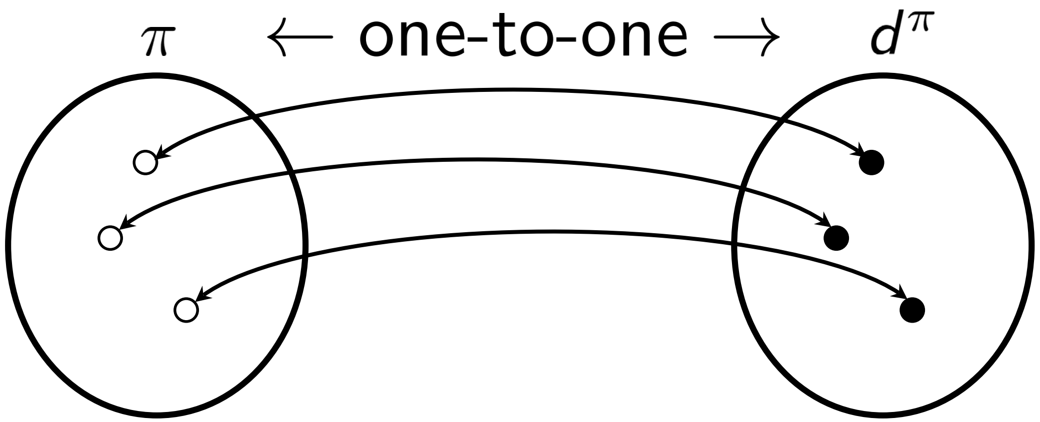
OptiDICE: Objective Function (2/4)

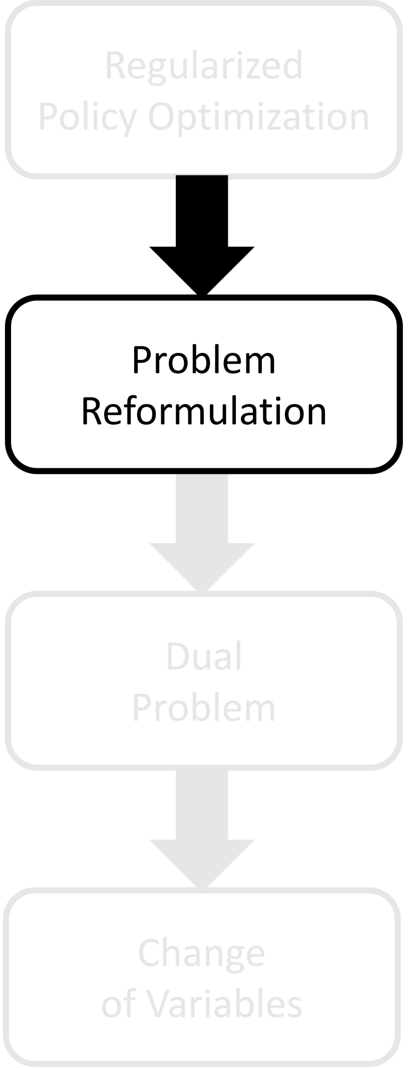
2. Reformulation for optimizing over stationary distributions:

$$\max_{\pi} \mathbb{E}_{(s,a)\sim d^{\pi}}[r(s,a)] - \alpha \mathbb{E}_{(s,a)\sim d^{D}}\left[f\left(\frac{d^{\pi}(s,a)}{d^{D}(s,a)}\right)\right]$$

$$\text{s.t.} (1-\gamma)p_0(s') + \gamma \sum_{s,a} d(s,a) T(s'|s,a) = \sum_{a'} d(s',a') \quad \forall s'$$

