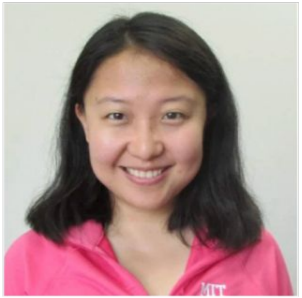
Dichotomy of Control: Separating What You Can Control from What You Cannot

Sherry Yang, Dale Schuurmans, Pieter Abbeel, Ofir Nachum



ICLR 2023 Notable **Top 5%**

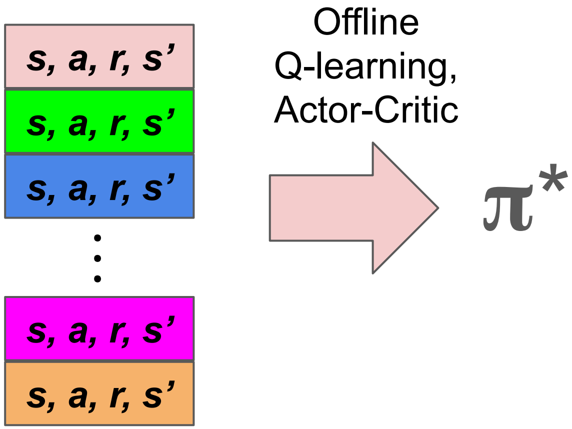
Google Brain
UNIVERSITY OF ALBERTA
Berkeley UNIVERSITY OF CALIFORNIA

Background: Offline Sequential Decision Making

Offline RL: Given an offline dataset, learn an optimal policy using RL algos.

Background: Offline Sequential Decision Making

Offline RL: Given an offline dataset, learn an optimal policy using RL algos.



Background: Offline Sequential Decision Making

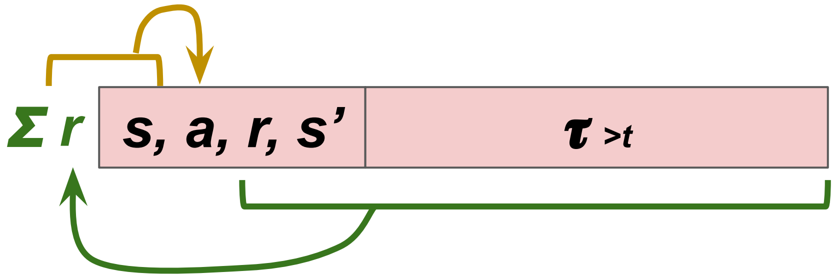
Offline RL: Given an offline dataset, learn an optimal policy using RL algos.

Return-Conditioned Supervised Learning: Imitate actions conditioned on future returns (Decision Transformer).

Background: Offline Sequential Decision Making

Offline RL: Given an offline dataset, learn an optimal policy using RL algos.

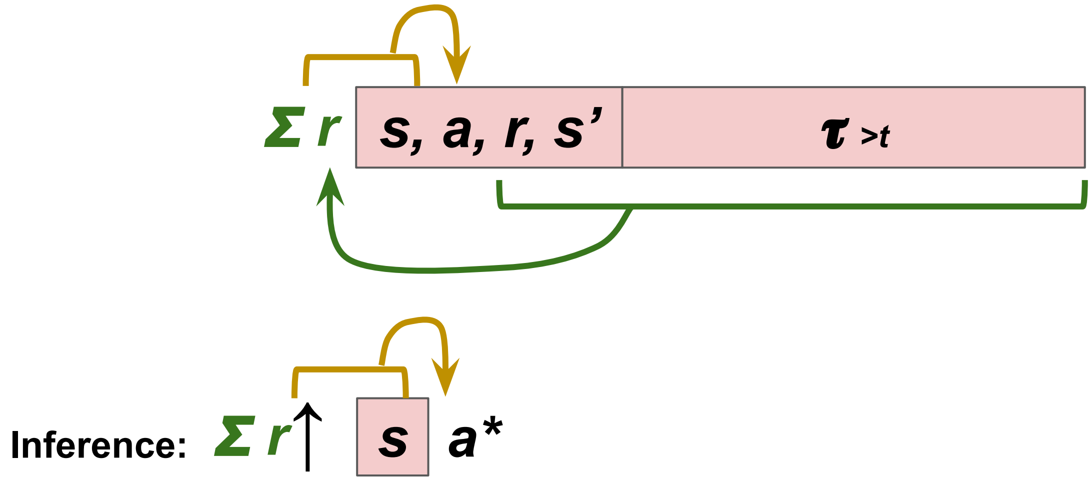
Return-Conditioned Supervised Learning: Imitate actions conditioned on future returns (Decision Transformer).



Background: Offline Sequential Decision Making

Offline RL: Given an offline dataset, learn an optimal policy using RL algos.

Return-Conditioned Supervised Learning: Imitate actions conditioned on future returns (Decision Transformer).



