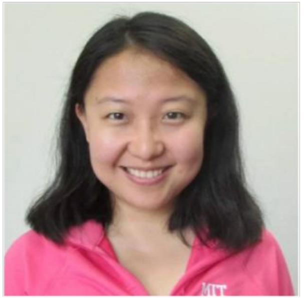
Dichotomy of Control: Separating What You Can Control from What You Can Not

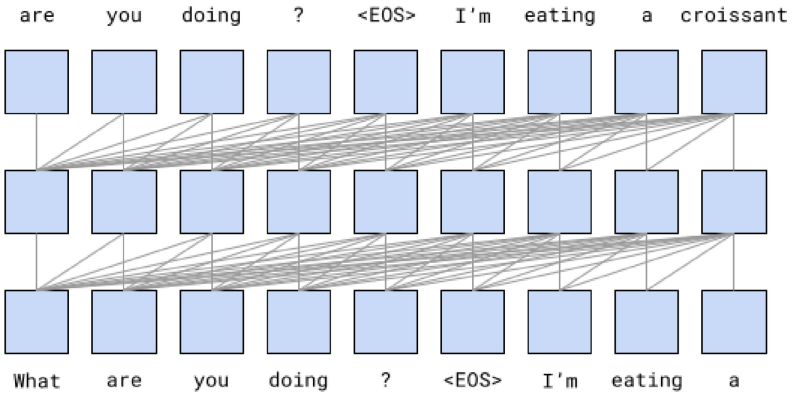
Sherry Yang
Dale Schuurmans
Pieter Abbeel
Ofir Nachum



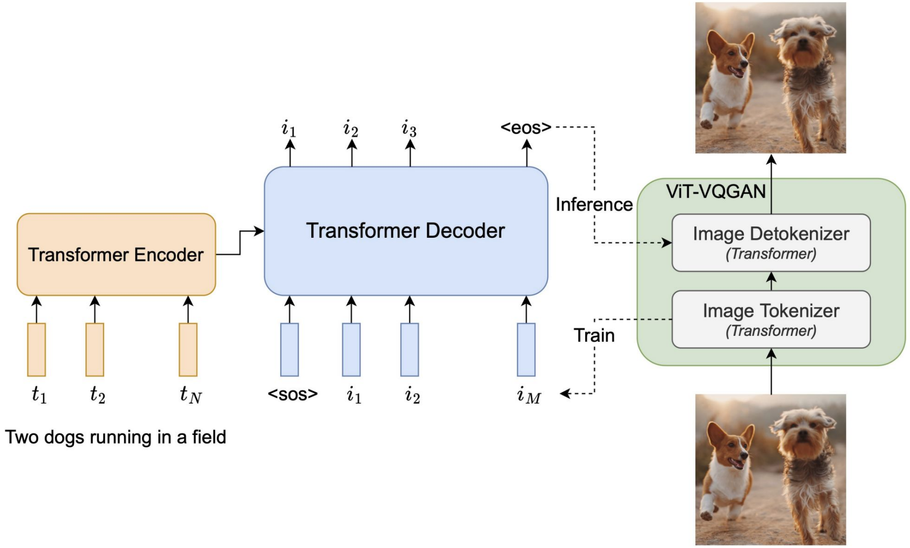

Paper: https://drive.google.com/file/d/109ktTzxF2FRkM8s412xYFVSguGvxEWCa/view?usp=sharing

Background

Training large-scale generative models has emerged as the dominant approach in NLP, vision, etc.



Thoppilan, et al. “LaMDa” (2022).

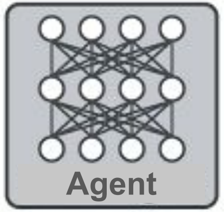


Yu, et al. “Parti” (2022).

Background

What about reinforcement learning (RL)?

Can we apply similar paradigms to RL?



