DualDICE

Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections

Ofir Nachum,* Yinlam Chow,* Bo Dai, Lihong Li

Google Research

*Equal contribution

Reinforcement Learning

Reinforcement Learning

- A policy acts on an environment.

Reinforcement Learning

- A policy acts on an environment.

Initial state distribution β

s_0

Reinforcement Learning

- A policy acts on an environment.

$\pi(-|s_0)$

$\downarrow$

a_0

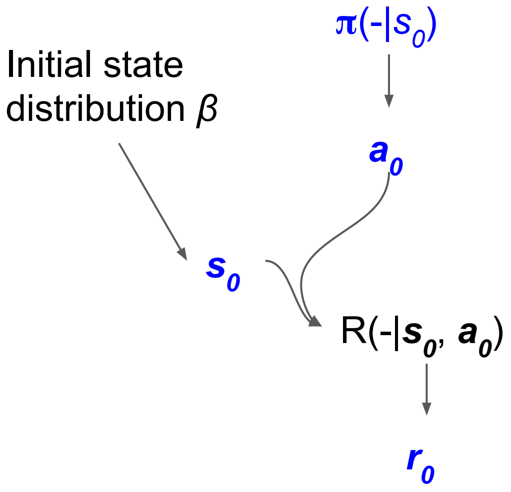
Initial state distribution β

$\searrow$

s_0

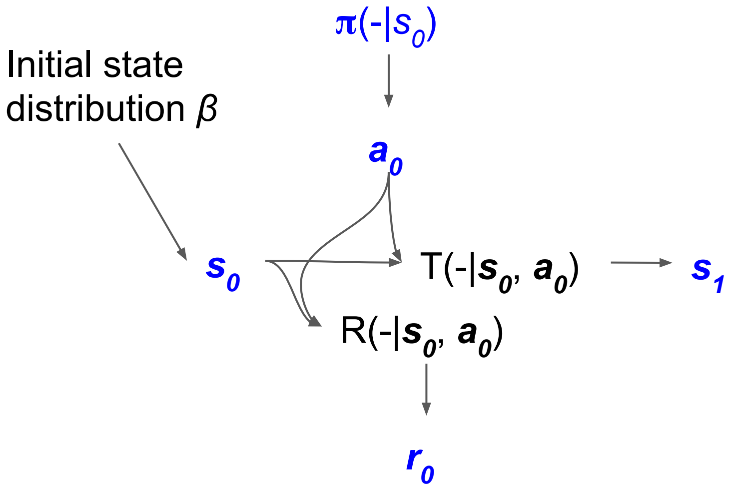
Reinforcement Learning

- A policy acts on an environment.



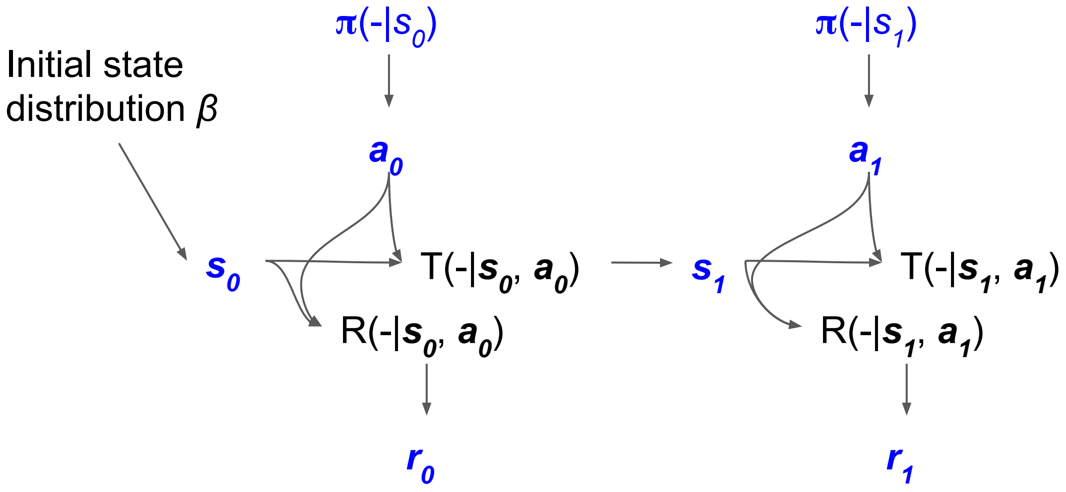
Reinforcement Learning

- A policy acts on an environment.



