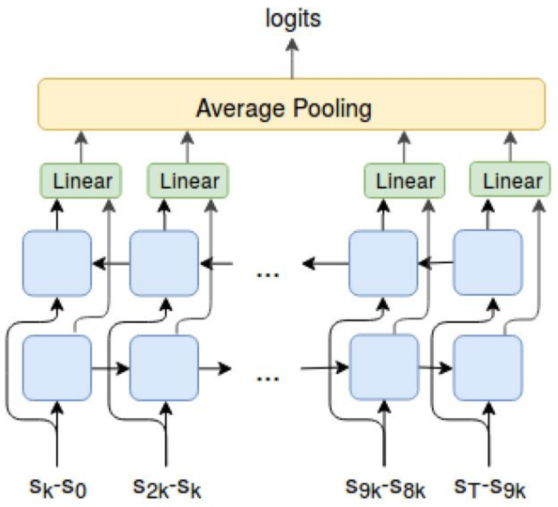
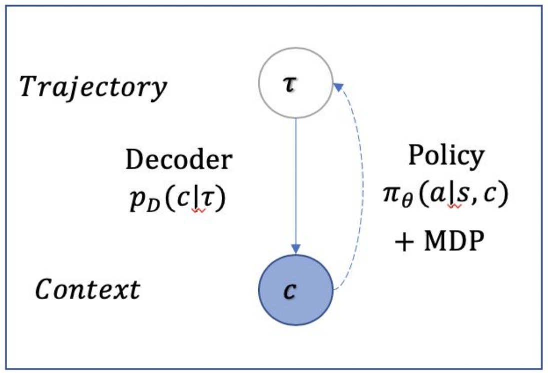
Variational Option Discovery Algorithms

Achiam, Edwards, Amodei, Abbeel

Topic: Hierarchical Reinforcement Learning

Presenter: Harris Chan





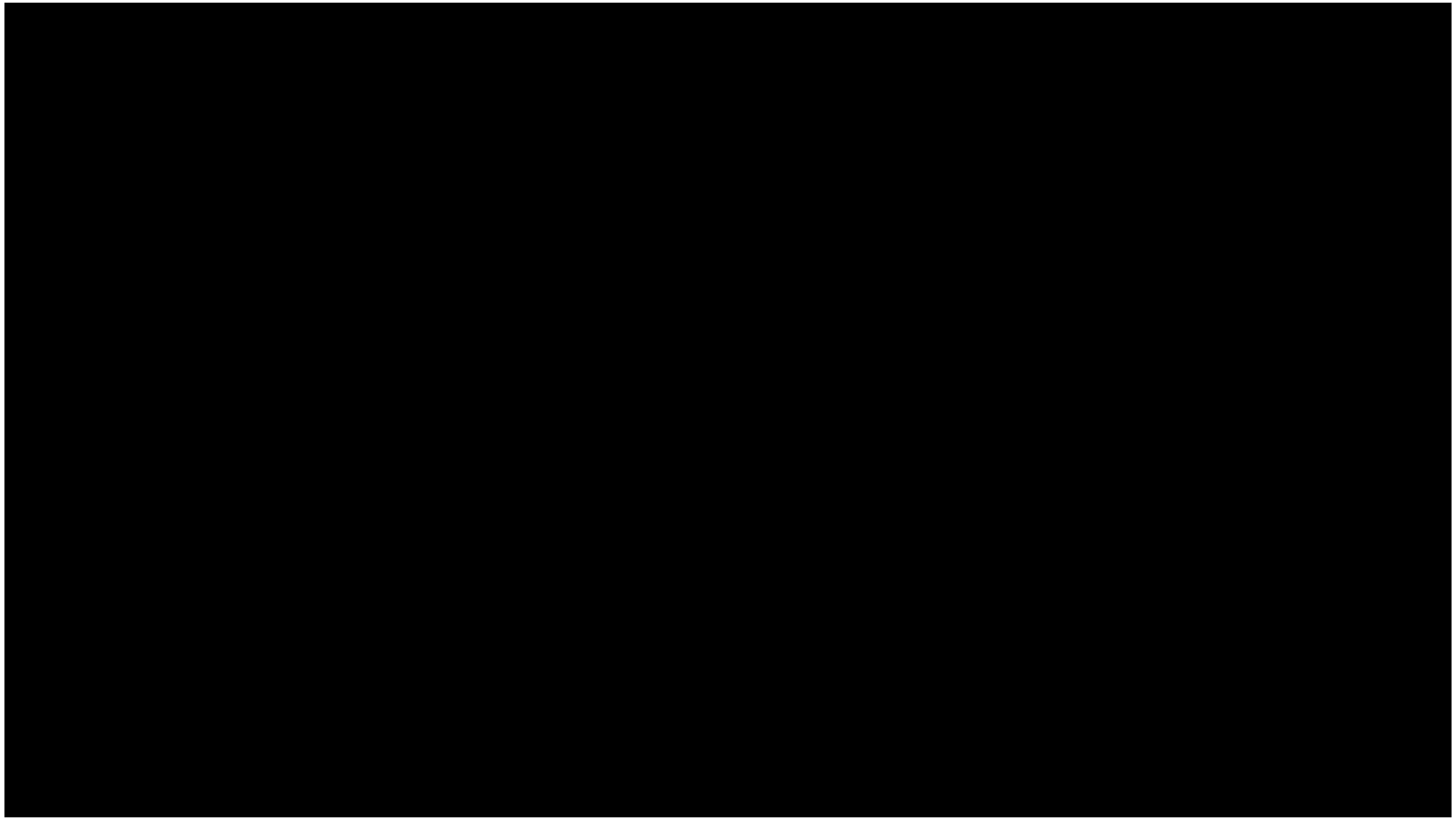
Overview

- **Motivation**: Reward-free option discovery
- **Contributions**
- **Background**: Universal Policies, Variational Autoencoder
- **Method**: Variational Option Discovery Algorithms, VALOR, Curriculum
- **Results**
- **Discussions & Limitations**

Overview

- **Motivation**: Reward-free option discovery
- **Contributions**
- **Background**: Universal Policies, Variational Autoencoder
- **Method**: Variational Option Discovery Algorithms, VALOR, Curriculum
- **Results**
- **Discussions & Limitations**

Humans find new ways to interact with environment



Motivation: Reward-Free Option Discovery

Reward-free Option Discovery: RL agent learn ***skills*** (options) **without** environment reward

Research Questions:

- How can we learn **diverse** set of skills?
- Do these skills match with human priors on what are useful skills?
- Can we use these learned skills for **downstream tasks?**

Limitations of Prior Related Works

- **Information Theoretic** approaches: mutual info between options and states, not full trajectories:

$$\max_{option} MI(option, f(sta$$

- **Multi-goal** Reinforcement learning (goal or instruction conditioned policies) requires:
  - Extrinsic reward signal (e.g. did the agent achieve the goal/instruction?)
  - Hand-crafted instruction space (e.g. XY coordinate of agent)
- **Intrinsic Motivations**: suffers from catastrophic forgetting
  - Intrinsic reward decays over time, may forget how to revisit

Overview

- **Motivation**: Reward-free option discovery
- **Contributions**
- **Background**: Universal Policies, Variational Autoencoder
- **Method**: Variational Option Discovery Algorithms, VALOR, Curriculum
- **Results**
- **Discussions & Limitations**