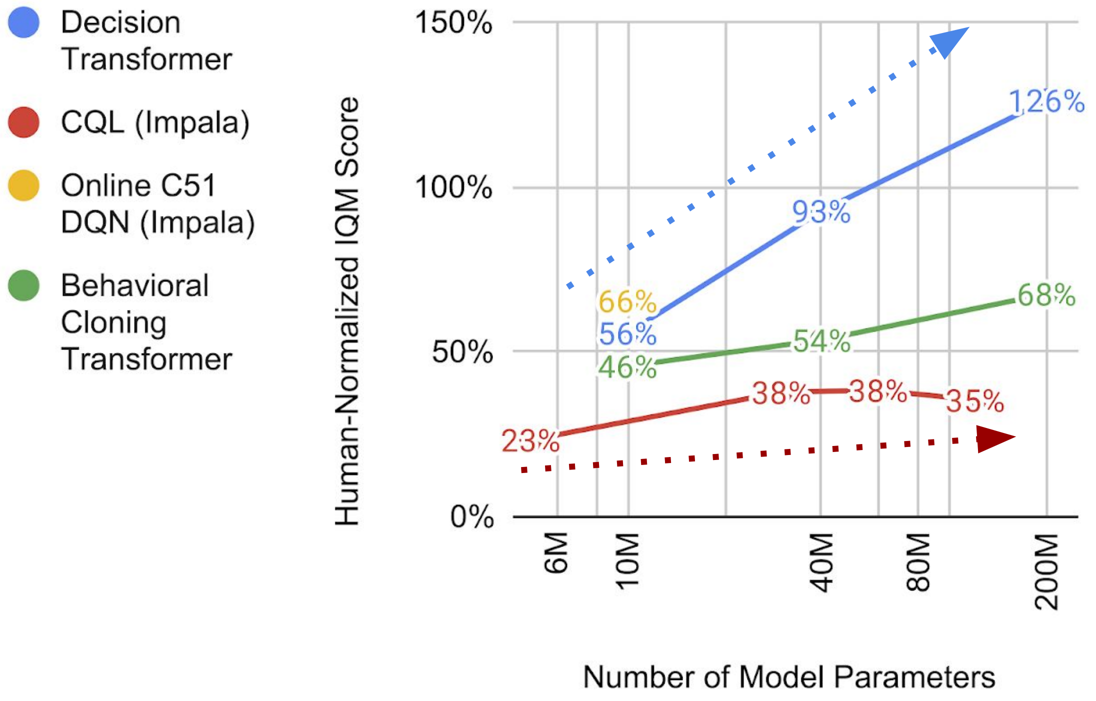
Background: Decision Transformers

MGDT: Build a generalist agent that acts in interactive environments

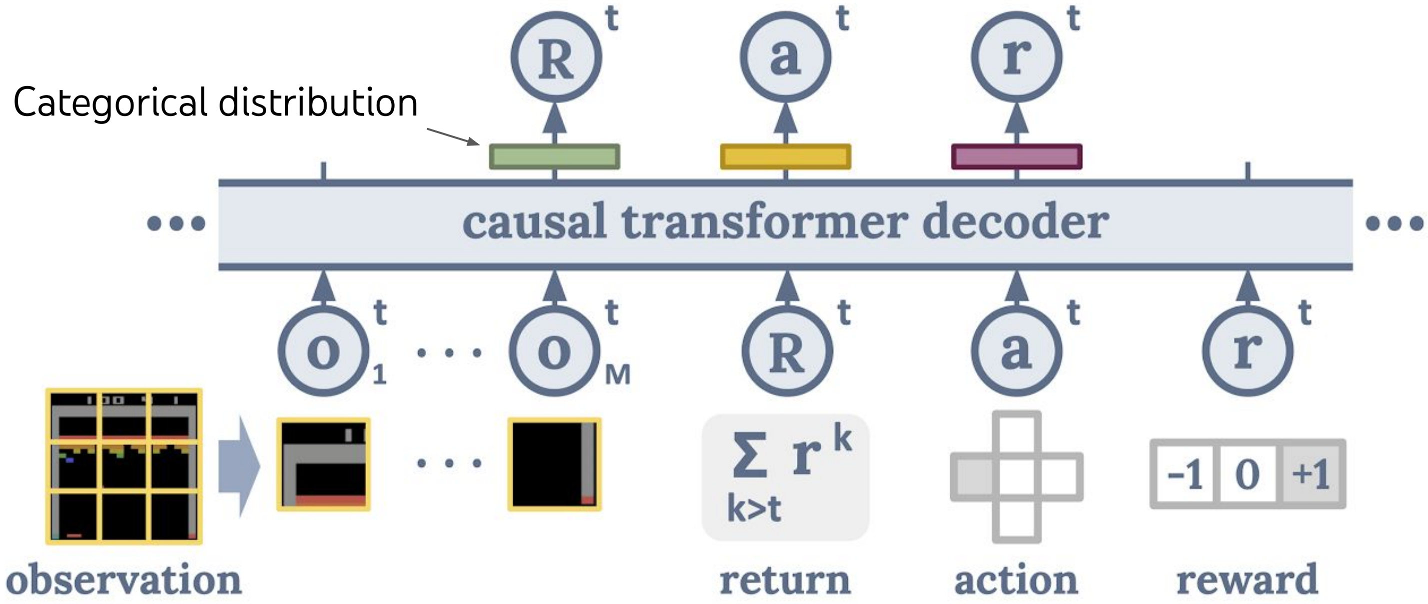
Results: One agent plays 41 Atari games. Rapid transfer to new games.