Bellman flow constraint





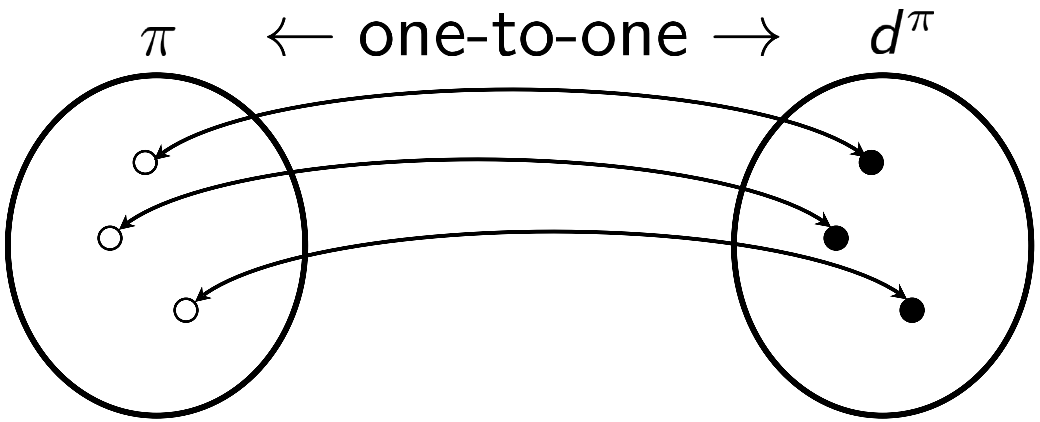
OptiDICE: Objective Function (2/4)

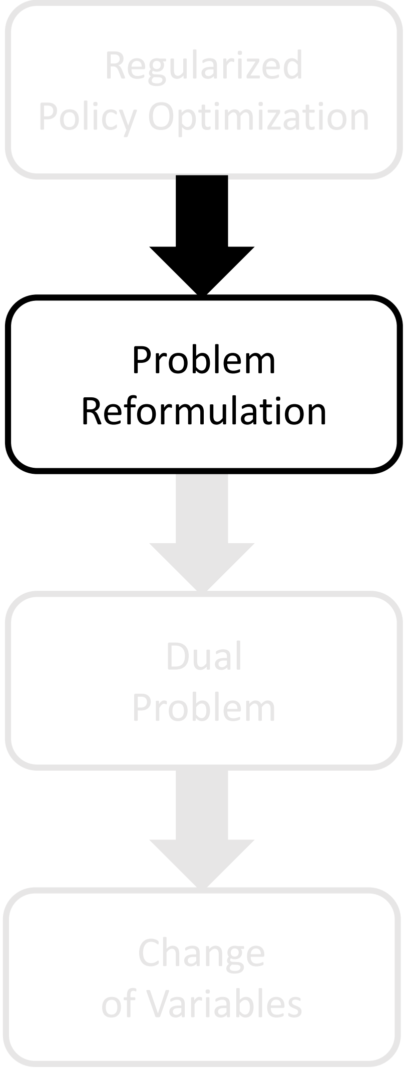
2. Reformulation for optimizing over stationary distributions:

$$\max_{d \geq 0} \mathbb{E}_{(s,a)\sim d}[r(s,a)] - \alpha \mathbb{E}_{(s,a)\sim d^D}\left[f\left(\frac{d(s,a)}{d^D(s,a)} \right) \right]$$

$$\text{s.t.} (1-\gamma) p_0(s') + \gamma \sum_{s,a} d(s,a) T(s'|s,a) = \sum_{a'} d(s',a') \quad \forall s'$$

Bellman flow constraint



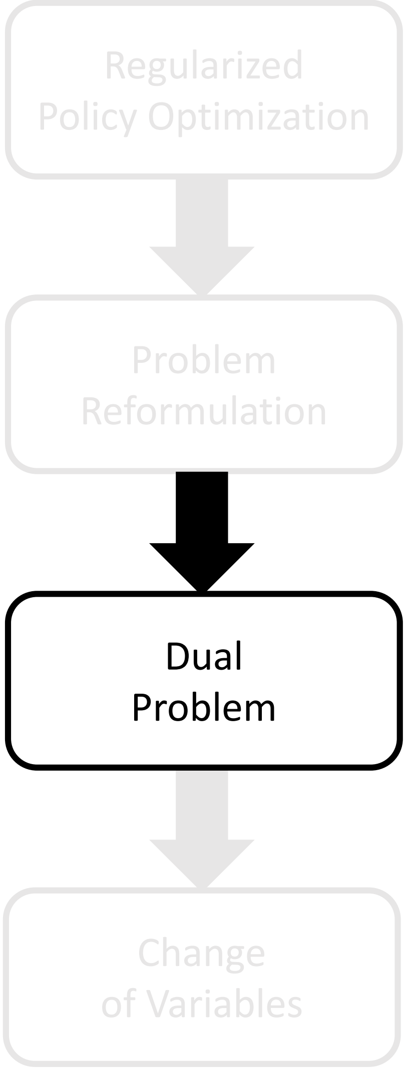


OptiDICE: Objective Function (3/4)

