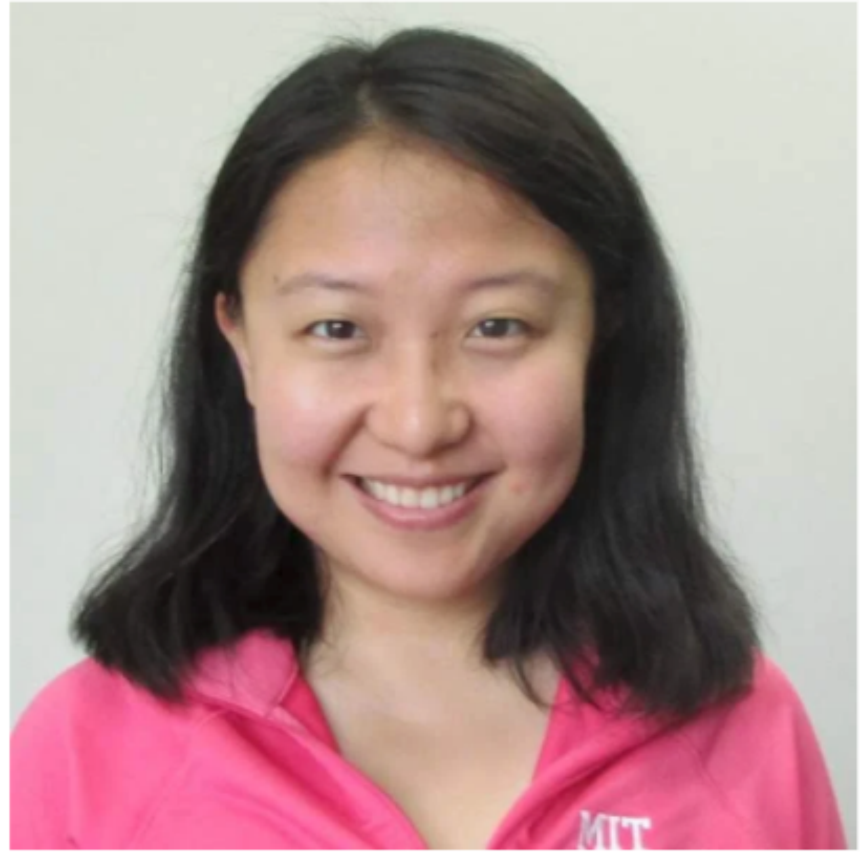
Offline Policy Selection under Uncertainty

Sherry Yang* Bo Dai* Ofir Nachum* George Tucker Dale Schuurmans





Paper: https://arxiv.org/abs/2012.06919
Code: https://github.com/google-research/dice_rl

Offline Policy Selection

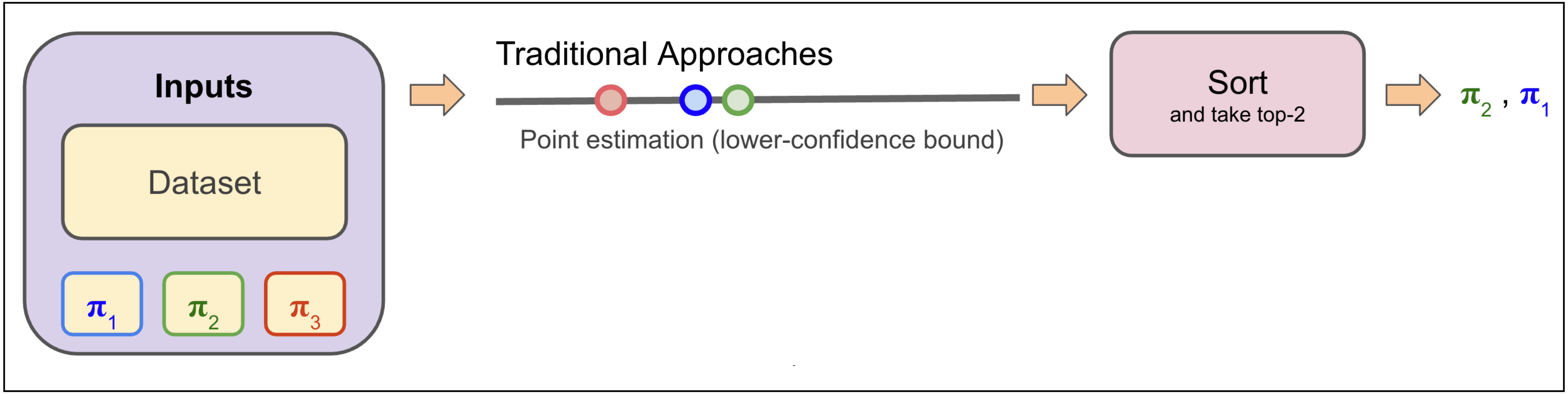
Offline policy selection:

- Compute a ranking O $\in$ Perm([1, N]) over $\{\pi_i\}_{i=1}^{N}$ given a fixed dataset D according to some utility function u: $\mathcal{O} \leftarrow \mathrm{ArgSortDescending}(\{u(\pi_i)\}_{i=1}^{N})$

Offline Policy Selection

Offline policy selection:

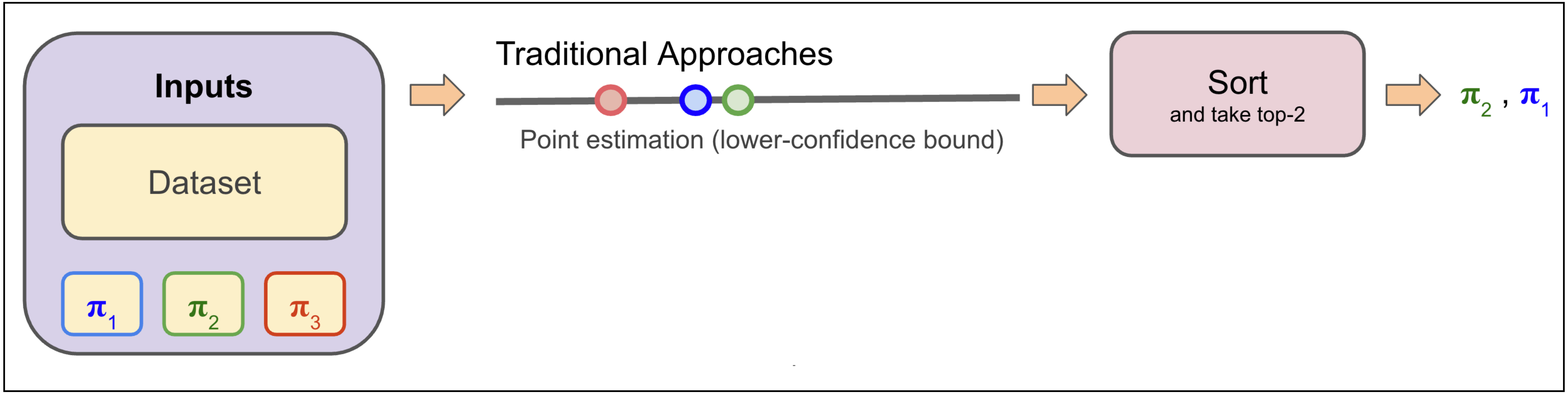
- Compute a ranking $O \in \text{Perm}([1, N])$ over $\{\pi_i\}_{i=1}^{N}$ given a fixed dataset D according to some utility function u: $\mathcal{O} \leftarrow \text{ArgSortDescending}(\{u(\pi_i)\}_{i=1}^{N})$



Offline Policy Selection

Offline policy selection:

- Compute a ranking O ∈ Perm([1, N]) over $\{\pi_i\}_{i=1}^{N}$ given a fixed dataset D according to some utility function u: $\mathcal{O} \leftarrow \mathrm{ArgSortDescending}(\{u(\pi_i)\}_{i=1}^{N})$