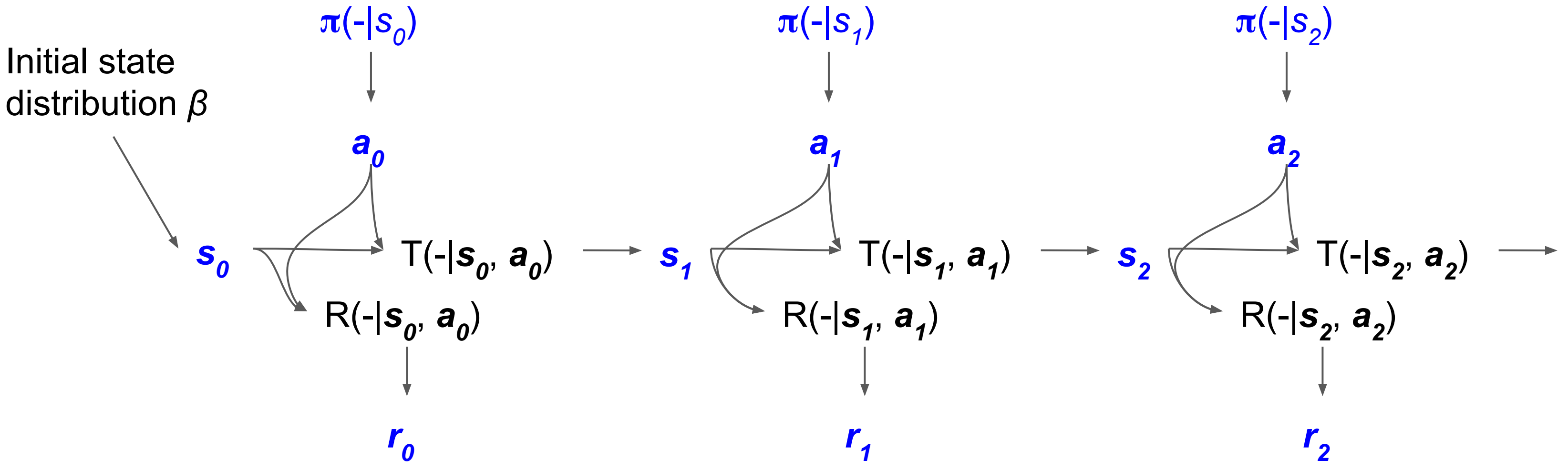
Reinforcement Learning

- A policy acts on an environment.