Background: Decision Transformer Scales

Have not plateaued



“o-R-a-r” Decision Transformers

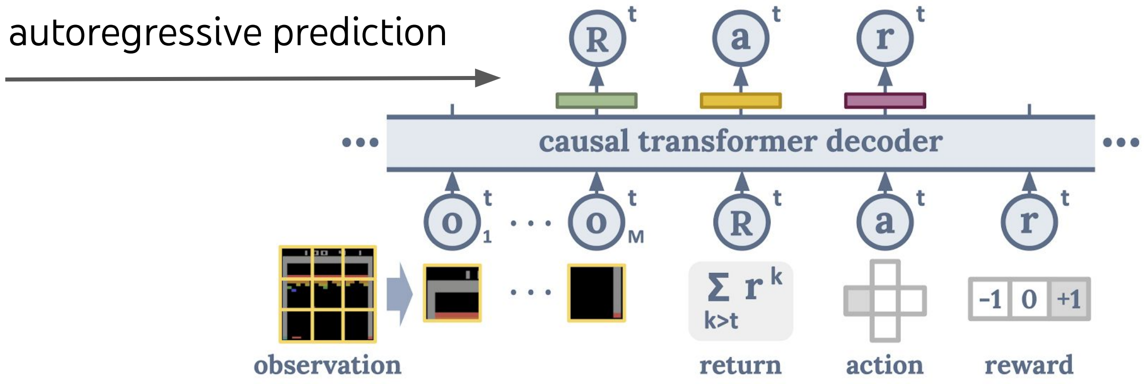


36 image patches

Discrete

Don’t Optimize for Return – Ask for Optimality

Inference

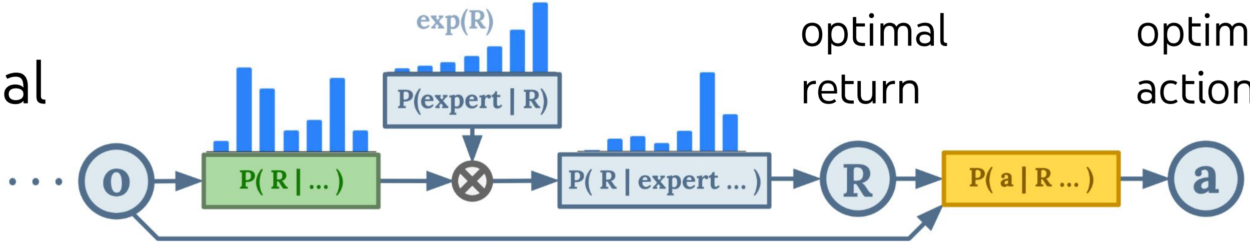


optimality bias

optimal return

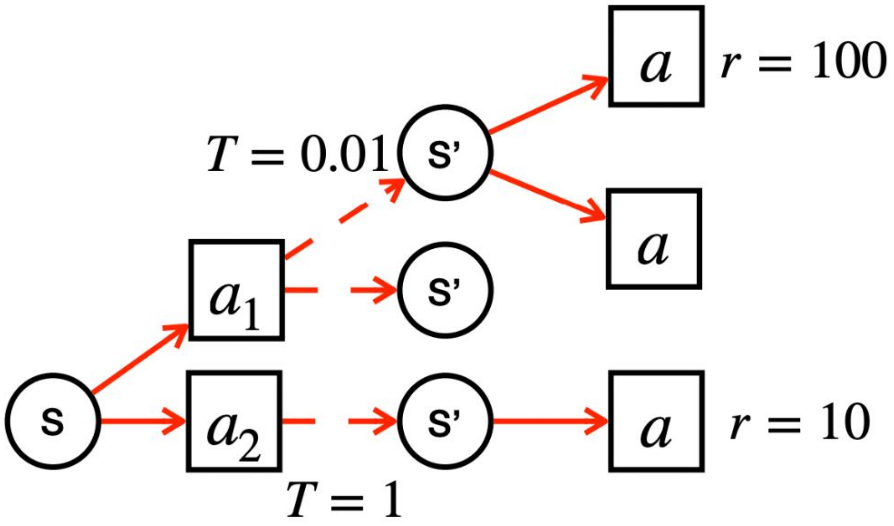
optimal action

Sample (near-)optimal target return

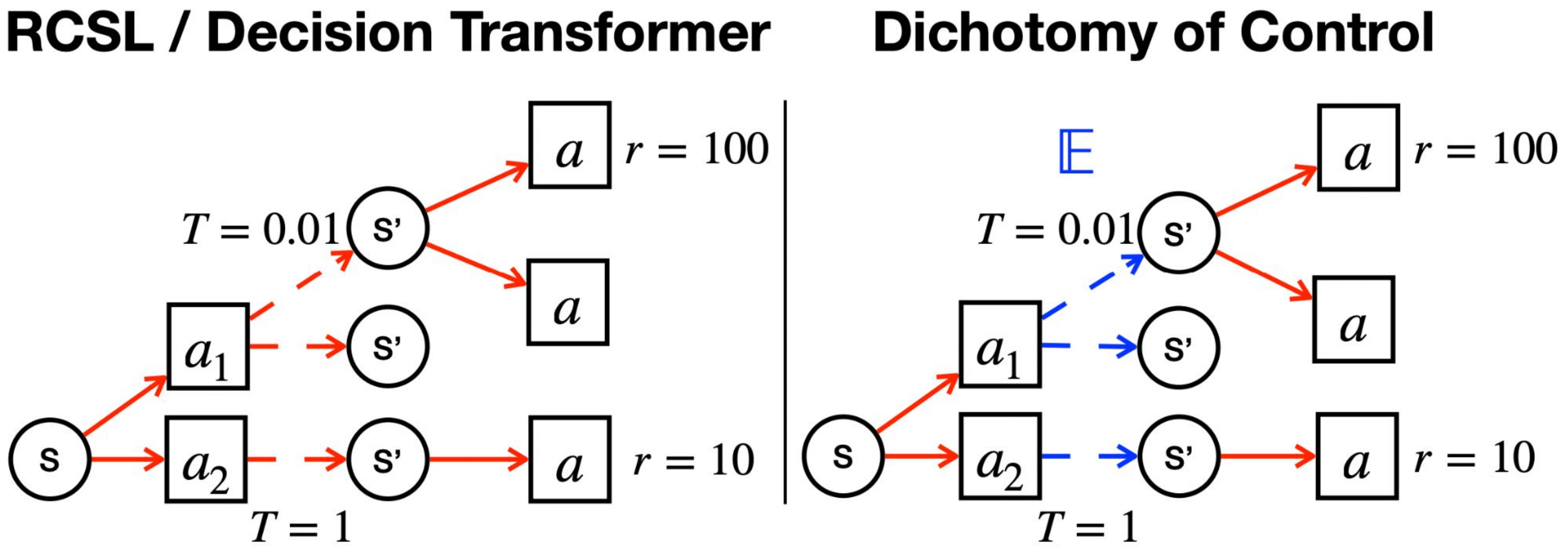


Issues with Decision Transformer - Stochasticity

RCSL / Decision Transformer



Dichotomy of Control: Control the Controllable



Return/Future Conditioned Supervised Learning

Return-conditioned supervised learning:

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

Future-conditioned supervised learning:

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

Dichotomy of Control

Max-likelihood as before

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

Dichotomy of Control

Max-likelihood as before

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

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

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

Cannot predict environment stochasticity from z

Dichotomy of Control: Practical Algorithm

$$
\begin{aligned}
&\mathrm{MI}(r_t; z|\tau_{0:t-1}, s_t, a_t) \\
&\quad = D_{\mathrm{KL}}\left(\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t] \| \Pr[r_t|\tau_{0:t-1}, s_t, a_t] \Pr[z|\tau_{0:t-1}, s_t, a_t]\right)
\end{aligned}
$$