Contributions

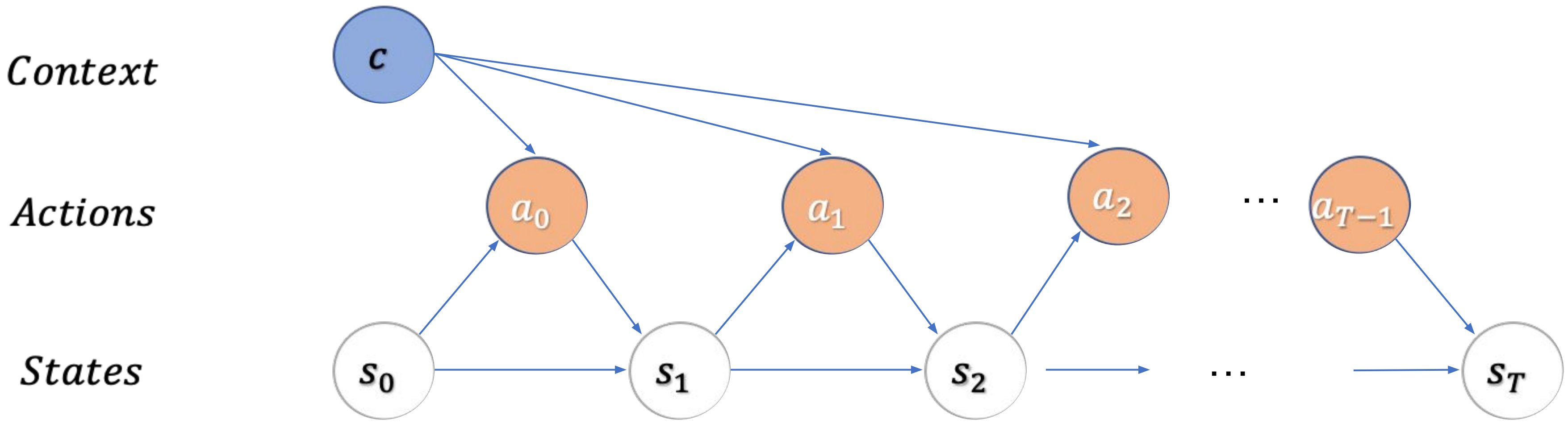
1. **Problem:** Reward-free options discovery, which aims to learn interesting behaviours without environment rewards (unsupervised)
2. Introduced a general framework **Variational Option Discovery** objective & algorithm
    1. Connected Variational Option Discovery and Variational Autoencoder (VAE)
3. Specific instantiation: **VALOR** and **Curriculum learning:**
    1. **VALOR:** a decoder architecture using Bi-LSTM over only (some) states in trajectory
    2. **Curriculum learning** for increasing number of skills when agent mastered current skills
4. Empirically tested on simulated robotics environments
    1. VALOR can learn diverse behaviours in variety of environments
    2. Learned policies are universal, can be interpolated and used in hierarchies

Overview

- **Motivation**: Reward-free option discovery
- **Contributions**
- **Background**: Universal Policies, Variational Autoencoder
- **Method**: Variational Option Discovery Algorithms, VALOR, Curriculum
- **Results**
- **Discussions & Limitations**

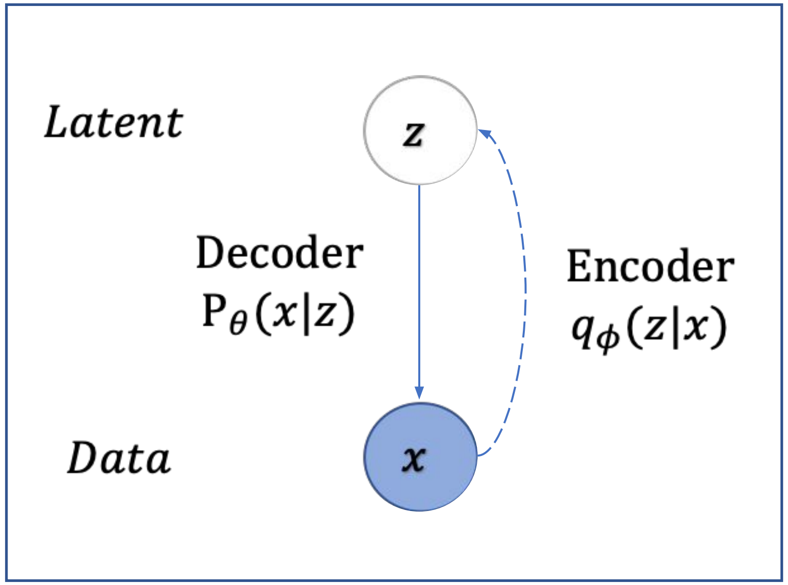
Background: Universal Policies

- **Aim:** Learn a policy $\boldsymbol{\pi(a|s,c)}$ conditioned on state s and **context** c
- Context is sampled at *beginning of episode* and *fixed throughout*



Background: Variational Autoencoders (VAE)

- **Aim:** Learn *encoder* $q_\phi(z|x)$ conditioned on **data x** for **latent variable z**, and *decoder* $p_\theta(x|z)$ conditioned on z.



Objective Function: Evidence Lowerbound (ELBO)

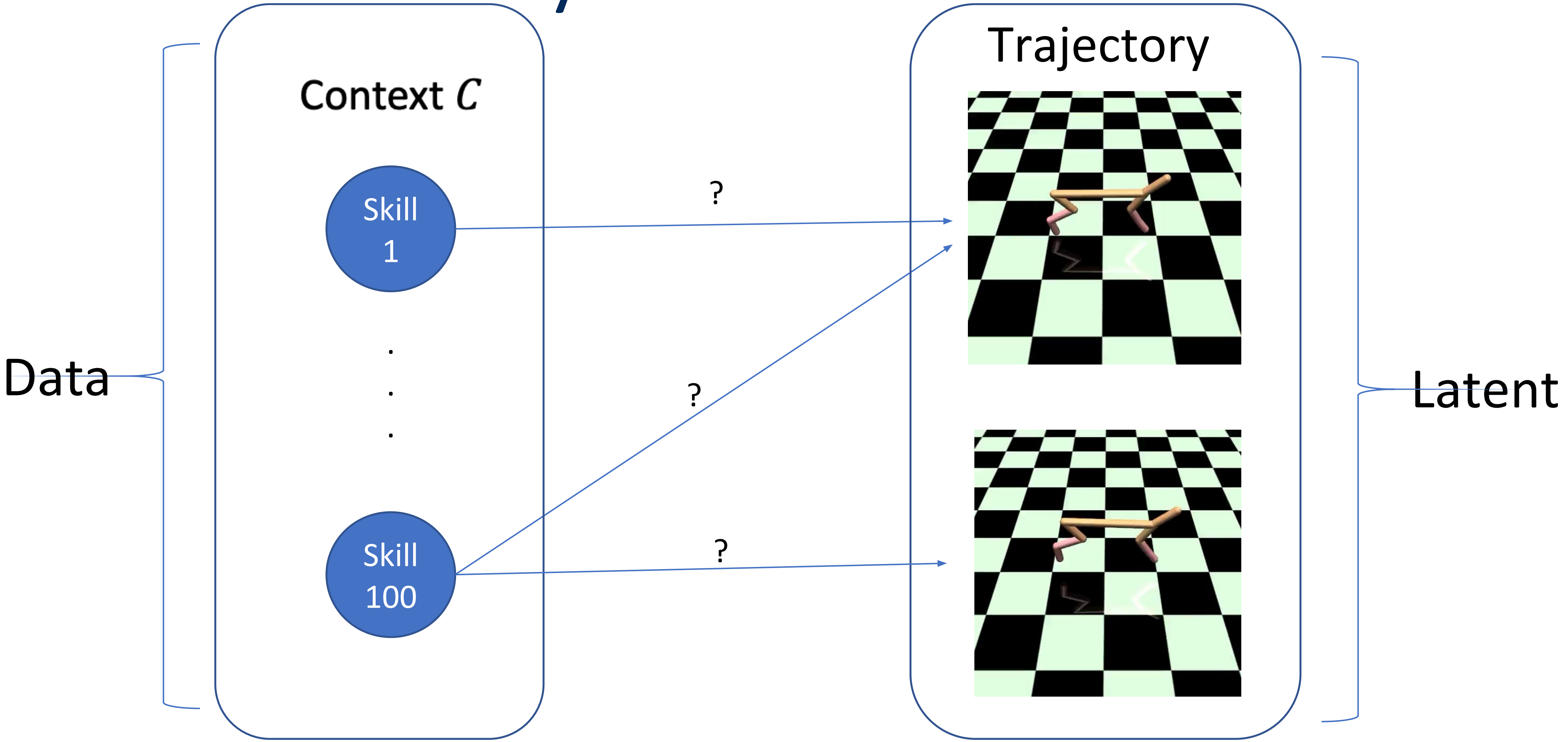
$$\max_{\phi,\theta} \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{z\sim q_\phi(\cdot|x)}[\log p_\theta(x|z)] - \beta D_{KL}(q_\phi(z|x)||p(z))\right]$$