3. Use Lagrangian of the constrained optimization problem:

$$\min_{\nu} \max_{d \geq 0} \mathbb{E}_{(s,a)\sim d}[r(s,a)] - \alpha \mathbb{E}_{(s,a)\sim d^D}\left[f\left(\frac{d(s,a)}{d^D(s,a)}\right)\right]$$

$$+ \underbrace{\sum_{s'} \nu(s')\Big((1-\gamma)p_0(s') + \gamma \sum_{s,a} d(s,a) T(s'|s,a) - \sum_{a'} d(s',a')\Big)}_{= (1-\gamma)\mathbb{E}_{s\sim p_0}[\nu(s)] + \mathbb{E}_{(s,a)\sim d}[\gamma \mathbb{E}_{s'\sim T(s,a)}[\nu(s')]] - \mathbb{E}_{(s,a)\sim d}[\nu(s)]}$$



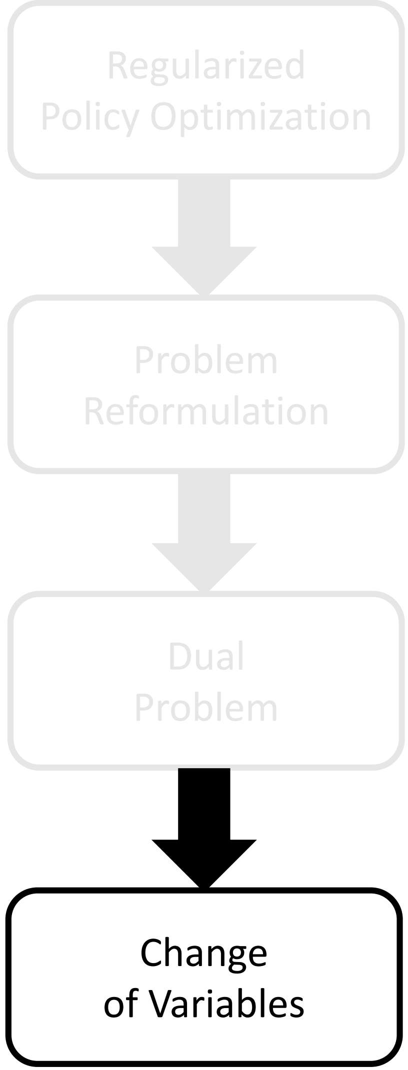
OptiDICE: Objective Function (4/4)

4. Reformulation and change-of-variables:

$$\min_{\nu} \max_{d \geq 0} \mathbb{E}_{(s,a)\sim d^D}\left[\frac{d(s,a)}{d^D(s,a)}\left(r(s,a) + \gamma \mathbb{E}_{s'\sim T(s,a)}[\nu(s')] - \nu(s)\right) - \alpha f\left(\frac{d(s,a)}{d^D(s,a)}\right)\right] + (1-\gamma)\mathbb{E}_{s\sim p_0}[\nu(s)]$$

$$= \min_{\nu} \max_{w \geq 0} \mathbb{E}_{(s,a)\sim d^D}\left[w(s,a)\left(r(s,a) + \gamma \mathbb{E}_{s'\sim T(s,a)}[\nu(s')] - \nu(s)\right) - \alpha f\left(w(s,a)\right)\right] + (1-\gamma)\mathbb{E}_{s\sim p_0}[\nu(s)]$$

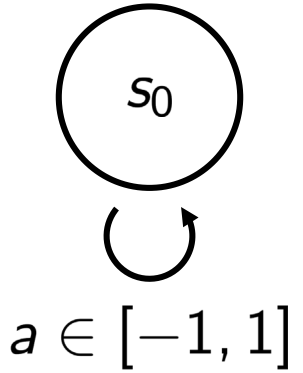
- Seek ***optimal*** *stationary distribution correction* $w^*(s,a) = \frac{d^{\pi^*}(s,a)}{d^D(s,a)}$.
- ***No*** *OOD actions* a', i.e., free from the overestimation.