Dichotomy of Control: Practical Algorithm

$$\begin{aligned}
&\mathrm{MI}(r_t; z|\tau_{0:t-1}, s_t, a_t) \\
&\quad = D_{\mathrm{KL}}\left(\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t] \| \Pr[r_t|\tau_{0:t-1}, s_t, a_t] \Pr[z|\tau_{0:t-1}, s_t, a_t]\right) \\
&\quad = \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]}\left[\log\left(\frac{\Pr[r_t|z, \tau_{0:t-1}, s_t, a_t]}{\Pr[r_t|\tau_{0:t-1}, s_t, a_t]}\right)\right]
\end{aligned}$$

Dichotomy of Control: Practical Algorithm

$$
\begin{aligned}
&\mathrm{MI}(r_t; z|\tau_{0:t-1}, s_t, a_t) \\
&\quad = D_{\mathrm{KL}}\left(\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t] \| \Pr[r_t|\tau_{0:t-1}, s_t, a_t]\Pr[z|\tau_{0:t-1}, s_t, a_t]\right) \\
&\quad = \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]}\left[\log\left(\frac{\Pr[r_t|z, \tau_{0:t-1}, s_t, a_t]}{\Pr[r_t|\tau_{0:t-1}, s_t, a_t]}\right)\right] \\
&\quad = \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]} \log \Pr[r_t|z, \tau_{0:t-1}, s_t, a_t] - \mathbb{E}_{\Pr[r_t|\tau_{0:t-1}, s_t, a_t]} \log \Pr[r_t|\tau_{0:t-1}, s_t, a_t].
\end{aligned}
$$

Dichotomy of Control: Practical Algorithm

$$
\begin{aligned}
&\mathrm{MI}(r_t; z|\tau_{0:t-1}, s_t, a_t) \\
&\quad = D_{\mathrm{KL}}\left(\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t] \| \Pr[r_t|\tau_{0:t-1}, s_t, a_t]\Pr[z|\tau_{0:t-1}, s_t, a_t]\right) \\
&\quad = \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]}\left[\log\left(\frac{\Pr[r_t|z, \tau_{0:t-1}, s_t, a_t]}{\Pr[r_t|\tau_{0:t-1}, s_t, a_t]}\right)\right] \\
&\quad = \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]} \log \Pr[r_t|z, \tau_{0:t-1}, s_t, a_t] - \mathbb{E}_{\Pr[r_t|\tau_{0:t-1}, s_t, a_t]} \log \Pr[r_t|\tau_{0:t-1}, s_t, a_t].
\end{aligned}
$$

$$
\omega(r_t|z, \tau_{0:t-1}, s_t, a_t) \propto \rho(r_t) \exp\left\{f(r_t, z, \tau_{0:t-1}, s_t, a_t)\right\}
$$

Dichotomy of Control: Practical Algorithm

$$
\begin{aligned}
&\mathrm{MI}(r_t; z|\tau_{0:t-1}, s_t, a_t) \\
&\quad = D_{\mathrm{KL}}\left(\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t] \| \Pr[r_t|\tau_{0:t-1}, s_t, a_t] \Pr[z|\tau_{0:t-1}, s_t, a_t]\right) \\
&\quad = \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]}\left[\log\left(\frac{\Pr[r_t|z, \tau_{0:t-1}, s_t, a_t]}{\Pr[r_t|\tau_{0:t-1}, s_t, a_t]}\right)\right] \\
&\quad = \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]} \log \Pr[r_t|z, \tau_{0:t-1}, s_t, a_t] - \mathbb{E}_{\Pr[r_t|\tau_{0:t-1}, s_t, a_t]} \log \Pr[r_t|\tau_{0:t-1}, s_t, a_t].
\end{aligned}
$$

$$\omega(r_t|z, \tau_{0:t-1}, s_t, a_t) \propto \rho(r_t) \exp\left\{f(r_t, z, \tau_{0:t-1}, s_t, a_t)\right\}$$

$$
\begin{aligned}
&\max_{\omega}\ \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]}\left[\log \omega(r_t|\tau_{0:t-1}, s_t, a_t)\right] \\
&= \max_{f} \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]}\left[f(r_t, z, \tau_{0:t-1}, s_t, a_t) - \log \mathbb{E}_{\rho(\tilde{r})}\left[\exp\{f(\tilde{r}, z, \tau_{0:t-1}, s_t, a_t)\}\right]\right]
\end{aligned}
$$