Prior
$p(z)$

Overview

- **Motivation**: Reward-free option discovery
- **Contributions**
- **Background**: Universal Policies, Variational Autoencoder
- **Method**: Variational Option Discovery Algorithms, VALOR, Curriculum
- **Results**
- **Discussions & Limitations**

Intuition: Why VAE + Universal Policies?



Variational Option Discovery Algorithms (VODA)

-

Aim: Learn ***universal policy*** $\boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{s}, \boldsymbol{c})$ such that a **decoder** $\boldsymbol{p_D}(\boldsymbol{c}|\boldsymbol{\tau})$ conditioned on trajectory $\tau = (s_0, a_0, s_1, a_1, \dots, s_T)$ can *distinguish* contexts

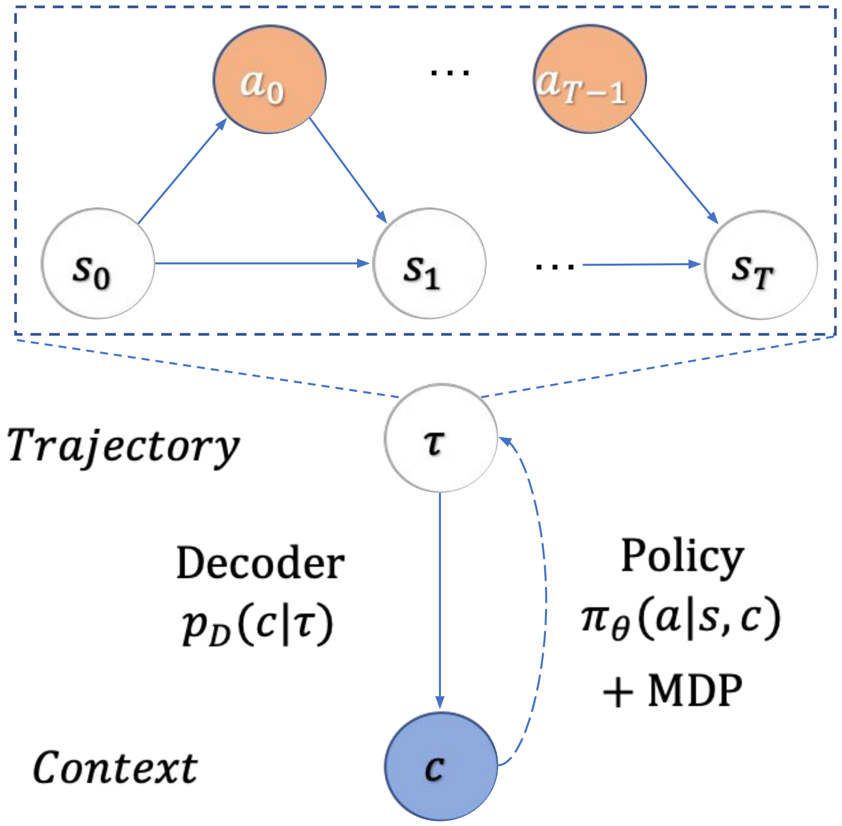
Objective Function:

$$\max_{\pi, D} \mathbb{E}_{c\sim G}\Big[\mathbb{E}_{\tau\sim\pi,c}\,\overbrace{[\log p_D(c|\tau)]}^{\textit{Decoder Reconstruction}} + \beta \mathcal{H}\underbrace{(\pi|c)}_{}\Big]$$

$$\mathcal{H}(\pi|c) \equiv \mathbb{E}_{\tau\sim\pi,c}\left[\textstyle\sum_t H(\pi(\cdot\,|s_t, c)\right]$$

Entropy Regularization

Variational Option Discovery Algorithms (VODA)



Algorithm:

1. Sample context $c \sim G$
2. Roll out trajectory $\tau \sim \pi_\theta(\cdot \mid \cdot, c)$

Create dataset $\mathcal{D} = \{c^i, \tau^i\}_{i=1,..,N}$

3. Update policy via RL to maximize:

$$\max_{\pi,D} \mathbb{E}_{c\sim G}\left[\mathbb{E}_{\tau\sim\pi,c}\left[\log p_D(c|\tau)\right] + \beta\mathcal{H}(\pi|c)\right]$$

4. Update decoder with supervised learning

$$\max_{D} \mathbb{E}_{c,\tau\sim\mathcal{D}}[\log p_D(c|\tau)]$$

Variational Option Discovery Algorithms (VODA)

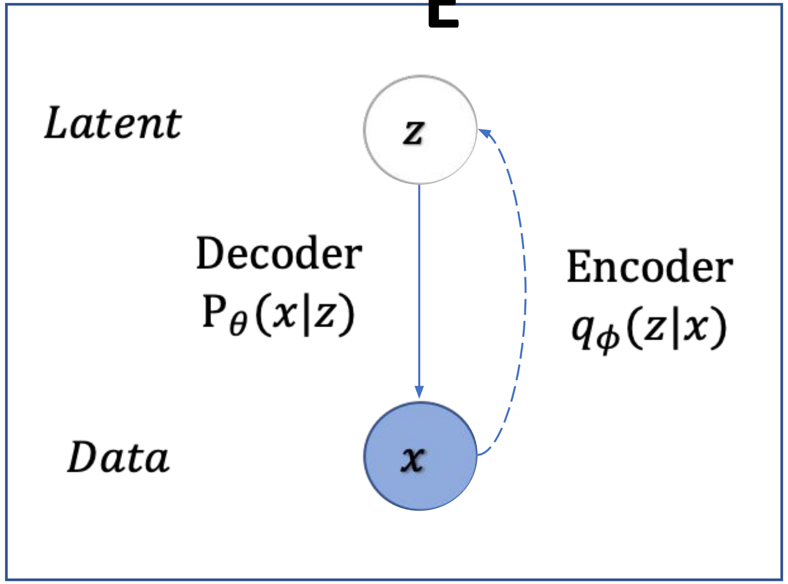
$$\max_{\pi,D} \mathbb{E}_{c\sim G}\left[\mathbb{E}_{\tau\sim\pi,c}\left[\log p_D(c|\tau)\right] + \beta\mathcal{H}(\pi|c)\right]$$

Algorithm 1 Template for Variational Option Discovery with Autoencoding Objective