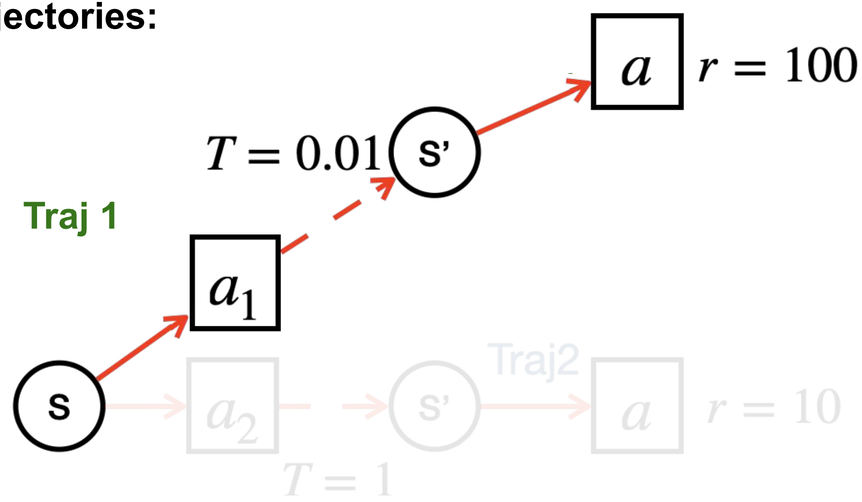
Background: Failures of RCSL

Stochastic Environments: High return arise from randomness in the environment rather than the actions themselves.

Background: Failures of RCSL

Stochastic Environments: High return arise from randomness in the environment rather than the actions themselves.

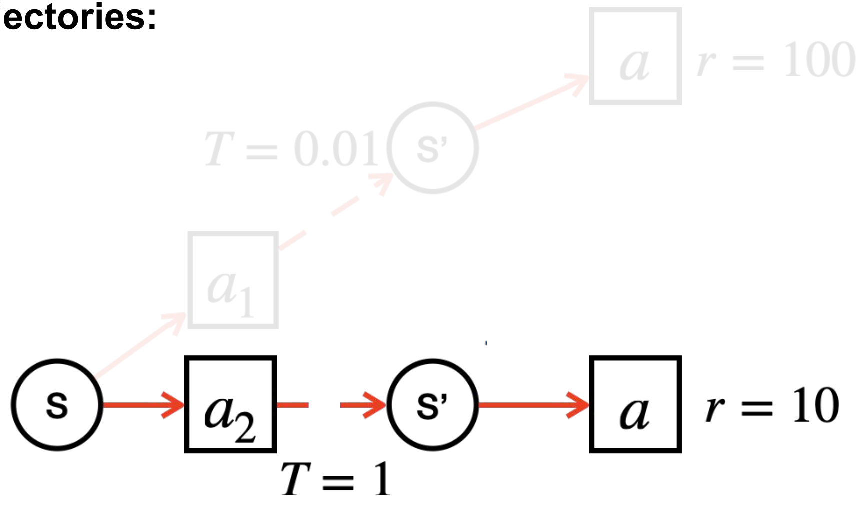
Two trajectories:



Background: Failures of RCSL

Stochastic Environments: High return arise from randomness in the environment rather than the actions themselves.

Two trajectories:

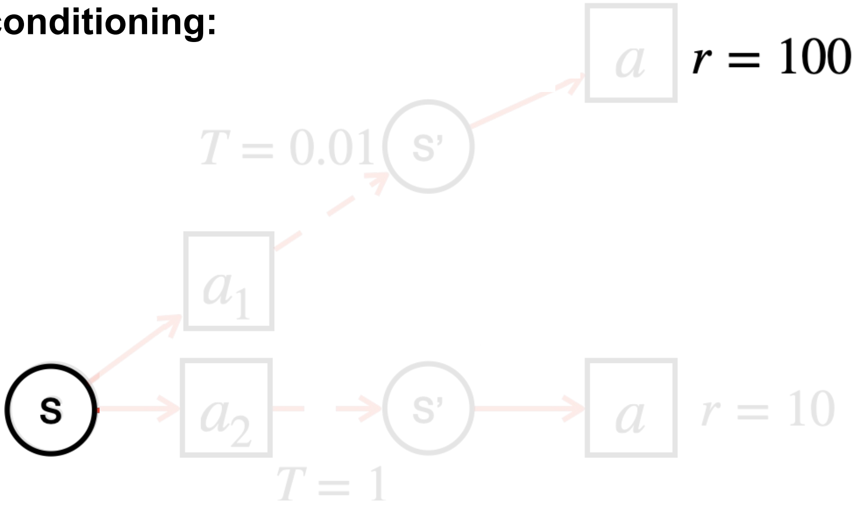


Traj 2

Background: Failures of RCSL

Failures of RCSL: Conditions on the high return that was a result of randomness in the environment.