Dichotomy of Control: Practical Algorithm

$$\begin{aligned}
&\mathrm{MI}(r_t; z|\tau_{0:t-1}, s_t, a_t) \\
&\quad = D_{\mathrm{KL}}\left(\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t] \| \Pr[r_t|\tau_{0:t-1}, s_t, a_t]\Pr[z|\tau_{0:t-1}, s_t, a_t]\right) \\
&\quad = \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]}\left[\log\left(\frac{\Pr[r_t|z, \tau_{0:t-1}, s_t, a_t]}{\Pr[r_t|\tau_{0:t-1}, s_t, a_t]}\right)\right] \\
&\quad = \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]} \log \Pr[r_t|z, \tau_{0:t-1}, s_t, a_t] - \mathbb{E}_{\Pr[r_t|\tau_{0:t-1}, s_t, a_t]} \log \Pr[r_t|\tau_{0:t-1}, s_t, a_t].
\end{aligned}$$

$$\omega(r_t|z, \tau_{0:t-1}, s_t, a_t) \propto \rho(r_t) \exp\left\{f(r_t, z, \tau_{0:t-1}, s_t, a_t)\right\}$$

$$\begin{aligned}
&\max_{\omega}\ \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]}\left[\log \omega(r_t|\tau_{0:t-1}, s_t, a_t)\right] \\
&= \max_{f} \mathbb{E}_{\Pr[r_t, z|\tau_{0:t-1}, s_t, a_t]}\left[f(r_t, z, \tau_{0:t-1}, s_t, a_t) - \log \mathbb{E}_{\rho(\tilde{r})}\left[\exp\{f(\tilde{r}, z, \tau_{0:t-1}, s_t, a_t)\}\right]\right]
\end{aligned}$$

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

Dichotomy of Control: Practical Algorithm

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^*`.`

Formalization: Inconsistency

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:* $V(z) = V_{\mathcal{M}}(\pi_z)$.

Formalization: Inconsistency

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). *The per-step reward and next-state transitions observed in the data distribution are the same as those of the environment. In other words, for any $\tau_{0:t-1}, s_t, a_t$ with $\Pr[\tau_{0:t-1}, s_t, a_t|\mathcal{D}] > 0$, we have $\Pr[\hat{r}_t = r_t|\tau_{0:t-1}, s_t, a_t, \mathcal{D}] = \mathcal{R}(\hat{r}_t|\tau_{0:t-1}, s_t, a_t)$ and $\Pr[\hat{s}_{t+1} = s_{t+1}|\tau_{0:t-1}, s_t, a_t, \mathcal{D}] = \mathcal{T}(\hat{s}_{t+1}|\tau_{0:t-1}, s_t, a_t)$ for all $\hat{r}_t, \hat{s}_{t+1}$.*

Assumption 3 (No optimization or approximation errors). *DoC yields policy π and value function V that are Bayes-optimal with respect to the training data distribution and q. In other words, $V(z) = \mathbb{E}_{\tau\sim\Pr[\cdot|z,\mathcal{D}]}[R(\tau)]$ and $\pi(\hat{a}|\tau_{0:t-1}, s_t, z) = \Pr[\hat{a} = a_t|\tau_{0:t-1}, s_t, z, \mathcal{D}]$.*

Formalization: Inconsistency

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). *The per-step reward and next-state transitions observed in the data distribution are the same as those of the environment. In other words, for any $\tau_{0:t-1}, s_t, a_t$ with $\Pr[\tau_{0:t-1}, s_t, a_t|\mathcal{D}] > 0$, we have $\Pr[\hat{r}_t = r_t|\tau_{0:t-1}, s_t, a_t, \mathcal{D}] = \mathcal{R}(\hat{r}_t|\tau_{0:t-1}, s_t, a_t)$ and $\Pr[\hat{s}_{t+1} = s_{t+1}|\tau_{0:t-1}, s_t, a_t, \mathcal{D}] = \mathcal{T}(\hat{s}_{t+1}|\tau_{0:t-1}, s_t, a_t)$ for all $\hat{r}_t, \hat{s}_{t+1}$.*