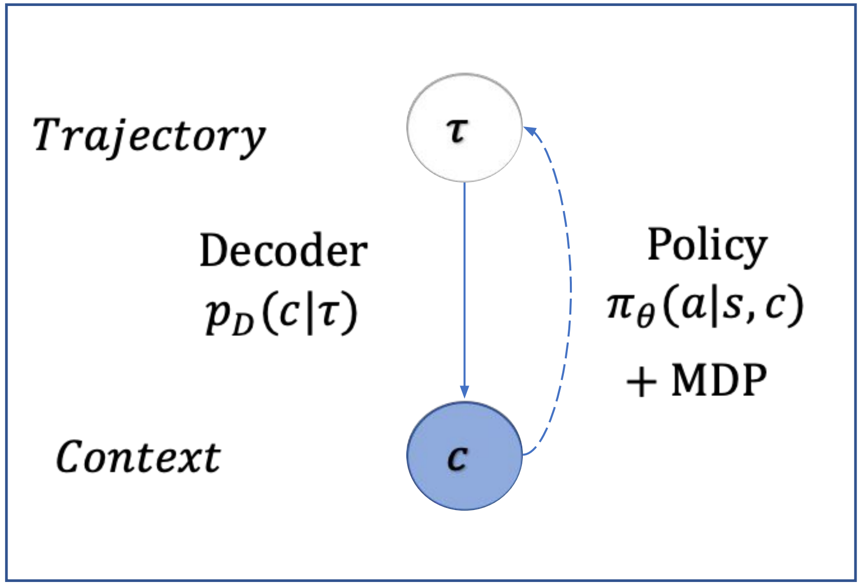
Generate initial policy π_{θ_0}, decoder D_{ϕ_0}
for $k = 0, 1, 2, ...$ **do**
  Sample context-trajectory pairs $\mathcal{D} = \{(c^i, \tau^i)\}_{i=1,...,N}$, by first sampling a context $c \sim G$ and then rolling out a trajectory in the environment, $\tau \sim \pi_{\theta_k}(\cdot|\cdot, c)$.
  Update policy with any reinforcement learning algorithm to maximize Eq. $\boxed{2}$, using batch $\mathcal{D}$
  Update decoder by supervised learning to maximize $\mathrm{E}\left[\log P_D(c|\tau)\right]$, using batch $\mathcal{D}$
end for

VAE vs VODA

VAE

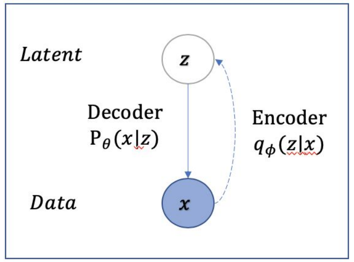


VODA

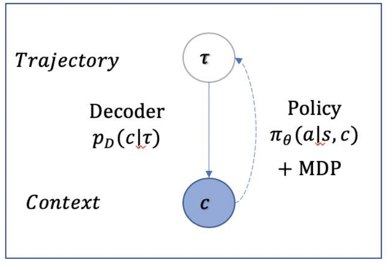


VAE vs VODA

VAE



VODA



$$\boxed{\max_{\phi,\theta} \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\log p_\theta(x|z)\right] - \beta D_{KL}(q_\phi(z|x)||p(z))\right]}$$

$$\boxed{\max_{\pi,D} \mathbb{E}_{c\sim G}\left[\mathbb{E}_{\tau\sim\pi,c}\left[\log p_D(c|\tau)\right] + \beta\mathcal{H}(\pi|c)\right]}$$

VAE			VALOR
Data x	x	⟷ c	Context c
Latent z	z	⟷ τ	Trajectory τ
Encoder $q_\phi(z|x)$		⟷	Policy $\pi(a|s,c)$ and MDP
"Reconstruction" Decoder $p_\theta(x|z)$		⟷	Decoder $p_D(c|\tau)$
$-D_{KL}(q_\phi(z|x)||p(z))$		⟷	Entropy Regularization $\mathcal{H}(\pi|c)$

"KL on prior"

How?

VAE vs VODA: Equivalence Proof

π_0 = Uniform random action policy

$$D_{KL}(q_\phi(z|x)||p(z)) \longleftrightarrow D_{KL}(P(\tau|\pi,c)||p(\tau|\pi_0)) \longrightarrow -\mathcal{H}(\pi|c)$$

Entropy Regularization

Constant Independent of π

Connection to existing works: VIC

Variational Intrinsic Controls (VIC):

$$\max_{G,\pi,D} \mathbb{E}_{s_0\sim\mu}\left[\mathbb{E}_{\substack{\tau\sim\pi,c\\ c\sim G(\cdot|s_0)}}[\log p_D(c|s_0, s_T)] + H(G(\cdot\,|s_0))\right]$$

1. Can optimizes G (But **not** done in experiments)

2. Context depends on initial state s_0

3. Decoder only sees **first** and **last** state

4. Entropy regularization on G, **no** policy entropy regularization

(VODA)

$$\max_{\pi,D} \mathbb{E}_{c\sim G}\left[\mathbb{E}_{\tau\sim\pi,c}[\log p_D(c|\tau)] + \beta\mathcal{H}(\pi|c)\right]$$

Connection to existing works: DIAYN

Diversity Is All You Need (DIAYN):

$$\max_{\pi,D} \mathbb{E}_{c\sim G}\left[\mathbb{E}_{\tau\sim\pi,c}\left[\sum_{t=0}^{T}(\log P_D(c|s_t) - \log G(c))\right] + \beta\mathcal{H}(\pi|c)\right]$$

1. Factorizes probability:

$$\log p_D(c|\tau) = \sum_{t=0}^{T} \log P_D(c|s_t)$$

2. G is **fixed** so can ignore this term

(VODA)

$$\max_{\pi,D} \mathbb{E}_{c\sim G}\left[\mathbb{E}_{\tau\sim\pi,c}\left[\log p_D(c|\tau)\right] + \beta\mathcal{H}(\pi|c)\right]$$

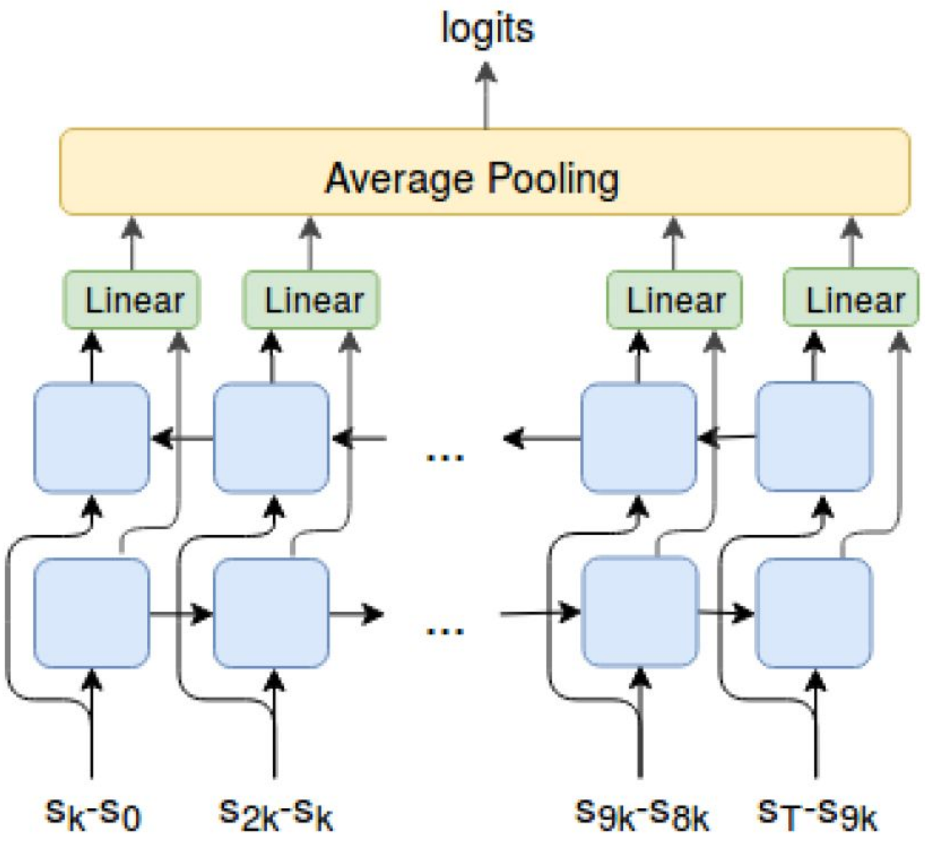
VALOR: Variational Autoencoding Learning of Options by Reinforcement