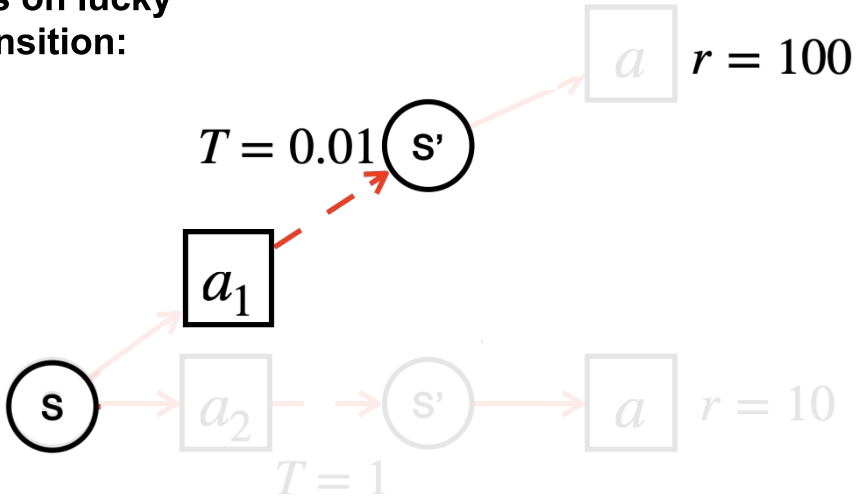
Return conditioning:



Background: Failures of RCSL

Failures of RCSL: Conditions on the high return that was a result of randomness in the environment.

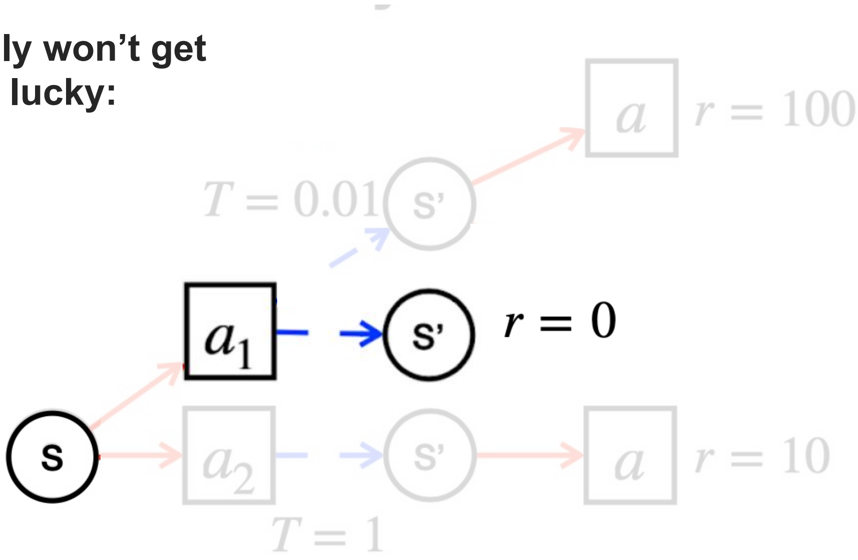
Relies on lucky transition:



Background: Failures of RCSL

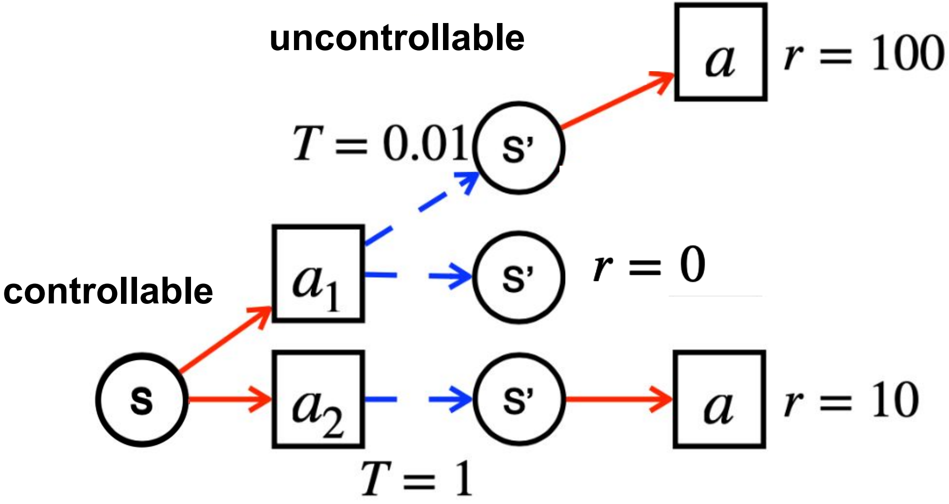
Failures of RCSL: Conditions on the high return that was a result of randomness in the environment.