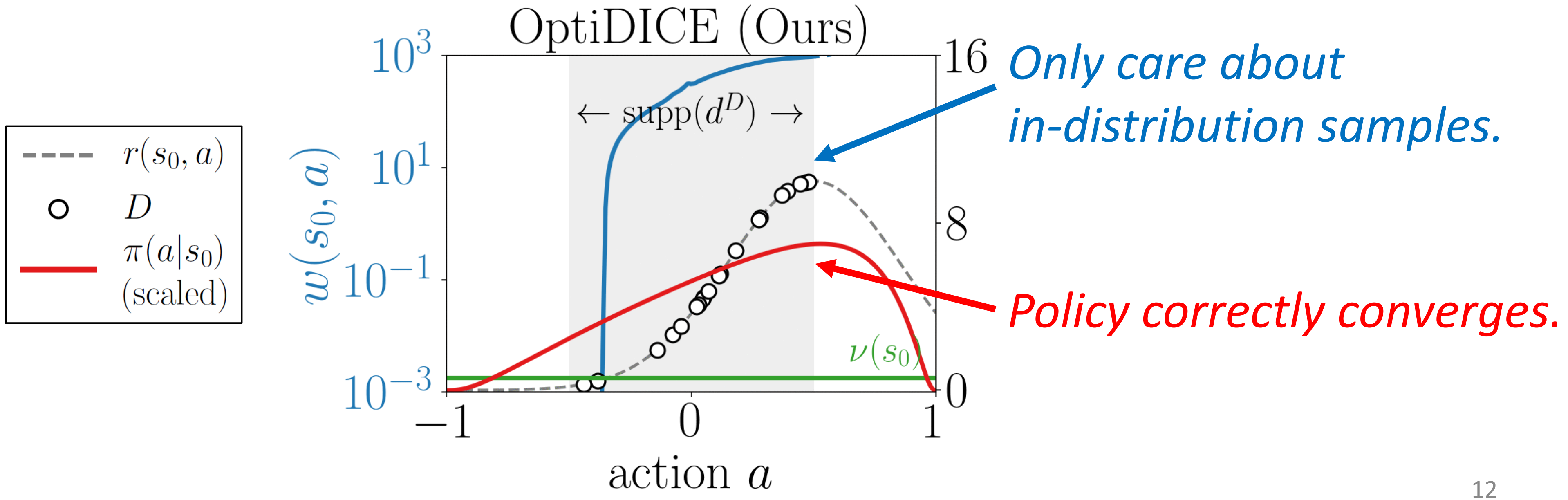


Toy Example

- OptiDICE

$$\min_{\nu} \mathbb{E}_{(s,a)\sim d^D}\left[w^*(s,a;\nu)\left(R(s,a)+\gamma\mathbb{E}_{s'\sim T(s,a)}[\nu(s')]-\nu(s)\right)-\alpha f\left(w^*(s,a;\nu)\right)\right]$$
$$+(1-\gamma)\mathbb{E}_{s\sim p_0}[\nu(s)]$$