- **Decoder Architecture**: Bi-LSTM

1. Only sees **states**
2. Not just average of per time-step computation (i.e. DIAYN)

$$\log P_D(c|\tau) \neq \log \sum_{t=0}^{T} f(s_t, c)$$

3. Every **K=11** states



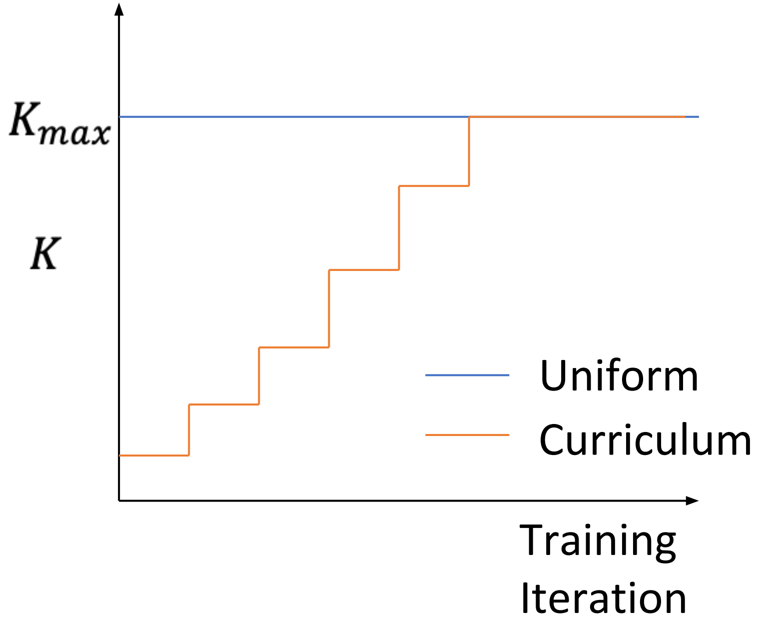
Curriculum on Contexts

- Standard approach (VIC, DIAYN): **Uniform**
  - sample discrete contexts with uniform distribution

$$c \sim \text{Uniform}(K_{max})$$

- Proposed **Curriculum**:
  - When $\mathbb{E}[\log P_D(c|\tau)] \approx 0.86$,

$$K \leftarrow \min(\text{ int }(1.5 \times K + 1), K_{max})$$



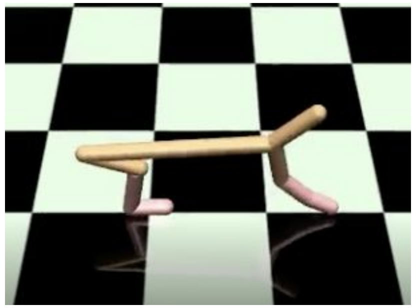
Overview

- **Motivation**: Reward-free option discovery
- **Contributions**
- **Background**: Universal Policies, Variational Autoencoder
- **Method**: Variational Option Discovery Algorithms, VALOR, Curriculum
- **Results**
- **Discussions & Limitations**

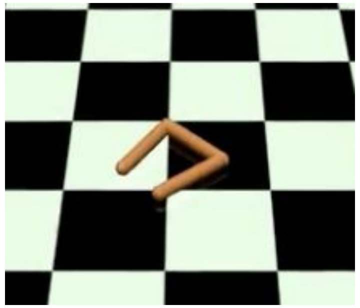
Experiments

1. What are the best practices when training VODAs?
    1. Does the **curriculum learning approach** help?
    2. Does **embedding** the discrete context help vs. **one-hot vector**?
2. What are the **qualitative results** from running VODA?
    1. Are the learned behaviors recognizably distinct to a human?
    2. Are there substantial differences between algorithms?
3. Are the learned behaviors useful for **downstream control tasks**?

Environments: Locomotion environments



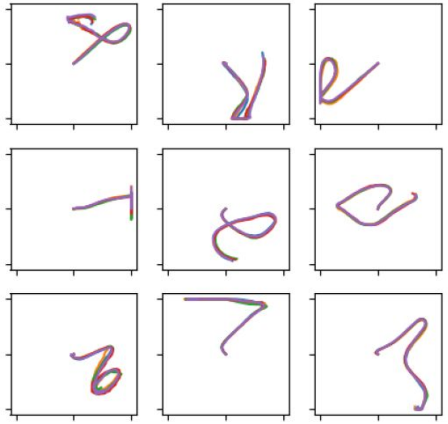
HalfCheetah



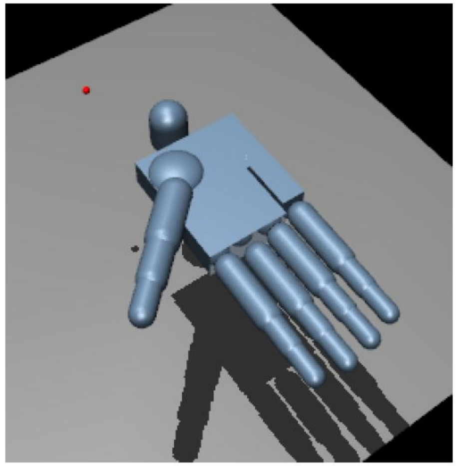
Swimmer

Ant

Note: State is given as vectors, not raw pixels

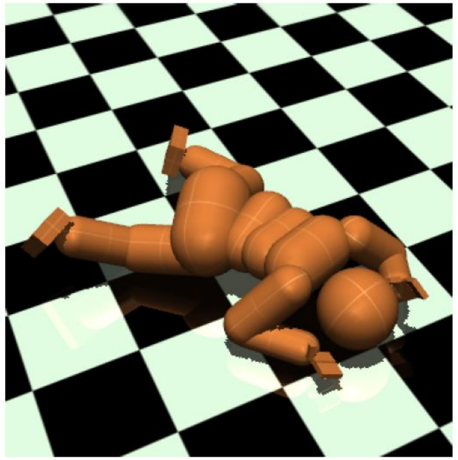


(a) X-Y traces of example modes in Point.



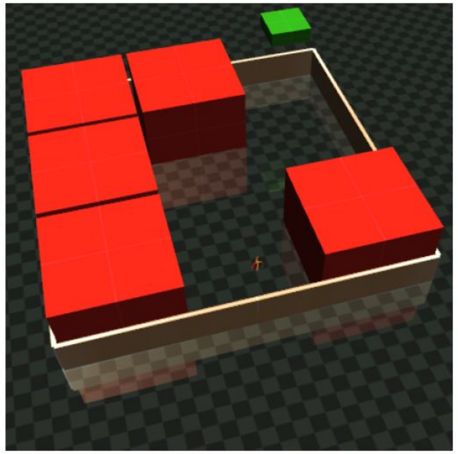
(b) Robot hand environment.

$\mathcal{S} \in \mathbb{R}^{48}, \mathcal{A} \in \mathbb{R}^{20}$



(c) Toddler environment.

$\mathcal{S} \in \mathbb{R}^{335}, \mathcal{A} \in \mathbb{R}^{35}$



(d) Ant-Maze environment.