But likely won't get so lucky:



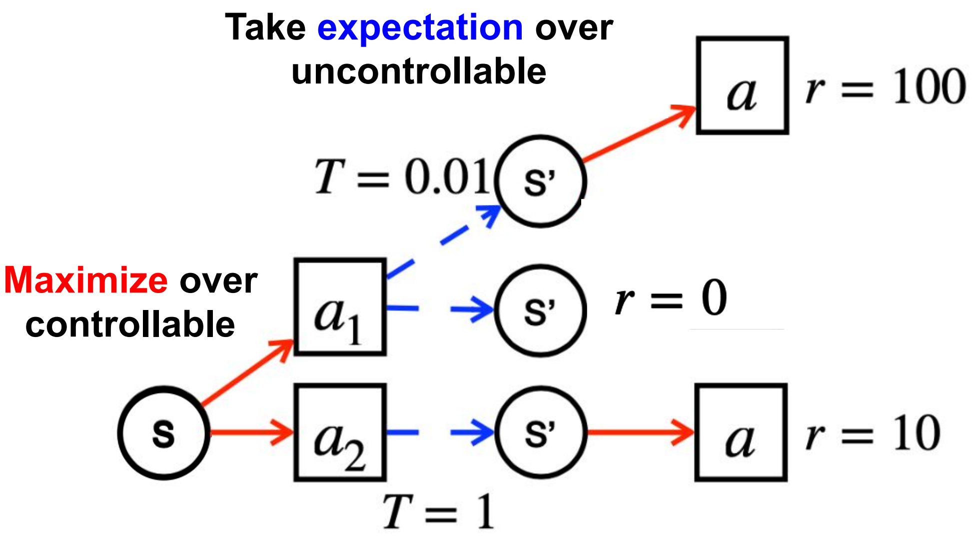
Background: Failures of RCSL

Failures of RCSL: No distinction between stochasticity of the policy (controllable) and stochasticity of the environment (uncontrollable).



Overcome Failures of RCSL

Dichotomy of Control: Separate stochasticity of the policy (controllable) and stochasticity of the environment (uncontrollable).



Overcome Failures of RCSL

Dichotomy of Control: Separate stochasticity of the policy (controllable) and stochasticity of the environment (uncontrollable).

> *"Grant me the serenity to accept the things one cannot change,*
>
> *courage to change the things one can,*
>
> *and the wisdom to know the difference"*
>
> — *Stoic Philosophy*

Outline

Formal Setup

Dichotomy of Control Objective

Consistency Guarantees

Experimental Results

Formal Setup: Return-Conditioned Supervised Learning

Given: Generic offline episodes $\tau := (s_t, a_t, r_t)_{t=0}^{H}$ and $z(\tau) = R(\tau) = \sum_{t=0}^{H} r_t$

Formal Setup: Return-Conditioned Supervised Learning

Given: Generic offline episodes $\tau := (s_t, a_t, r_t)_{t=0}^{H}$ and $z(\tau) = R(\tau) = \sum_{t=0}^{H} r_t$

RCSL: Learn policy π by maximum likelihood:

$$\mathcal{L}_{\mathrm{RCSL}}(\pi) := \mathbb{E}_{\tau \sim \mathcal{D}} \left[\sum_{t=0}^{H} -\log \pi(a_t | \tau_{0:t-1}, s_t, z(\tau)) \right]$$

Non-Markov

$$r_t \sim \mathcal{R}(\tau_{0:t-1}, s_t, a_t)$$
$$s_{t+1} \sim \mathcal{T}(\tau_{0:t-1}, s_t, a_t)$$

Formal Setup: Return-Conditioned Supervised Learning

Given: Generic offline episodes $\tau := (s_t, a_t, r_t)_{t=0}^{H}$ and $z(\tau) = R(\tau) = \sum_{t=0}^{H} r_t$

RCSL: Learn policy π by maximum likelihood:

$$\mathcal{L}_{\mathrm{RCSL}}(\pi) := \mathbb{E}_{\tau \sim \mathcal{D}} \left[\sum_{t=0}^{H} - \log \pi(a_t | \tau_{0:t-1}, s_t, z(\tau)) \right]$$

Inconsistency: Policy conditioned on z does not achieve z **in expectation**

$$V_{\mathcal{M}}(\pi_z) := \mathbb{E}_{\tau \sim \Pr[\cdot | \pi_z, \mathcal{M}]} [R(\tau)] \qquad V_{\mathcal{M}}(\pi_z) \neq z$$

Formal Setup: Return-Conditioned Supervised Learning

Given: Generic offline episodes $\tau := (s_t, a_t, r_t)_{t=0}^{H}$ and $z(\tau) = R(\tau) = \sum_{t=0}^{H} r_t$

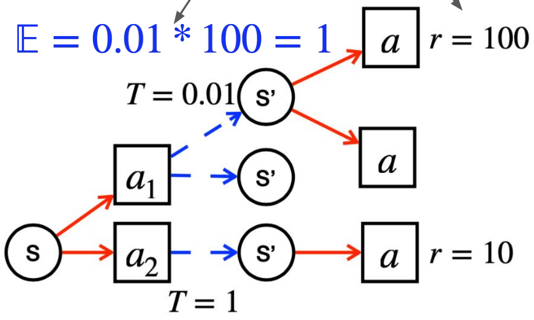
RCSL: Learn policy π by maximum likelihood:

$$\mathcal{L}_{\mathrm{RCSL}}(\pi) := \mathbb{E}_{\tau\sim\mathcal{D}}\left[\sum_{t=0}^{H} -\log \pi(a_t|\tau_{0:t-1}, s_t, z(\tau))\right]$$