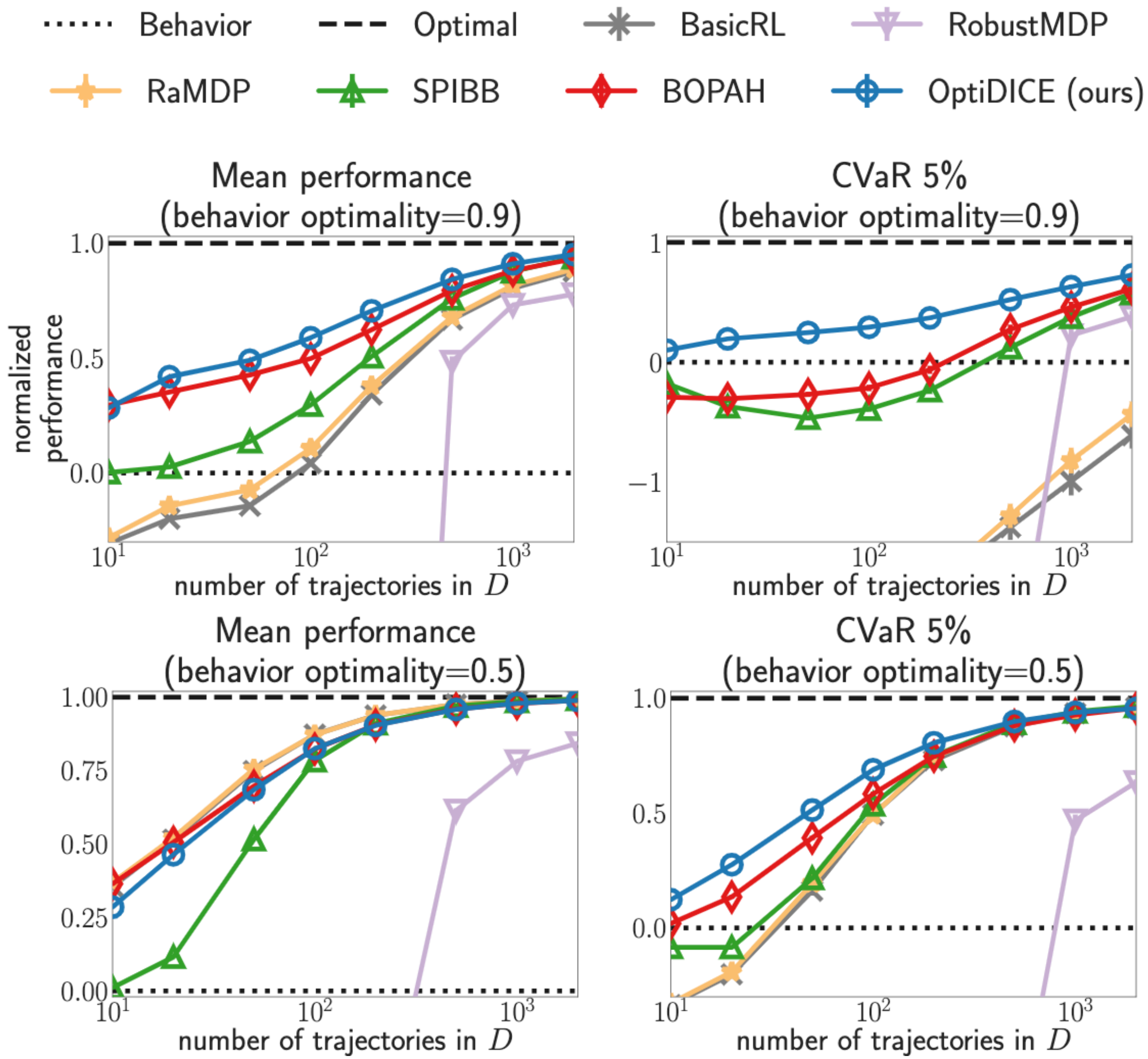




Experiment: Random MDPs

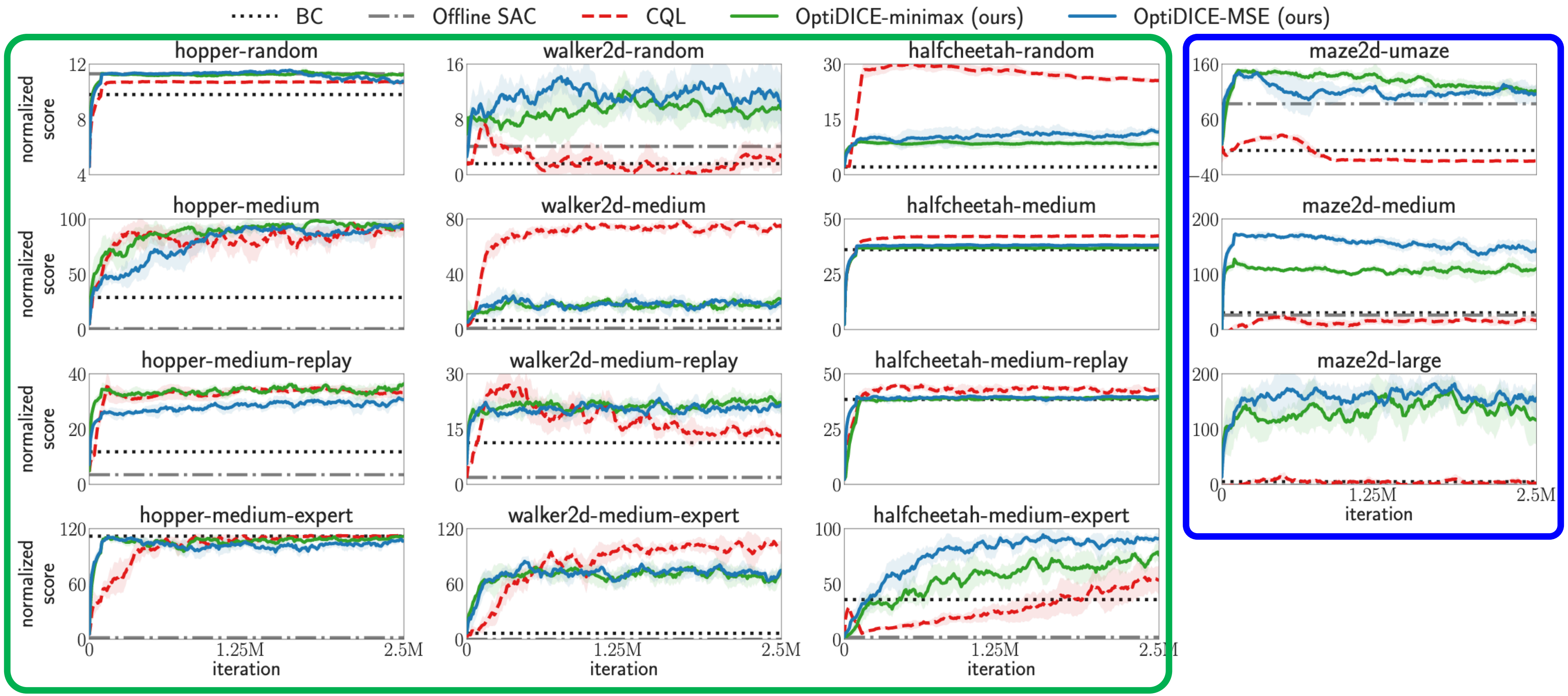
- Performance measure
  - Mean performance
  - Conditional Value at Risk (CVaR)
    - Worst case analysis

- OptiDICE
  - performs **on par with baselines on its mean.**
  - **performs the best in CVaR**.