Implementation Details (Brief)

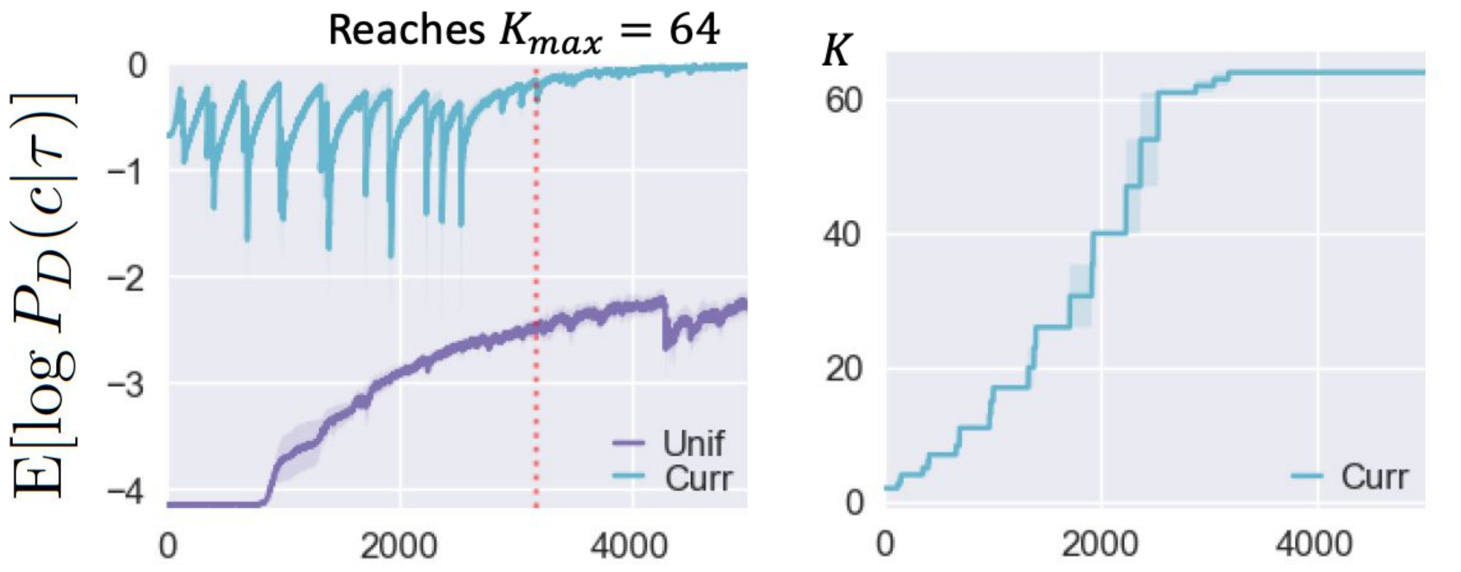
- **Policy** $\pi(a|s,c)$: $LSTM(64)$ then $MLP(32)$ with tanh activations
- **Decoder** $p_D(c|\tau)$:
  - VALOR: Bidirectional LSTM with hidden size 64 for each direction
  - VIC, DIAYN: MLP with hidden size $(180, 180)$
- **Embedding context**: size 32, $\beta = 0.001$
- **Policy Optimization:** vanilla Policy Gradient, and approx. entropy reg.

$$\nabla_\theta J(\pi_\theta) = \underset{\substack{c \sim G \\ \tau \sim \pi, c}}{\mathrm{E}} \left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t, c) \hat{A}_t \right] \quad , \quad \nabla_\theta \mathcal{H}(\pi, c) \approx \sum_{t=0}^{T-1} \underset{s_t \sim \pi, c}{\mathrm{E}} \left[\nabla_\theta H(\pi(\cdot|s_t, c)) \right]$$

Curriculum learning on contexts does help

- **Env:** HalfCheetah

- Using curriculum allows the agent to master $K_{max} = 64$ contexts faster than sampling uniformly



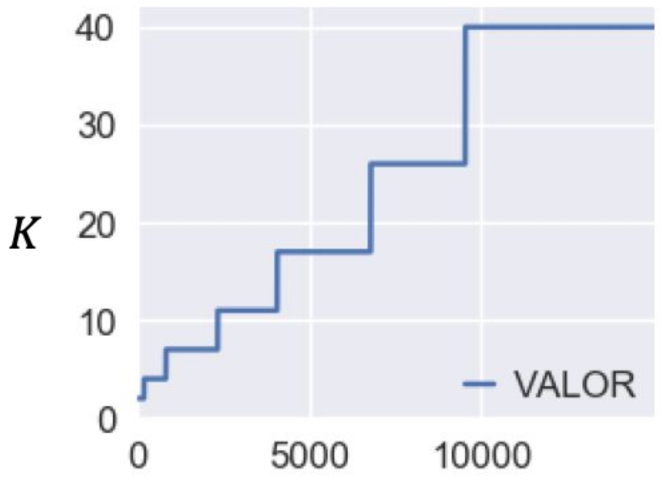
(b) Uniform vs Curriculum

(c) Curriculum, current K

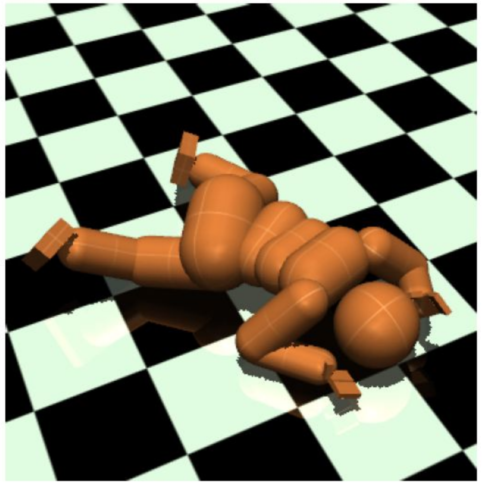
…But struggle in high dimensional environment

Env: Toddler

- After 15K iterations, only $K = 40$ behaviours have been learned



(g) Toddler, current K.

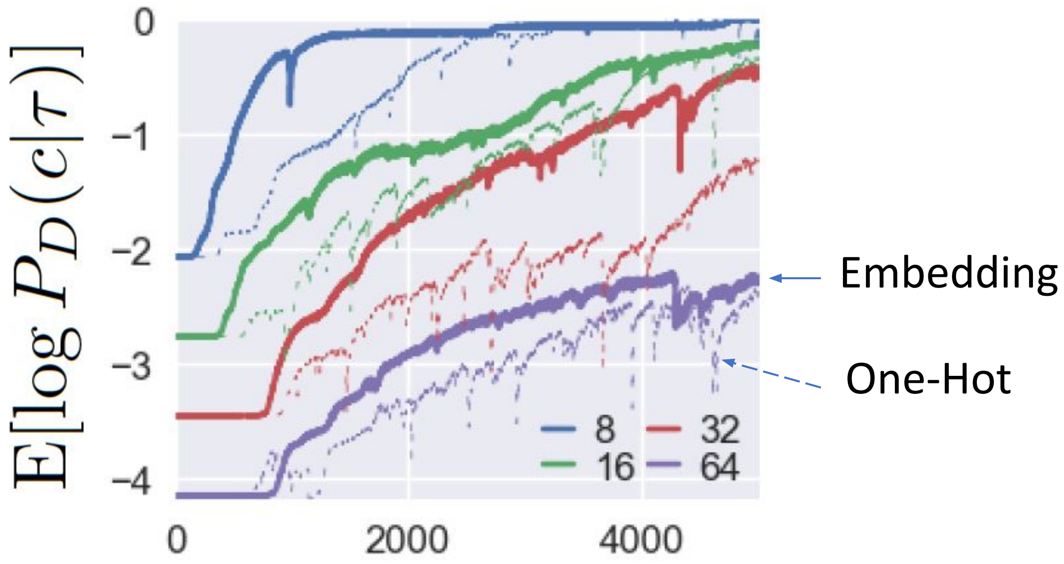


(c) Toddler environment.