Inconsistency: Policy conditioned on z does not achieve z **in expectation**

$$V_{\mathcal{M}}(\pi_z) := \mathbb{E}_{\tau\sim\Pr[\cdot|\pi_z,\mathcal{M}]}[R(\tau)] \qquad V_{\mathcal{M}}(\pi_z) \neq z$$

$\mathbb{E} = 0.01 * 100 = 1$



Formal Setup: Return-Conditioned Supervised Learning

Given: Generic offline episodes $\tau := (s_t, a_t, r_t)_{t=0}^{H}$ and $z(\tau) = R(\tau) = \sum_{t=0}^{H} r_t$

RCSL: Learn policy π by maximum likelihood:

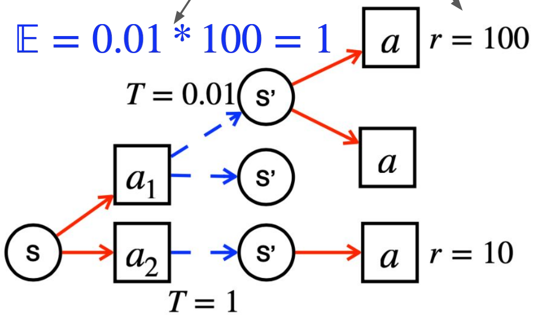
$$\mathcal{L}_{\mathrm{RCSL}}(\pi) := \mathbb{E}_{\tau \sim \mathcal{D}}\left[\sum_{t=0}^{H} -\log \pi(a_t | \tau_{0:t-1}, s_t, z(\tau))\right]$$

Inconsistency: Policy conditioned on z does not achieve z **in expectation**

$$V_{\mathcal{M}}(\pi_z) := \mathbb{E}_{\tau \sim \Pr[\cdot | \pi_z, \mathcal{M}]}[R(\tau)] \qquad V_{\mathcal{M}}(\pi_z) \neq z$$

Depends on entire τ

$\mathbb{E} = 0.01 * 100 = 1$



Formal Setup: Return-Conditioned Supervised Learning

Given: Generic offline episodes $\tau := (s_t, a_t, r_t)_{t=0}^{H}$ and $z(\tau) = R(\tau) = \sum_{t=0}^{H} r_t$

RCSL: Learn policy π by maximum likelihood:

$$\mathcal{L}_{\text{RCSL}}(\pi) := \mathbb{E}_{\tau \sim \mathcal{D}} \left[\sum_{t=0}^{H} - \log \pi(a_t | \tau_{0:t-1}, s_t, z(\tau)) \right]$$

Inconsistency: Policy conditioned on z does not achieve z **in expectation**

$$V_{\mathcal{M}}(\pi_z) := \mathbb{E}_{\tau \sim \Pr[\cdot | \pi_z, \mathcal{M}]} [R(\tau)] \qquad V_{\mathcal{M}}(\pi_z) \neq z$$

Attempt: Condition policy on stochastic future

$$\mathcal{L}_{\text{VAE}}(\pi, q, p) := \mathbb{E}_{\tau \sim \mathcal{D}, \boxed{z \sim q(z|\tau)}} \left[\sum_{t=0}^{H} - \log \pi(a_t | \tau_{0:t-1}, s_t, z) \right] + \beta \cdot \mathbb{E}_{\tau \sim \mathcal{D}} \left[D_{\text{KL}}(q(z|\tau) \| p(z|s_0)) \right]$$

Formal Setup: Return-Conditioned Supervised Learning

Given: Generic offline episodes $\tau := (s_t, a_t, r_t)_{t=0}^{H}$ and $z(\tau) = R(\tau) = \sum_{t=0}^{H} r_t$

RCSL: Learn policy π by maximum likelihood:

$$\mathcal{L}_{\mathrm{RCSL}}(\pi) := \mathbb{E}_{\tau \sim \mathcal{D}} \left[\sum_{t=0}^{H} -\log \pi(a_t | \tau_{0:t-1}, s_t, z(\tau)) \right]$$

Inconsistency: Policy conditioned on z does not achieve z **in expectation**

$$V_{\mathcal{M}}(\pi_z) := \mathbb{E}_{\tau \sim \Pr[\cdot | \pi_z, \mathcal{M}]} [R(\tau)] \qquad V_{\mathcal{M}}(\pi_z) \neq z$$