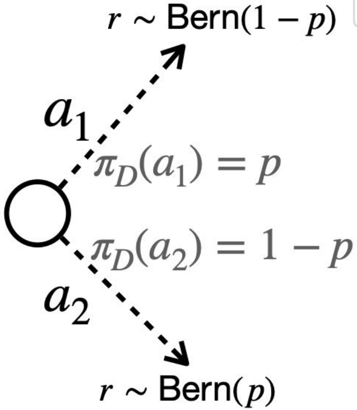
Assumption 3 (No optimization or approximation errors). *DoC yields policy π and value function V that are Bayes-optimal with respect to the training data distribution and q. In other words, $V(z) = \mathbb{E}_{\tau \sim \Pr[\cdot|z,\mathcal{D}]}[R(\tau)]$ and $\pi(\hat{a}|\tau_{0:t-1}, s_t, z) = \Pr[\hat{a} = a_t|\tau_{0:t-1}, s_t, z, \mathcal{D}]$.*

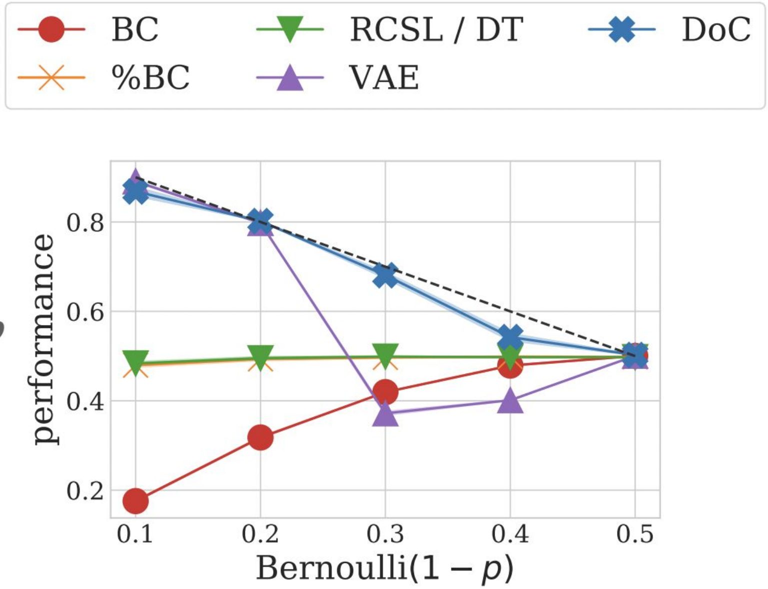
Theorem 4. *Suppose DoC yields π, V, q with q satisfying the MI constraints:*

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

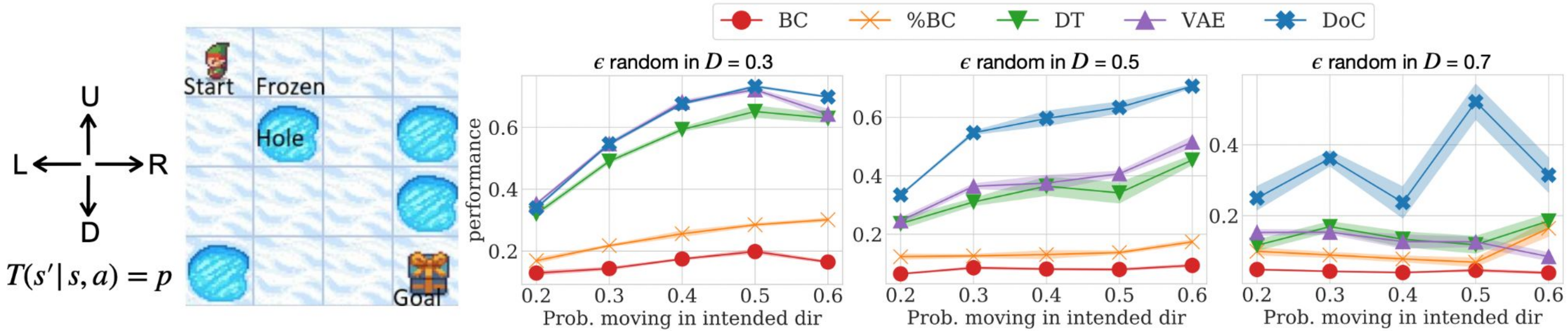
for all $\tau_{0:t-1}, s_t, a_t$ with $\Pr[\tau_{0:t-1}, s_t, a_t|\mathcal{D}] > 0$. Then under Assumptions 2 and 3, V and π are consistent for any z with $\Pr[z|q, \mathcal{D}] > 0$.

Experiments: Stochastic Bandit





Experiments: Stochastic Gridwalk, MuJoCo



Experiments: Stochastic Gridwalk, MuJoCo