$\mathcal{S} \in \mathbb{R}^{335}, \mathcal{A} \in \mathbb{R}^{35}$

Embedding context is better than one-hot

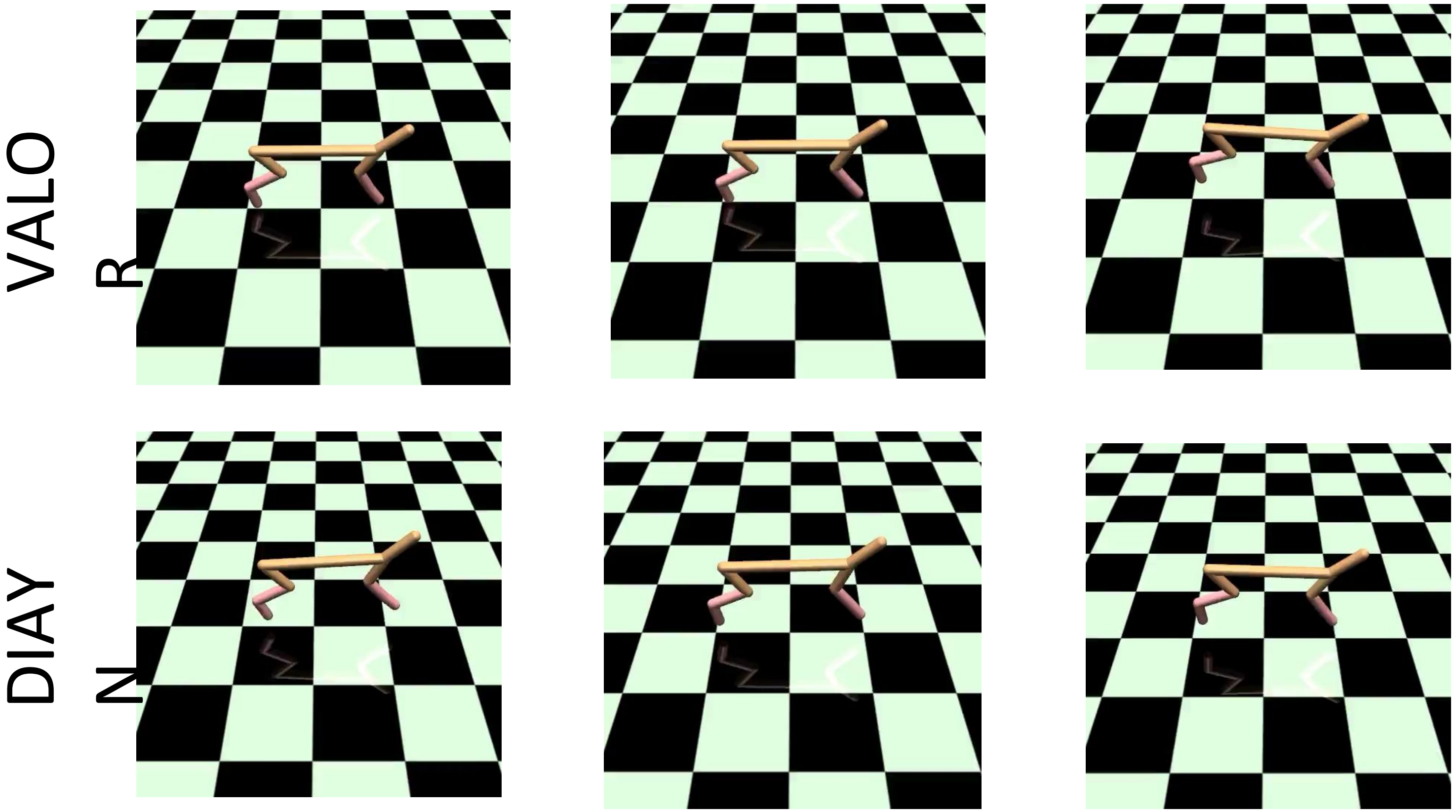
- **Env:** HalfCheetah



(a) Uniform, for various K

Qualitatively learns some interesting behaviors

- VALOR/VIC able to find locomotion gaits that travel in variety of speeds/directions
- DIAYN learns behaviours that 'attain target state' (fixed/unmoving target state)
  - **Note:** Original DIAYN use SAC

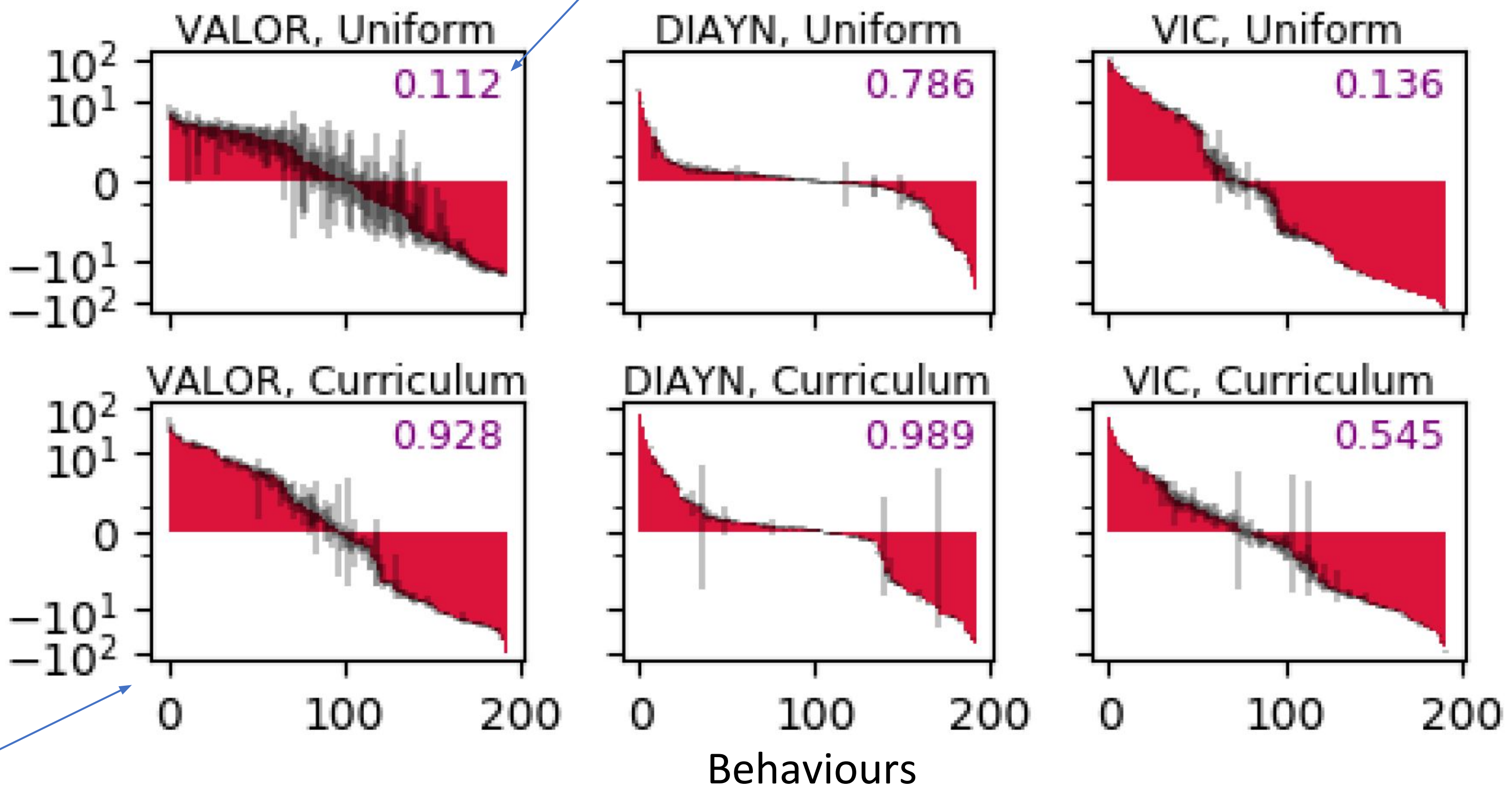


Source: https://varoptdisc.github.io/

Qualitative results (Quantified)