Attempt: Condition policy on stochastic future

$$\mathcal{L}_{\mathrm{VAE}}(\pi, q, p) := \mathbb{E}_{\tau \sim \mathcal{D}, z \sim q(z|\tau)} \left[\sum_{t=0}^{H} -\log \pi(a_t | \tau_{0:t-1}, s_t, z) \right] + \beta \cdot \mathbb{E}_{\tau \sim \mathcal{D}} \left[D_{\mathrm{KL}}(q(z|\tau) \| p(z|s_0)) \right]$$

z still contains entire τ

Dichotomy of Control Objective

Attempt: Condition policy on stochastic future

$$\mathcal{L}_{\mathrm{VAE}}(\pi, q, p) := \mathbb{E}_{\tau\sim\mathcal{D}, z\sim q(z|\tau)}\left[\sum_{t=0}^{H} -\log \pi(a_t|\tau_{0:t-1}, s_t, z)\right] + \beta \cdot \mathbb{E}_{\tau\sim\mathcal{D}}\left[D_{\mathrm{KL}}(q(z|\tau)\|p(z|s_0))\right]$$

z still contains entire τ

Dichotomy of Control: Condition on future without stochastic environment info.

$$\mathcal{L}_{\mathrm{DoC}}(\pi, q) := \mathbb{E}_{\tau\sim\mathcal{D}, z\sim q(z|\tau)}\left[\sum_{t=0}^{H} -\log \pi(a_t|\tau_{0:t-1}, s_t, z)\right]$$

$$\text{s.t.} \quad \mathrm{MI}(r_t; z \mid \tau_{0:t-1}, s_t, a_t) = 0, \mathrm{MI}(s_{t+1}; z \mid \tau_{0:t-1}, s_t, a_t) = 0,$$
$$\forall\, \tau_{0:t-1}, s_t, a_t \text{ and } 0 \le t \le H,$$

Dichotomy of Control Objective

Attempt: Condition policy on stochastic future

$$\mathcal{L}_{\text{VAE}}(\pi, q, p) := \mathbb{E}_{\tau\sim\mathcal{D}, z\sim q(z|\tau)}\left[\sum_{t=0}^{H} -\log \pi(a_t|\tau_{0:t-1}, s_t, z)\right] + \beta \cdot \mathbb{E}_{\tau\sim\mathcal{D}}\left[D_{\text{KL}}(q(z|\tau)\|p(z|s_0))\right]$$

z still contains entire τ

Dichotomy of Control: Condition on future without stochastic environment info.

$$\mathcal{L}_{\text{DoC}}(\pi, q) := \mathbb{E}_{\tau\sim\mathcal{D}, z\sim q(z|\tau)}\left[\sum_{t=0}^{H} -\log \pi(a_t|\tau_{0:t-1}, s_t, z)\right]$$

$$\text{s.t.}\quad \text{MI}(r_t; z \mid \tau_{0:t-1}, s_t, a_t) = 0, \text{MI}(s_{t+1}; z \mid \tau_{0:t-1}, s_t, a_t) = 0,$$
$$\forall\ \tau_{0:t-1}, s_t, a_t \text{ and } 0 \le t \le H,$$

Cannot predict future environment stochasticity from z

Dichotomy of Control Objective

Attempt: Condition policy on stochastic future

$$\mathcal{L}_{\text{VAE}}(\pi, q, p) := \mathbb{E}_{\tau\sim\mathcal{D}, z\sim q(z|\tau)}\left[\sum_{t=0}^{H} -\log \pi(a_t|\tau_{0:t-1}, s_t, z)\right] + \beta \cdot \mathbb{E}_{\tau\sim\mathcal{D}}\left[D_{\text{KL}}(q(z|\tau)\|p(z|s_0))\right]$$

z still contains entire τ

Dichotomy of Control: Condition on future without stochastic environment info.

$$\mathcal{L}_{\text{DoC}}(\pi, q) := \mathbb{E}_{\tau\sim\mathcal{D}, z\sim q(z|\tau)}\left[\sum_{t=0}^{H} -\log \pi(a_t|\tau_{0:t-1}, s_t, z)\right]$$

$$\text{s.t.}\quad \text{MI}(r_t; z \mid \tau_{0:t-1}, s_t, a_t) = 0, \text{MI}(s_{t+1}; z \mid \tau_{0:t-1}, s_t, a_t) = 0,$$

$$\forall\, \tau_{0:t-1}, s_t, a_t \text{ and } 0 \leq t \leq H,$$

$$+\beta \cdot \sum_{t=0}^{H} \mathbb{E}_{\tau\sim\mathcal{D}, z\sim q(z|\tau)}\left[f(r_t, s_{t+1}, z, \tau_{0:t-1}, s_t, a_t) - \log \mathbb{E}_{\rho(\tilde{r}, \tilde{s}')}\left[\exp\{f(\tilde{r}, \tilde{s}', z, \tau_{0:t-1}, s_t, a_t)\}\right]\right]$$

Dichotomy of Control Objective

Inference: Choose the z with the highest expected return.

Dichotomy of Control Objective

Inference: Choose the z with the highest expected return.

(1) Sample a large number of potential values of z,

(2) Estimate the expected return for each of these values of z,

Dichotomy of Control Objective

Inference: Choose the z with the highest expected return.