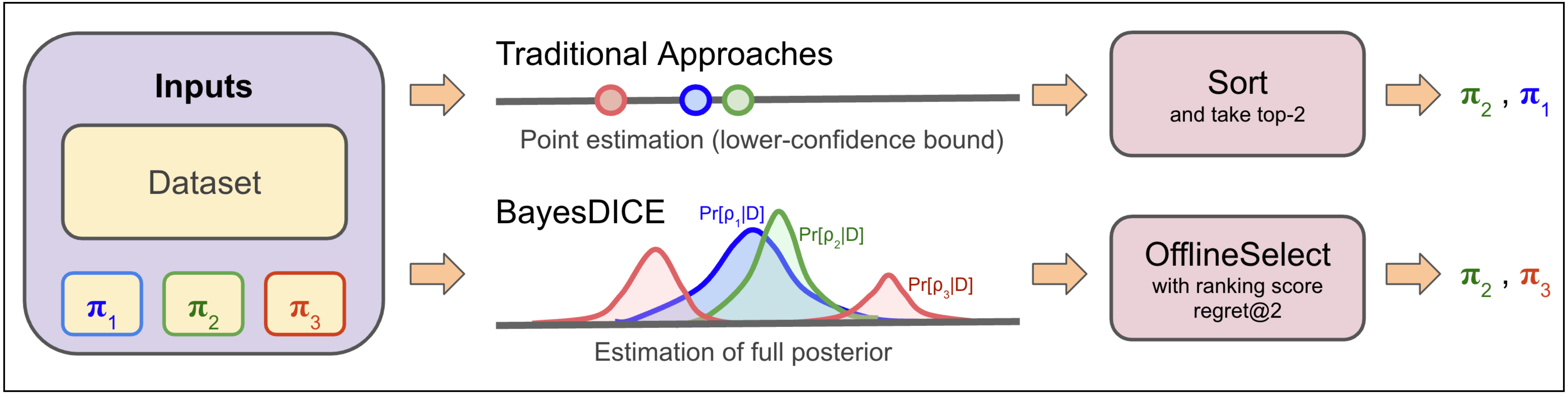


- Practical ranking criteria: top-k precision, top-k accuracy, top-k regret, top-k correlation, CVaR, …

Offline Policy Selection

Offline policy selection:

- Compute a ranking O ∈ Perm([1, N]) over $\{\pi_i\}_{i=1}^{N}$ given a fixed dataset D according to some utility function u: $\mathcal{O} \leftarrow \mathrm{ArgSortDescending}(\{u(\pi_i)\}_{i=1}^{N})$



- Practical ranking criteria: top-k precision, top-k accuracy, top-k regret, top-k correlation, CVaR, …

BayesDICE

Recall off-policy evaluation:

$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[R(s,a)]$ where $\mathcal{P}_*^\pi d^\pi(s,a) := \pi(a|s) \sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a}) d^\pi(\tilde{s},\tilde{a})$

BayesDICE

Recall off-policy evaluation:

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[R(s,a)] \text{ where } \mathcal{P}_*^\pi d^\pi(s,a) := \pi(a|s)\sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a})d^\pi(\tilde{s},\tilde{a})$$

DICE point estimator:

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\zeta^*(s,a)\cdot R(s,a)] \text{ where } \zeta^*(s,a) := \frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}$$

[1] Nachum, et al. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections.
[2] Yang, et al. Off-policy evaluation via the regularized lagrangian.

BayesDICE

Recall off-policy evaluation:

$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[R(s,a)]$ where $\mathcal{P}_*^\pi d^\pi(s,a) := \pi(a|s)\sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a})d^\pi(\tilde{s},\tilde{a})$

DICE point estimator:

$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\zeta^*(s,a)\cdot R(s,a)]$ where $\zeta^*(s,a) := \frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}$

BayesDICE learns $q(\zeta^\pi|\mathcal{D}) \propto p(\mathcal{D}|\zeta^\pi)\,p(\zeta^\pi)$:

- By optimizing $\min_{q\in\mathcal{P}} -\mathbb{E}_{q(\zeta^\pi)}[\log p(\mathcal{D}|\zeta^\pi)] + KL(q\|p)$

[1] Nachum, et al. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections.
[2] Yang, et al. Off-policy evaluation via the regularized lagrangian.

BayesDICE

Recall off-policy evaluation:

$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[R(s,a)]$ where $\mathcal{P}_*^\pi d^\pi(s,a) := \pi(a|s)\sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a})d^\pi(\tilde{s},\tilde{a})$

DICE point estimator:

$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\mathcal{D}}[\zeta^*(s,a)\cdot R(s,a)]$ where $\zeta^*(s,a) := \frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}$

constraints

BayesDICE learns $q(\zeta^\pi|\mathcal{D}) \propto p(\mathcal{D}|\zeta^\pi)\, p(\zeta^\pi)$:

- By optimizing $\min_{q\in\mathcal{P}} -\mathbb{E}_{q(\zeta^\pi)}[\log p(\mathcal{D}|\zeta^\pi)] + KL(q\|p)$

[1] Nachum, et al. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections.
[2] Yang, et al. Off-policy evaluation via the regularized lagrangian.