- Each vertical bar:
  - Average score for 5 trajectories 1 context
  - 192 Behaviours from 64 contexts × 3 seeds
- VALOR & VIC finds policies that can move non-trivial amount in HalfCheetah; Many DIAYN behaviours do not move the agent
- Curriculum sometimes increase range

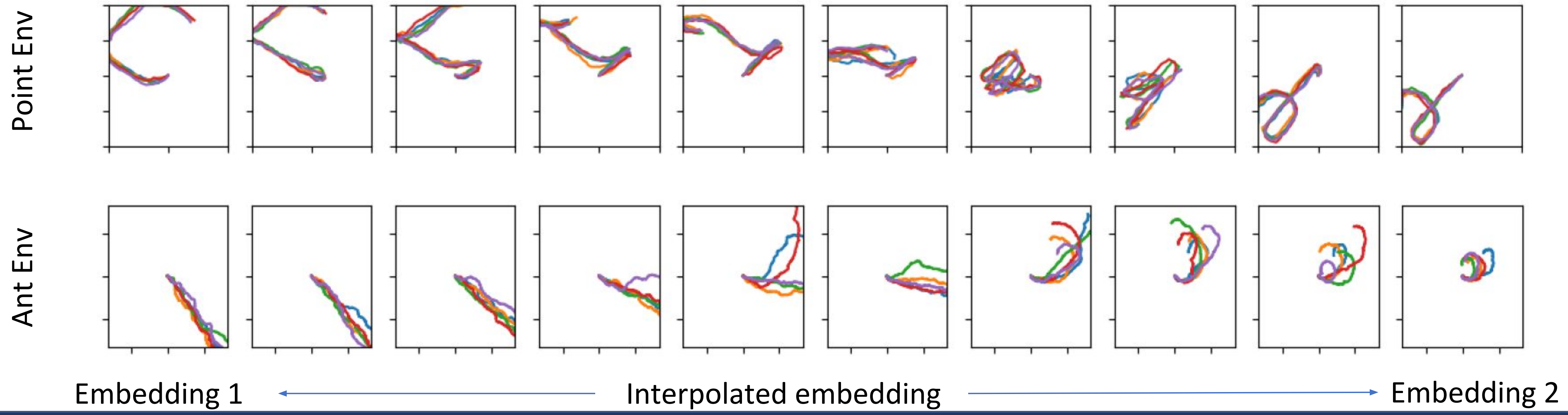
$$\mathrm{E}[P_D(\tau|c)]$$



(a) Final x-coordinate in Cheetah.

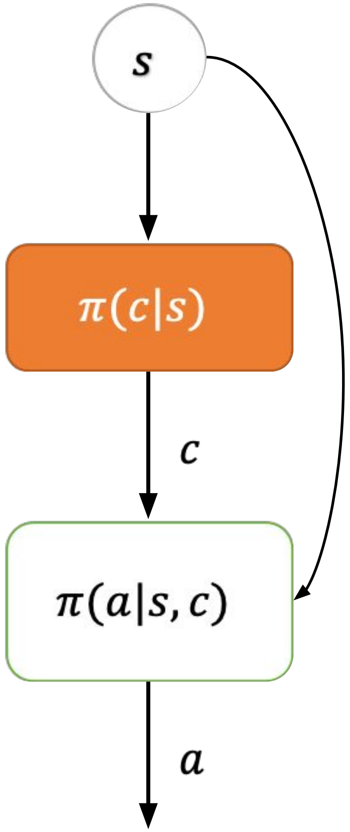
Can somewhat interpolate behaviours

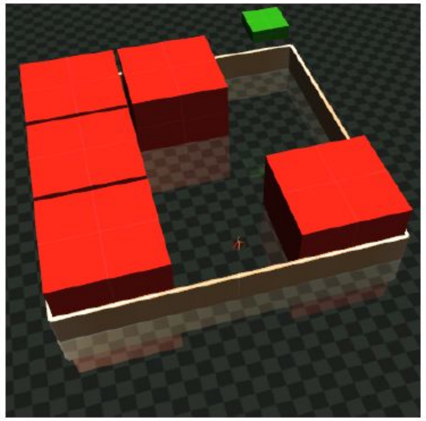
- Interpolating between context embeddings yields reasonably smooth behaviours
- X-Y Traces for behaviours learned by VALOR



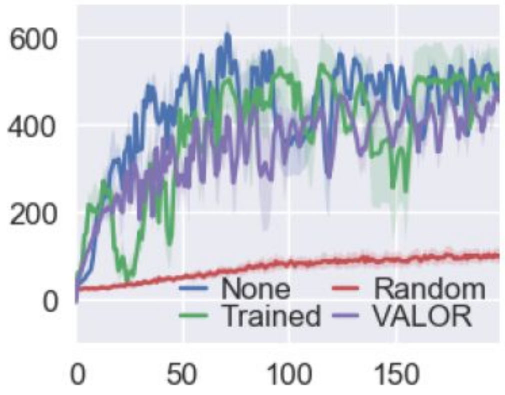
Experiment: Downstream tasks on Ant-Maze

- Take frozen policy trained with VALOR as lower level agent
- Train upper level policy $\pi(c|s)$ using A2C
- Performed similarly to:
    - Training both from scratch
    - No lower level $\pi(a|s)$
- Fixed random network for policy as lower level performs poorly





(d) Ant-Maze environment.



(h) Ant-Maze return.

Overview

- **Motivation**: Reward-free option discovery
- **Contributions**
- **Background**: Universal Policies, Variational Autoencoder
- **Method**: Variational Option Discovery Algorithms, VALOR, Curriculum
- **Results**
- **Discussions & Limitations**

Discussion and Limitations

- Learned behaviours are *unnatural*
  - Due to using purely information theoretic approach?
- Struggle in high dimensional environments (e.g. Toddler)
- Need better performance metrics for evaluating discovered behaviours
- Hierarchies built on top of learned contexts do not outperform task-specific policies learned from scratch
  - But at least universal enough to be able to adapt to more complex tasks
- Specific curriculum on context equation seems unprincipled/hacky

Follow Up Works

- (ICLR'20) **Dynamics-Aware Unsupervised Discovery of Skills (DADS)**: unsupervised discovery of skills and incorporated into model-based planning

$$\max\ MI(s'; c|s)$$

- **State Marginal Matching with Mixtures of Policies:** Learns to maximize the entropy in the visited states when marginalized out context. Includes entropy of states condition on context

$$H(s|c) = \mathbb{E}_{c\sim G,\, s\sim\pi(\cdot|c)}\left[-\log p(s|c)\right]$$

Future Research Directions

- Fix “unnaturalness” of learned behaviours: incorporate human priors?
  - Distinguish trajectories in ways which corresponds to human intuition
  - Leverage demonstration? Human-in-the-loop feedback?
- Architectures: Use **Transformers** instead of Bi-LSTM for decoder
  - As done in NLP: ELMO (Bi-LSTM) vs BERT (Transformer)

Contributions

1. **Problem:** Reward-free options discovery, which aims to learn interesting behaviours without environment rewards (unsupervised)
2. Introduced a general framework **Variational Option Discovery** objective & algorithm
    1. Connected Variational Option Discovery and Variational Autoencoder (VAE)
3. Specific instantiation: **VALOR** and **Curriculum learning:**
    1. **VALOR:** a decoder architecture using Bi-LSTM over only (some) states in trajectory
    2. **Curriculum learning** for increasing number of skills when agent mastered current skills
4. Empirically tested on simulated robotics environments
    1. VALOR can learn diverse behaviours in variety of environments
    2. Learned policies are universal, can be interpolated and used in hierarchies

References

1. Achiam, et al. Variational Option Discovery Algorithms
2. (VIC) Variational Intrinsic Control
3. (DIAYN) Diversity Is All You Need
4. Rich Sutton's page on Options Discovery