(1) Sample a large number of potential values of z,

(2) Estimate the expected return for each of these values of z,

$$\mathcal{L}_{\text{aux}}(V, p) = \mathbb{E}_{\tau\sim\mathcal{D}, z\sim q(z|\tau)}\left[(V(z) - R(\tau))^2 + D_{\text{KL}}(\text{stopgrad}(q(z|\tau)) \| p(z|s_0)) \right].$$

Dichotomy of Control Objective

Inference: Choose the z with the highest expected return.

(1) Sample a large number of potential values of z,

(2) Estimate the expected return for each of these values of z,

$$\mathcal{L}_{\text{aux}}(V, p) = \mathbb{E}_{\tau \sim \mathcal{D}, z \sim q(z|\tau)} \left[(V(z) - R(\tau))^2 + D_{\text{KL}}(\text{stopgrad}(q(z|\tau)) \| p(z|s_0)) \right].$$

Algorithm 1 Inference with Dichotomy of Control

Inputs Policy $\pi(\cdot|\cdot,\cdot,\cdot)$, prior $p(\cdot)$, value function $V(\cdot)$, initial state s_0, number of samples hyperparameter K.

Initialize $z^*; V^*$ ▷ `Track the best latent and its value.`

for $k = 1$ to K **do**

  Sample $z_k \sim p(z|s_0)$ ▷ `Sample a latent from the learned prior.`

  if $V(z_k) > V^*$ **then**

    $z^* = z_k; V^* = V$ ▷ `Set best latent to the one with the highest value.`

return $\pi(\cdot|\cdot,\cdot,z^*)$ ▷ `Policy conditioned on the best` z^*.

Outline

Formal Setup

Dichotomy of Control Objective

Consistency Guarantees

Experimental Results

Consistency Guarantees

Definition 1 (Consistency). *A future-conditioned policy* π *and value function* V *are* **consistent** *for a specific conditioning input* z *if the expected return of* z *predicted by* V *is equal to the true expected return of* π_z *in the environment:* $V(z) = V_{\mathcal{M}}(\pi_z)$.

Consistency Guarantees

Definition 1 (Consistency). *A future-conditioned policy π and value function V are* **consistent** *for a specific conditioning input z if the expected return of z predicted by V is equal to the true expected return of π_z in the environment: $V(z) = V_{\mathcal{M}}(\pi_z)$.*

Assumption 2 (Data and environment agreement).

Consistency Guarantees

Definition 1 (Consistency). *A future-conditioned policy* π *and value function* V *are* **consistent** *for a specific conditioning input* z *if the expected return of* z *predicted by* V *is equal to the true expected return of* π_z *in the environment:* $V(z) = V_{\mathcal{M}}(\pi_z)$.

Assumption 2 (Data and environment agreement).

Assumption 3 (No optimization or approximation errors).

Consistency Guarantees

Definition 1 (Consistency). *A future-conditioned policy* π *and value function* V *are* **consistent** *for a specific conditioning input* z *if the expected return of* z *predicted by* V *is equal to the true expected return of* π_z *in the environment:* $V(z) = V_{\mathcal{M}}(\pi_z)$.

Assumption 2 (Data and environment agreement).

Assumption 3 (No optimization or approximation errors).

Theorem 4. *Suppose DoC yields* π, V, q *with* q *satisfying the MI constraints:*

$$\mathrm{MI}(r_t; z|\tau_{0:t-1}, s_t, a_t) = \mathrm{MI}(s_{t+1}; z|\tau_{0:t-1}, s_t, a_t) = 0, \tag{10}$$

for all $\tau_{0:t-1}, s_t, a_t$ *with* $\Pr[\tau_{0:t-1}, s_t, a_t|\mathcal{D}] > 0$. *Then under Assumptions 2 and 3,* V *and* π *are consistent for any* z *with* $\Pr[z|q, \mathcal{D}] > 0$.

Consistency Guarantees

Definition 1 (Consistency). *A future-conditioned policy π and value function V are* **consistent** *for a specific conditioning input z if the expected return of z predicted by V is equal to the true expected return of π_z in the environment: $V(z) = V_{\mathcal{M}}(\pi_z)$.*

Assumption 2 (Data and environment agreement).

Assumption 3 (No optimization or approximation errors).

Theorem 4. *Suppose DoC yields π, V, q with q satisfying the MI constraints:*

$$\mathrm{MI}(r_t; z|\tau_{0:t-1}, s_t, a_t) = \mathrm{MI}(s_{t+1}; z|\tau_{0:t-1}, s_t, a_t) = 0, \quad \text{Non-Markovian} \quad (10)$$

for all $\tau_{0:t-1}, s_t, a_t$ with $\Pr[\tau_{0:t-1}, s_t, a_t|\mathcal{D}] > 0$. Then under Assumptions 2 and 3, V and π are consistent for any z with $\Pr[z|q, \mathcal{D}] > 0$.