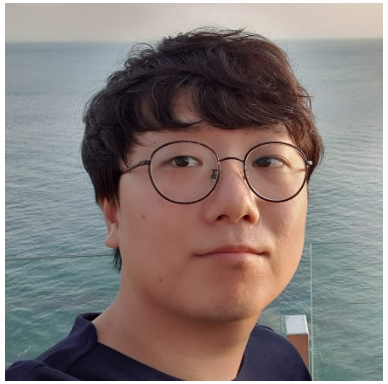


Experiment: D4RL Dataset

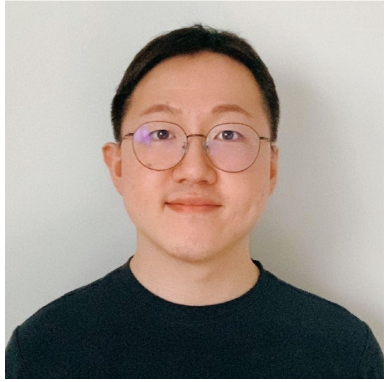
- OptiDICE performs **the best in Maze2D.**
- OptiDICE performs **on par with CQL in MuJoCo.**



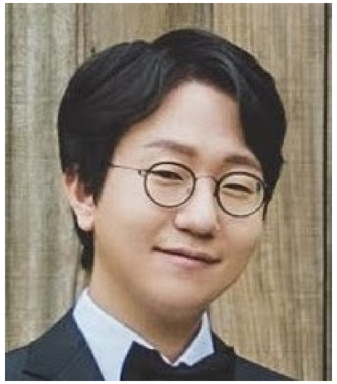
Thanks for Listening!



Jongmin
Lee*



Wonseok
Jeon*



Byung-Jun
Lee

Joelle
Pineau

Kee-Eung
Kim