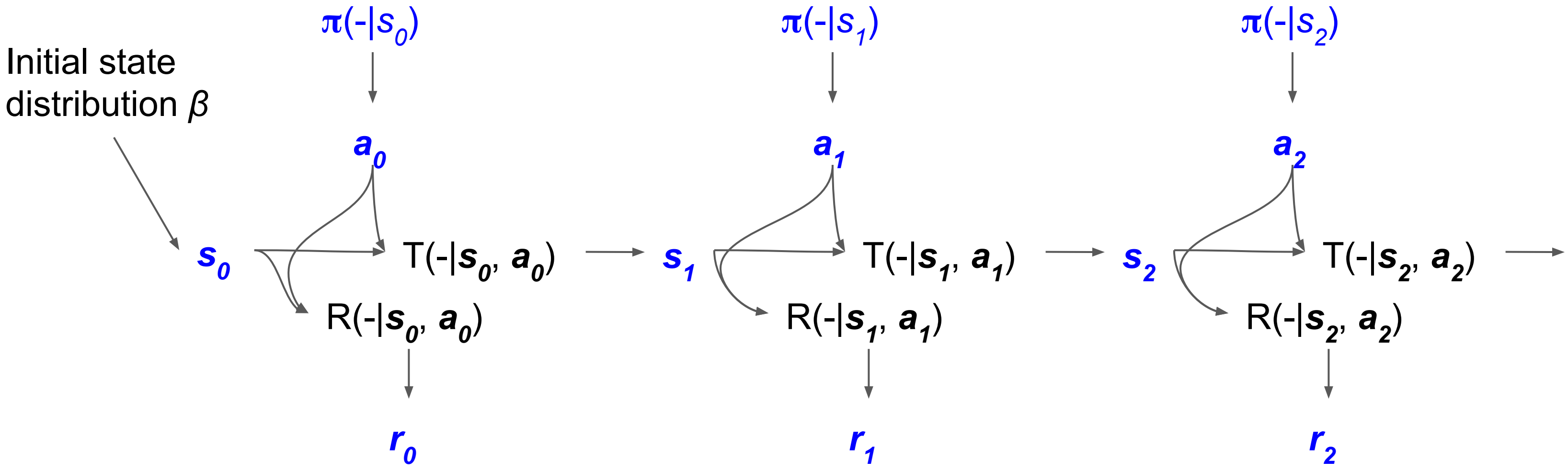
Reinforcement Learning

- A policy acts on an environment.



Reinforcement Learning

- A policy acts on an environment.



- Question: What is the value (average reward) of the policy?

Off-policy Policy Estimation

Off-policy Policy Estimation

- Want to estimate average discounted per-step reward of policy,

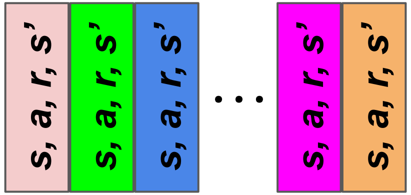
$$\rho(\pi) := (1 - \gamma) \cdot \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s, a) \mid s_0 \sim \beta_0, a_t \sim \pi(s_t), s_{t+1} \sim T(s_t, a_t)\right]$$

Off-policy Policy Estimation

- Want to estimate average discounted per-step reward of policy,

$$\rho(\pi) := (1-\gamma) \cdot \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s,a) \mid s_0 \sim \beta_0, a_t \sim \pi(s_t), s_{t+1} \sim T(s_t, a_t)\right]$$

- Only have access to finite experience dataset $\mathcal{D} := \{(s^{(i)}, a^{(i)}, r^{(i)}, s^{(i)\prime})\}_{i=1}^{N}$



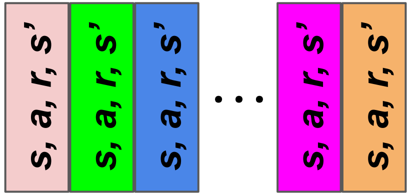
where transitions are from some unknown distribution $(s,a) \sim d^{\mathcal{D}}, s' \sim T(s,a)$

Off-policy Policy Estimation

- Want to estimate average discounted per-step reward of policy,

$$\rho(\pi) := (1-\gamma) \cdot \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s,a) \mid s_0 \sim \beta_0, a_t \sim \pi(s_t), s_{t+1} \sim T(s_t, a_t)\right]$$

- Only have access to finite experience dataset $\mathcal{D} := \{(s^{(i)}, a^{(i)}, r^{(i)}, s^{(i)\prime})\}_{i=1}^{N}$



where transitions are from some unknown distribution $(s,a) \sim d^{\mathcal{D}}, s' \sim T(s,a)$

- Don't even know the behavior policy!

Reduction of OPE to Density Ratio Estimation

Reduction of OPE to Density Ratio Estimation

- Can write $\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)]$ where d^π is discounted on-policy distribution

$$d^\pi(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \cdot \Pr[s_t = s, a_t = a \mid s_0 \sim \beta_0, \pi]$$

Reduction of OPE to Density Ratio Estimation

- Can write $\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)]$ where d^π is discounted on-policy distribution

$$d^\pi(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \cdot \Pr[s_t = s, a_t = a \mid s_0 \sim \beta_0, \pi]$$

- Using importance weighting trick, we have,

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] = \mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[\frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)} \cdot r(s,a)\right]$$

Reduction of OPE to Density Ratio Estimation

- Can write $\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)]$ where d^π is discounted on-policy distribution

$$d^\pi(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \cdot \Pr[s_t = s, a_t = a \mid s_0 \sim \beta_0, \pi]$$

- Using importance weighting trick, we have,

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] = \mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[\frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)} \cdot r(s,a)\right]$$

