Background: Offline Sequential Decision Making

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

1. **Traj 1**
   - \( a_1 \) -> \( s' \) -> \( a \)  
   - \( T = 0.01 \)
   - \( r = 100 \)

2. **Traj 2**
   - \( a_2 \) -> \( s' \) -> \( a \)  
   - \( T = 1 \)
   - \( r = 10 \)
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

**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:

\[
T = 0.01 \quad s' \quad a_1 \quad s \quad a_2 \quad s' \quad a \quad r = 100 \\
T = 1 \quad a_2 \quad s' \quad a \quad r = 10
\]
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:
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 $z(\tau) = R(\tau) = \sum_{t=0}^H r_t$
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 $\pi$ by maximum likelihood:

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

Non-Markov

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

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

$$V_{\mathcal{M}}(\pi_z) := \mathbb{E}_{\tau \sim \text{Pr}[\cdot | \pi_z, \mathcal{M}]} [R(\tau)] \quad V_{\mathcal{M}}(\pi_z) \neq z$$
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 $\pi$ by maximum likelihood:

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

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

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

Diagram:
- $T = 0.01$
- $E = 0.01 \times 100 = 1$
- $r = 100$
- $a_1 \rightarrow s' \rightarrow a$
- $a_2 \rightarrow s' \rightarrow a$
- $r = 10$
- $T = 1$
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 $\pi$ 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

$$V_M(\pi_z) := \mathbb{E}_{\tau \sim \text{Pr}[\cdot | \pi_z, M]} [R(\tau)]$$

$$V_M(\pi_z) \neq z$$ Depends on entire $\tau$

Diagram:
- State $s$, action $a$, reward $r$
- Transition $T = 0.01$
- Reward $R = 0.01 * 100 = 100$
- Reward $R = 10$
- State $s'$, action $a$, reward $r$
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 $\pi$ 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

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

**Attempt:** Condition policy on stochastic future

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

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

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

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

Attempt: Condition policy on stochastic future

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

$z$ still contains entire $\tau$
Dichotomy of Control Objective

**Attempt:** Condition policy on stochastic future

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

\( z \) still contains entire \( \tau \)

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

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

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

\( \forall \tau_{0:t-1}, s_t, a_t \) and \( 0 \leq t \leq H, \)
Dichotomy of Control Objective

**Attempt**: Condition policy on stochastic future

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

\[z \text{ still contains entire } \tau\]

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

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

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

Cannot predict future environment stochasticity from \(z\)
Dichotomy of Control Objective

**Attempt**: Condition policy on stochastic future

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

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

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

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

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

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

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

**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 $K$.

1. Initialize $z^*; V^*$
2. for $k = 1$ to $K$
   1. Sample $z_k \sim p(z|s_0)$
   2. if $V(z_k) > V^*$ then
      1. $z^* = z_k; V^* = V$
   3. return $\pi(\cdot|\cdot, \cdot, z^*)$

▷ Track the best latent and its value.
▷ Sample a latent from the learned prior.
▷ Set best latent to the one with the highest value.
▷ 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 $\pi$ 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 $\pi_z$ in the environment: $V(z) = V_M(\pi_z)$. 
Consistency Guarantees

**Definition 1** (Consistency). A future-conditioned policy $\pi$ 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 $\pi_z$ in the environment: $V(z) = V_M(\pi_z)$.

**Assumption 2** (Data and environment agreement).
Consistency Guarantees

**Definition 1** (Consistency). A future-conditioned policy $\pi$ 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 $\pi_z$ in the environment: $V(z) = V_M(\pi_z)$.

**Assumption 2** (Data and environment agreement).

**Assumption 3** (No optimization or approximation errors).
Consistency Guarantees

**Definition 1** (Consistency). A future-conditioned policy $\pi$ 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 $\pi_z$ in the environment: $V(z) = V_M(\pi_z)$.

**Assumption 2** (Data and environment agreement).

**Assumption 3** (No optimization or approximation errors).

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

$$MI(r_t; z|\tau_0:t-1, s_t, a_t) = MI(s_{t+1}; z|\tau_0:t-1, s_t, a_t) = 0,$$

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


Consistency Guarantees

**Definition 1** (Consistency). A future-conditioned policy $\pi$ 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 $\pi_z$ in the environment: $V(z) = V_M(\pi_z)$.

**Assumption 2** (Data and environment agreement).

**Assumption 3** (No optimization or approximation errors).

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

$$\text{MI}(r_t; z|\tau_0:t-1, s_t, a_t) = \text{MI}(s_{t+1}; z|\tau_0:t-1, s_t, a_t) = 0, \quad \text{Non-Markovian} \quad (10)$$

for all $\tau_0:t-1, s_t, a_t$ with $\Pr[\tau_0:t-1, s_t, a_t|D] > 0$. Then under Assumptions 2 and 3, $V$ and $\pi$ are consistent for any $z$ with $\Pr[z|q, D] > 0$.

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

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

for all $s_t, a_t$ with $\Pr[s_t, a_t|D] > 0$. Then under Assumptions 2, 5, and 6, $V$ and $\pi$ are consistent for any $z$ with $\Pr[z|q, D] > 0$. 
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 $\text{Bern}(1 - p)$ 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 ($\epsilon$). 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