BayesDICE

Recall off-policy evaluation:

$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[R(s,a)]$ where $\mathcal{P}_*^\pi d^\pi(s,a) := \pi(a|s)\sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a})d^\pi(\tilde{s},\tilde{a})$

DICE point estimator:

$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\mathcal{D}}[\zeta^*(s,a)\cdot R(s,a)]$ where $\zeta^*(s,a) := \frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}$

constraints

BayesDICE learns $q(\zeta^\pi|\mathcal{D}) \propto p(\mathcal{D}|\zeta^\pi)\,p(\zeta^\pi)$:

- By optimizing $\min_{q\in\mathcal{P}} -\mathbb{E}_{q(\zeta^\pi)}[\log p(\mathcal{D}|\zeta^\pi)] + KL(q\|p)$
- Posterior regularization:

$$\min_q \xi + KL(q\|p) \quad \text{s.t.} \quad q \in \mathcal{P}\cap\{\xi = -\mathbb{E}_{q(\zeta^\pi)}[\log p(\mathcal{D}|\zeta^\pi)]\}$$

[1] Nachum, et al. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections.

[2] Yang, et al. Off-policy evaluation via the regularized lagrangian.

BayesDICE

Constraint embeddings:

- Density constraints: $\mathcal{P}_*^\pi d^\pi(s,a) := \pi(a|s) \sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a}) d^\pi(\tilde{s},\tilde{a})$

  Equivalently: $\Delta_d(s,a) := (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a) - d(s,a) = 0$

[1] Farias and Roy. The linear programming approach to approximate dynamic programming.
[2] Small, et al. A Hilbert space embedding for distributions.
[3] Nachum, et al. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections.

BayesDICE

Constraint embeddings:

- Density constraints: $\mathcal{P}_*^\pi d^\pi(s,a) := \pi(a|s)\sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a})d^\pi(\tilde{s},\tilde{a})$
  Equivalently: $\Delta_d(s,a) := (1-\gamma)\mu_0(s)\pi(a|s) + \gamma\cdot\mathcal{P}_*^\pi d(s,a) - d(s,a) = 0$
- Introduce function space embedding ϕ :

$$\langle\phi,\Delta_d\rangle := \mathbb{E}_{(1-\gamma)\mu_0(s)\pi(a|s)+\gamma\cdot\mathcal{P}_*^\pi d(s,a)}[\phi(s,a)] - \mathbb{E}_{d(s,a)}[\phi(s,a)]$$

[1] Farias and Roy. The linear programming approach to approximate dynamic programming.
[2] Small, et al. A Hilbert space embedding for distributions.
[3] Nachum, et al. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections.

BayesDICE

Constraint embeddings:

- Density constraints: $\mathcal{P}_*^\pi d^\pi(s,a) := \pi(a|s)\sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a})d^\pi(\tilde{s},\tilde{a})$
  Equivalently: $\Delta_d(s,a) := (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a) - d(s,a) = 0$
- Introduce function space embedding ϕ :
  $$\langle \phi, \Delta_d \rangle := \mathbb{E}_{(1-\gamma)\mu_0(s)\pi(a|s)+\gamma\cdot\mathcal{P}_*^\pi d(s,a)}\left[\phi(s,a)\right] - \mathbb{E}_{d(s,a)}\left[\phi(s,a)\right]$$
- BayesDICE objective:
  $$\min_{q\in\mathcal{P}} -\mathbb{E}_{q(\zeta^\pi)}\left[\log p\left(\mathcal{D}|\zeta^\pi\right)\right] + KL\left(q\|p\right)$$
  $$\min_{q} \frac{\lambda}{\epsilon}\mathbb{E}_q\left[\ell(\zeta,\mathcal{D})\right] + KL\left(q\|p\right) \text{ where } \ell\left(\zeta,\mathcal{D}\right) = \langle\phi,\Delta_d\rangle^\top\langle\phi,\Delta_d\rangle$$
  $$= \max_{\beta\in\mathcal{H}_\phi} \beta^\top\langle\phi,\Delta_d\rangle - \beta^\top\beta$$

[1] Farias and Roy. The linear programming approach to approximate dynamic programming.
[2] Small, et al. A Hilbert space embedding for distributions.
[3] Nachum, et al. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections.

Experiments: CI Estimation

Experiments: Policy Selection

Experiments: Policy Selection

Thank you. Checkout

Paper: https://arxiv.org/pdf/2012.06919.pdf

Code: https://github.com/google-research/google-research/tree/master/rl_repr