- Given finite dataset, this corresponds to weighted average,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{d^\pi(s^{(i)},a^{(i)})}{d^\mathcal{D}(s^{(i)},a^{(i)})} \cdot r^{(i)}$$

Reduction of OPE to Density Ratio Estimation

- Can write $\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)]$ where d^π is discounted on-policy distribution

$$d^\pi(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \cdot \Pr[s_t = s, a_t = a \mid s_0 \sim \beta_0, \pi]$$

- Using importance weighting trick, we have,

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] = \mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[\frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)} \cdot r(s,a)\right]$$

- Given finite dataset, this corresponds to weighted average,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{d^\pi(s^{(i)},a^{(i)})}{d^\mathcal{D}(s^{(i)},a^{(i)})} \cdot r^{(i)}$$

- Problem reduces to estimating weights (density ratios) $w_{\pi/\mathcal{D}}(s,a) := \dfrac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}$

Reduction of OPE to Density Ratio Estimation

- Can write $\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)]$ where d^π is discounted on-policy distribution

$$d^\pi(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \cdot \Pr[s_t = s, a_t = a \mid s_0 \sim \beta_0, \pi]$$

- Using importance weighting trick, we have,

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] = \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\boxed{\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}} \cdot r(s,a)\right]$$

- Given finite dataset, this corresponds to weighted average,

$$\frac{1}{N}\sum_{i=1}^{N}\boxed{\frac{d^\pi(s^{(i)},a^{(i)})}{d^{\mathcal{D}}(s^{(i)},a^{(i)})}} \cdot r^{(i)}$$

- Problem reduces to estimating weights (density ratios) $\boxed{w_{\pi/\mathcal{D}}(s,a) := \frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}}$

Reduction of OPE to Density Ratio Estimation

- Can write $\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)]$ where d^π is discounted on-policy distribution

$$d^\pi(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \cdot \Pr[s_t = s, a_t = a \mid s_0 \sim \beta_0, \pi]$$

- Using importance weighting trick, we have,

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] = \mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[\boxed{\frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}} \cdot r(s,a)\right]$$

- Given finite dataset, this corresponds to weighted average,

$$\frac{1}{N}\sum_{i=1}^{N}\boxed{\frac{d^\pi(s^{(i)},a^{(i)})}{d^\mathcal{D}(s^{(i)},a^{(i)})}} \cdot r^{(i)}$$

- Problem reduces to estimating weights (density ratios) $\boxed{w_{\pi/\mathcal{D}}(s,a) := \frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}}$

- Difficult because we don’t have access to environment and we don’t have explicit knowledge of $d^D(s,a)$, only samples.

The DualDICE Objective

- Define zero-reward Bellman operator as $\mathcal{B}_\pi \nu(s,a) := \gamma \mathbb{E}_{s' \sim T(s,a), a' \sim \pi(s')}[\nu(s', a')]$

The DualDICE Objective

- Define zero-reward Bellman operator as $\mathcal{B}_\pi \nu(s,a) := \gamma \mathbb{E}_{s' \sim T(s,a), a' \sim \pi(s')}[\nu(s',a')]$

$$\min_{\nu : S \times A \to \mathbb{R}} \frac{1}{2} \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}}[(\nu(s,a) - \mathcal{B}_\pi \nu(s,a))^2] - (1-\gamma) \cdot \mathbb{E}_{\substack{s_0 \sim \beta_0 \\ a_0 \sim \pi(s_0)}}[\nu(s_0,a_0)]$$

The DualDICE Objective

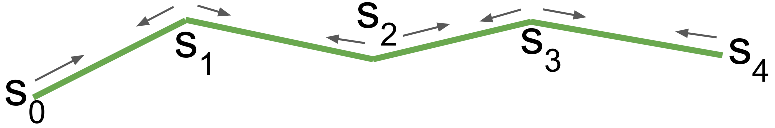
- Define zero-reward Bellman operator as $\mathcal{B}_\pi \nu(s,a) := \gamma \mathbb{E}_{s' \sim T(s,a), a' \sim \pi(s')}[\nu(s', a')]$

$$\min_{\nu: S \times A \to \mathbb{R}} \frac{1}{2} \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} \overbrace{[(\nu(s,a) - \mathcal{B}_\pi \nu(s,a))}^{\text{minimize squared Bellman error}}{}^2] - (1-\gamma) \cdot \mathbb{E}_{\substack{s_0 \sim \beta_0 \\ a_0 \sim \pi(s_0)}} [\nu(s_0, a_0)]$$