Theorem 7. *Suppose DoC yields π, V, q with q satisfying the MI constraints:*

$$\mathrm{MI}(r_t; z|s_t, a_t) = \mathrm{MI}(s_{t+1}; z|s_t, a_t) = 0, \quad \text{Markovian} \quad (11)$$

for all s_t, a_t with $\Pr[s_t, a_t|\mathcal{D}] > 0$. Then under Assumptions 2, 5, and 6, V and π are consistent for any z with $\Pr[z|q, \mathcal{D}] > 0$.

Outline

Formal Setup

Dichotomy of Control Objective

Consistency Guarantees

Experimental Results

Experiments: Stochastic Bandit

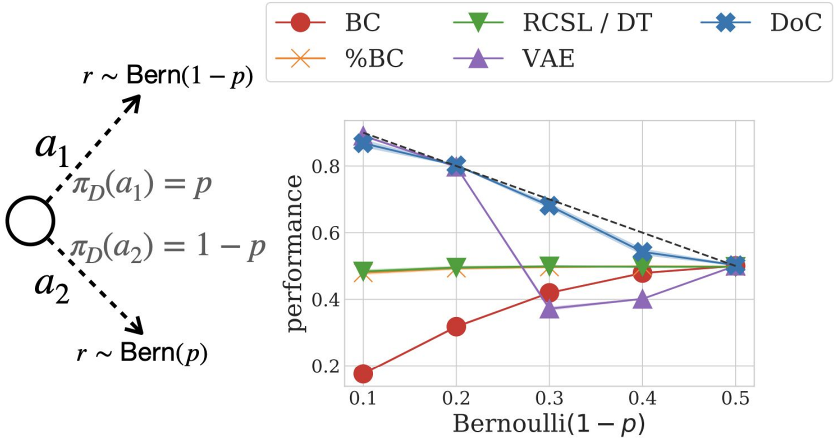


Figure 2: [Left] Bernoulli bandit where the better arm a_1 with reward $\texttt{Bern}(1-p)$ for $p < 0.5$ is pulled with probability $\pi_D(a_1) = p$ in the offline data. [Right] Average rewards achieved by DoC and baselines across 5 environment seeds. RCSL is highly suboptimal when p is small, whereas DoC achieves close to Bayes-optimal performance (dotted line) for all values of p.

Experiments: Stochastic Gridwalk

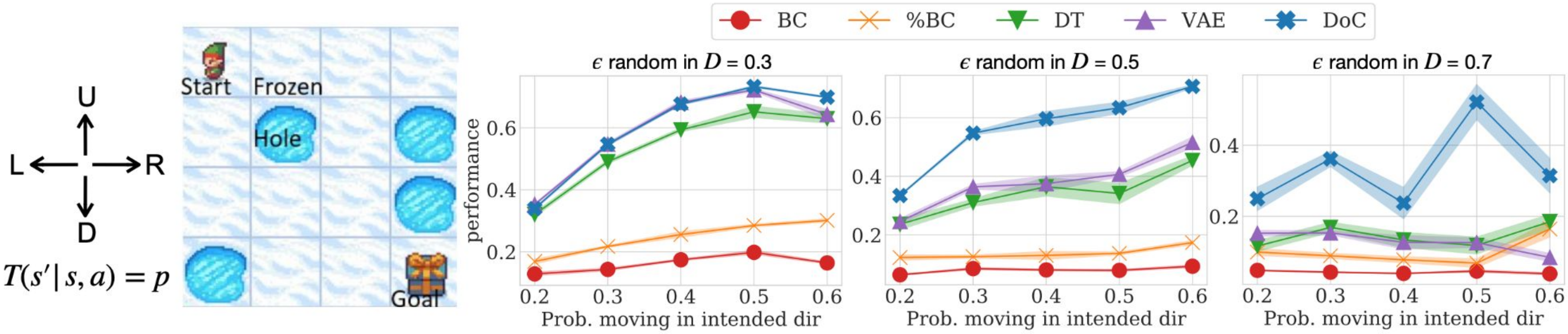


Figure 3: [Left] Visualization of the stochastic FrozenLake task. The agent has a probability p of moving in the intended direction and $1 - p$ of slipping to either sides. [Right] Average performance (across 5 seeds) of DoC and baselines on FrozenLake with different levels of stochasticity (p) and offline dataset quality (ϵ). DoC outperforms DT and future VAE, where the gain is more salient when the offline data is less optimal ($\epsilon = 0.5$ and $\epsilon = 0.7$).

Recap

Alternative to offline RL: RCSL **Inconsistent in stochastic environments.**

Dichotomy of Control **Mutual information constrained objective.**

Consistency analysis and experiments **Achieves consistency and works in practice.**

Remaining Open Questions

What else can offline RL do but RCSL cannot? **Stitching - composing suboptimal trajectories.**

Application in real-world stochastic environments? **Dialogue.**

Scale DoC to large-scale, multi-task settings? **Foundation models for decision making (arxiv)**

Thank you. Check out our paper and poster.

Today, May 2, 2023, 11:30 am - 1:30 pm,
#119