The DualDICE Objective

- Define zero-reward Bellman operator as $\mathcal{B}_\pi \nu(s,a) := \gamma \mathbb{E}_{s' \sim T(s,a), a' \sim \pi(s')}[\nu(s', a')]$

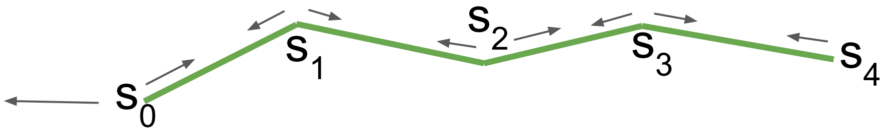
$$\min_{\nu : S \times A \to \mathbb{R}} \frac{1}{2} \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} \overbrace{[(\nu(s,a) - \mathcal{B}_\pi \nu(s,a))^2]}^{\text{minimize squared Bellman error}} - (1-\gamma) \cdot \mathbb{E}_{\substack{s_0 \sim \beta_0 \\ a_0 \sim \pi(s_0)}} [\nu(s_0, a_0)]$$



The DualDICE Objective

- Define zero-reward Bellman operator as $\mathcal{B}_\pi\nu(s,a) := \gamma\mathbb{E}_{s'\sim T(s,a),a'\sim\pi(s')}[\nu(s',a')]$

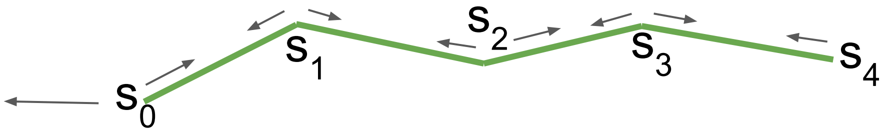
$$\min_{\nu:S\times A\to\mathbb{R}} \frac{1}{2}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\overbrace{[(\nu(s,a)-\mathcal{B}_\pi\nu(s,a))^2]}^{\text{minimize squared Bellman error}} - (1-\gamma)\cdot\mathbb{E}_{\substack{s_0\sim\beta_0\\ a_0\sim\pi(s_0)}}\overbrace{[\nu(s_0,a_0)]}^{\text{maximize initial “nu-values”}}$$



The DualDICE Objective

- Define zero-reward Bellman operator as $\mathcal{B}_\pi\nu(s,a) := \gamma\mathbb{E}_{s'\sim T(s,a),a'\sim\pi(s')}[\nu(s',a')]$

$$\min_{\nu:S\times A\to\mathbb{R}} \frac{1}{2}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\overbrace{[(\nu(s,a)-\mathcal{B}_\pi\nu(s,a))^2]}^{\text{minimize squared Bellman error}} - (1-\gamma)\cdot \mathbb{E}_{\substack{s_0\sim\beta_0\\ a_0\sim\pi(s_0)}} \overbrace{[\nu(s_0,a_0)]}^{\text{maximize initial “nu-values”}}$$

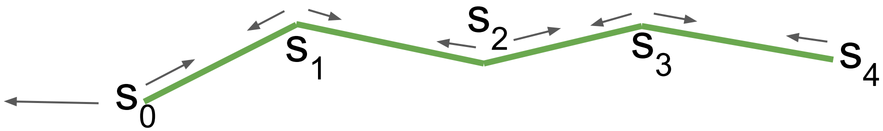


$$\Rightarrow \nu^*(s,a) - \mathcal{B}_\pi\nu^*(s,a) = w_{\pi/\mathcal{D}}(s,a)$$

The DualDICE Objective

- Define zero-reward Bellman operator as $\mathcal{B}_\pi \nu(s,a) := \gamma \mathbb{E}_{s' \sim T(s,a), a' \sim \pi(s')}[\nu(s',a')]$

$$\min_{\nu: S \times A \to \mathbb{R}} \frac{1}{2} \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}} \overbrace{\left[(\nu(s,a) - \mathcal{B}_\pi \nu(s,a))^2 \right]}^{\text{minimize squared Bellman error}} - (1-\gamma) \cdot \mathbb{E}_{\substack{s_0 \sim \beta_0 \\ a_0 \sim \pi(s_0)}} \overbrace{[\nu(s_0,a_0)]}^{\text{maximize initial “nu-values”}}$$

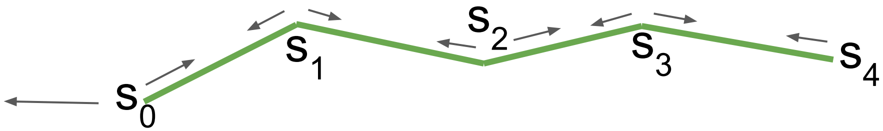


$$\boxed{\Rightarrow \nu^*(s,a) - \mathcal{B}_\pi \nu^*(s,a) = w_{\pi/\mathcal{D}}(s,a)}$$

The DualDICE Objective

- Define zero-reward Bellman operator as $\mathcal{B}_\pi \nu(s,a) := \gamma \mathbb{E}_{s' \sim T(s,a), a' \sim \pi(s')}[\nu(s',a')]$

$$\min_{\nu: S \times A \to \mathbb{R}} \frac{1}{2} \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}} \overbrace{\left[(\nu(s,a) - \mathcal{B}_\pi \nu(s,a))^2\right]}^{\text{minimize squared Bellman error}} - (1-\gamma) \cdot \mathbb{E}_{\substack{s_0 \sim \beta_0 \\ a_0 \sim \pi(s_0)}} \overbrace{[\nu(s_0, a_0)]}^{\text{maximize initial “nu-values”}}$$



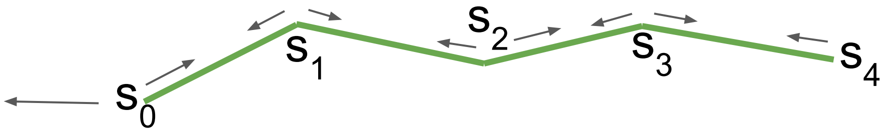
$$\boxed{\Rightarrow \nu^*(s,a) - \mathcal{B}_\pi \nu^*(s,a) = w_{\pi/\mathcal{D}}(s,a)}$$

- **Nice:** Objective is based on expectations from d^D, β, and π, which we have access to.

The DualDICE Objective

- Define zero-reward Bellman operator as $\mathcal{B}_\pi \nu(s,a) := \gamma \mathbb{E}_{s' \sim T(s,a), a' \sim \pi(s')}[\nu(s',a')]$

$$\min_{\nu: S \times A \to \mathbb{R}} \frac{1}{2} \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}} \overbrace{[(\nu(s,a) - \mathcal{B}_\pi \nu(s,a))^2]}^{\text{minimize squared Bellman error}} - (1-\gamma) \cdot \mathbb{E}_{\substack{s_0 \sim \beta_0 \\ a_0 \sim \pi(s_0)}} \overbrace{[\nu(s_0, a_0)]}^{\text{maximize initial “nu-values”}}$$



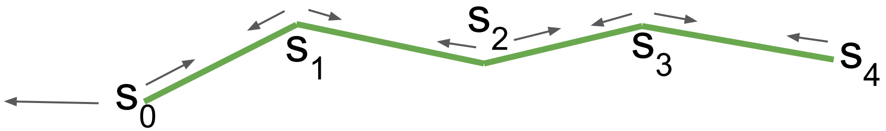
$$\boxed{\Rightarrow \nu^*(s,a) - \mathcal{B}_\pi \nu^*(s,a) = w_{\pi/\mathcal{D}}(s,a)}$$

- **Nice:** Objective is based on expectations from d^D, β, and π, which we have access to.
- **Extension 1:** Can remove appearance of Bellman operator from **both** objective and solution by application of Fenchel conjugate!

The DualDICE Objective

- Define zero-reward Bellman operator as $\mathcal{B}_\pi \nu(s,a) := \gamma \mathbb{E}_{s' \sim T(s,a), a' \sim \pi(s')}[\nu(s',a')]$

$$\min_{\nu: S \times A \to \mathbb{R}} \frac{1}{2} \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} \overbrace{[(\nu(s,a) - \mathcal{B}_\pi \nu(s,a))^2]}^{\text{minimize squared Bellman error}} - (1-\gamma) \cdot \mathbb{E}_{\substack{s_0 \sim \beta_0 \\ a_0 \sim \pi(s_0)}} \overbrace{[\nu(s_0, a_0)]}^{\text{maximize initial “nu-values”}}$$

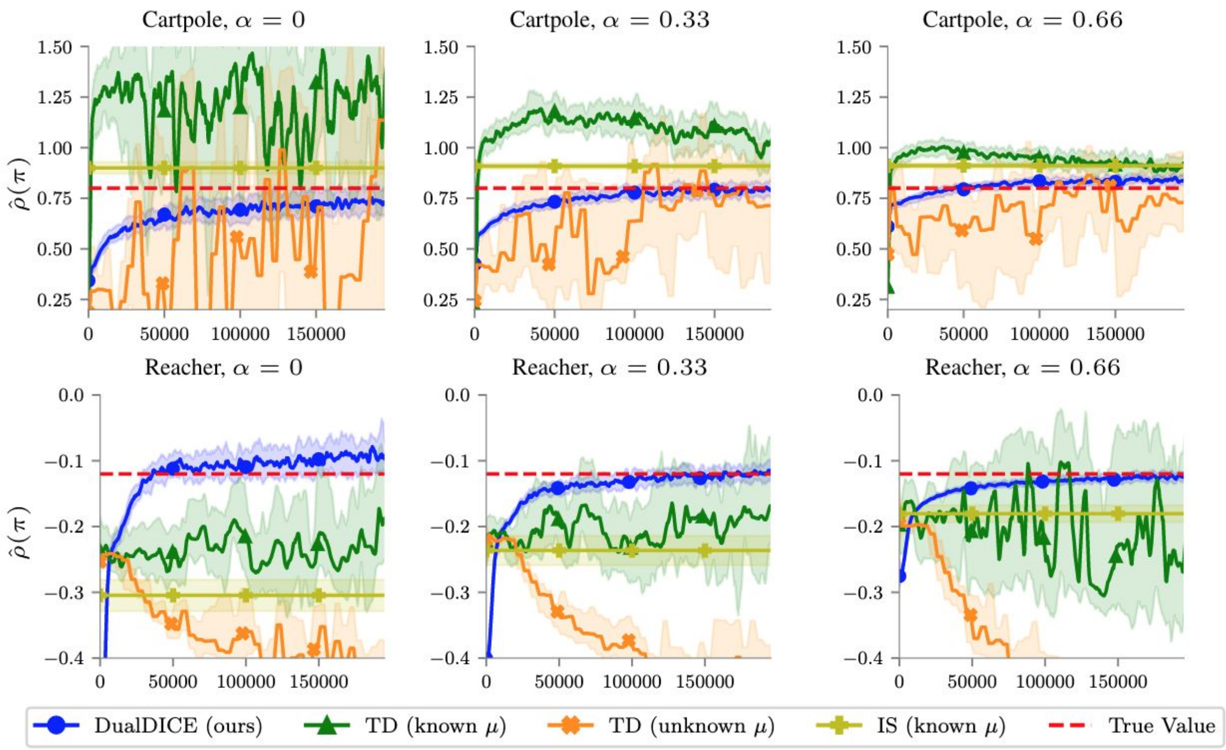


$$\boxed{\Rightarrow \nu^*(s,a) - \mathcal{B}_\pi \nu^*(s,a) = w_{\pi/\mathcal{D}}(s,a)}$$

- **Nice:** Objective is based on expectations from d^D, β, and π, which we have access to.
- **Extension 1:** Can remove appearance of Bellman operator from **both** objective and solution by application of Fenchel conjugate!
- **Extension 2:** Can generalize this result to **any convex function** (not just square)!

DualDICE Results

- DualDICE accuracy during training compared to existing methods.



DualDICE Results

East Exhibition Hall B+C

Poster #205

- DualDICE accuracy during training compared to existing methods.