Investigating different exploration strategies for model-based reinforcement learning

Aseem Saxena, Joe Nguyen, Skand

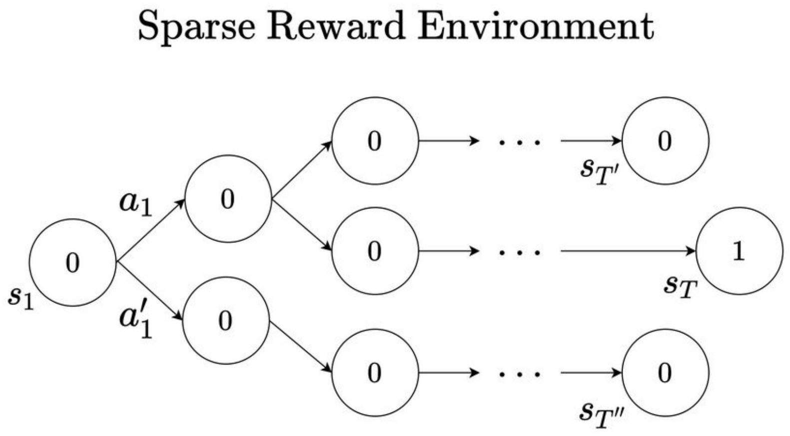
Advance PGM Final Project

Motivation

Agents need to explore states to find the successful paths leading to the target states

In sparse-reward setting, random exploration takes *long* time to find important states

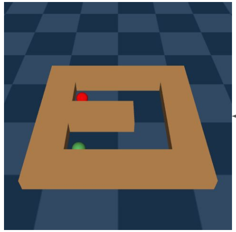
Agents need to explore *intelligently*



Outline

- Model based RL – a primer
- Three *knowledge* based exploration methods
  - Curiosity
  - Disagreement (Plan2Explore)
  - Monte Carlo dropout
- Qualitative results on state space coverage
- Quantitative results on downstream task performance
- Ablation Study
- Limitation and Conclusion

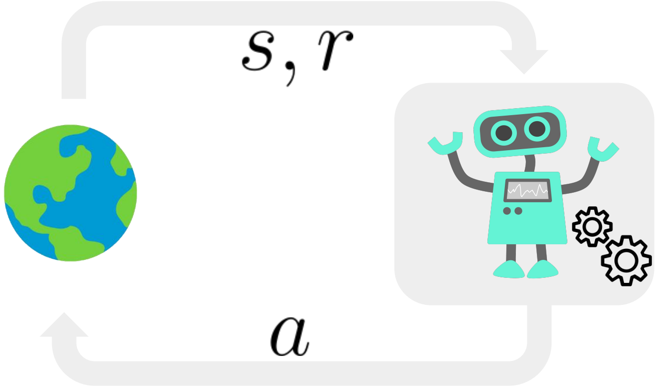
Reinforcement learning



state, reward

<u>Sparse</u>: e.g. whether it reaches the goal or not (0/1)

<u>Dense</u>: e.g. the negative distance between the current position and goal position



s, r

a

$\pi(a|s)$

policy

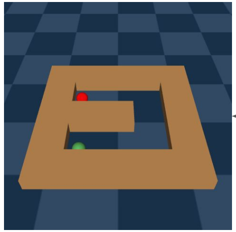
action

Goal: maximize the expected sum of rewards (r)

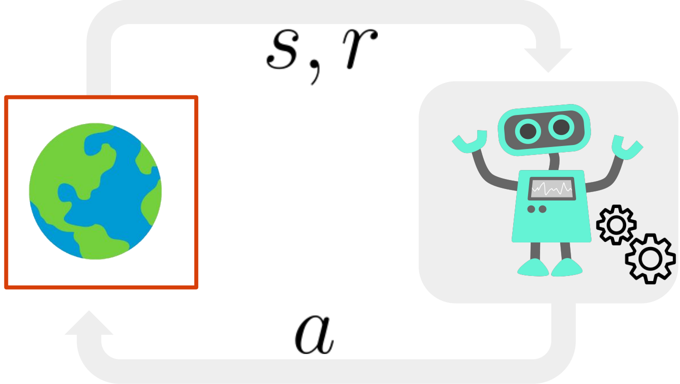
Reinforcement learning: Model-based RL

<u>Sparse</u>: e.g. whether it reaches the goal or not (0/1)

<u>Dense</u>: e.g. the negative distance between the current position and goal position



state, reward



$\pi(a|s)$

policy

action

Interaction with the environment takes time.

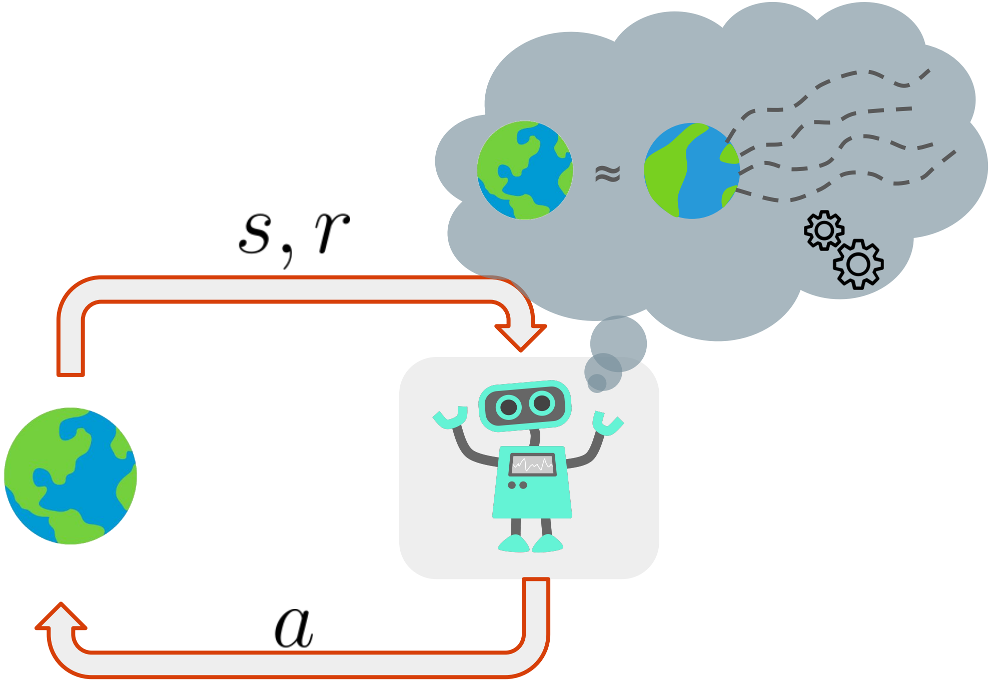
To make data efficient, we simulate the environment by a *world model*, from which we can rollout the trajectory to learn the policy

-> **Model-based RL**

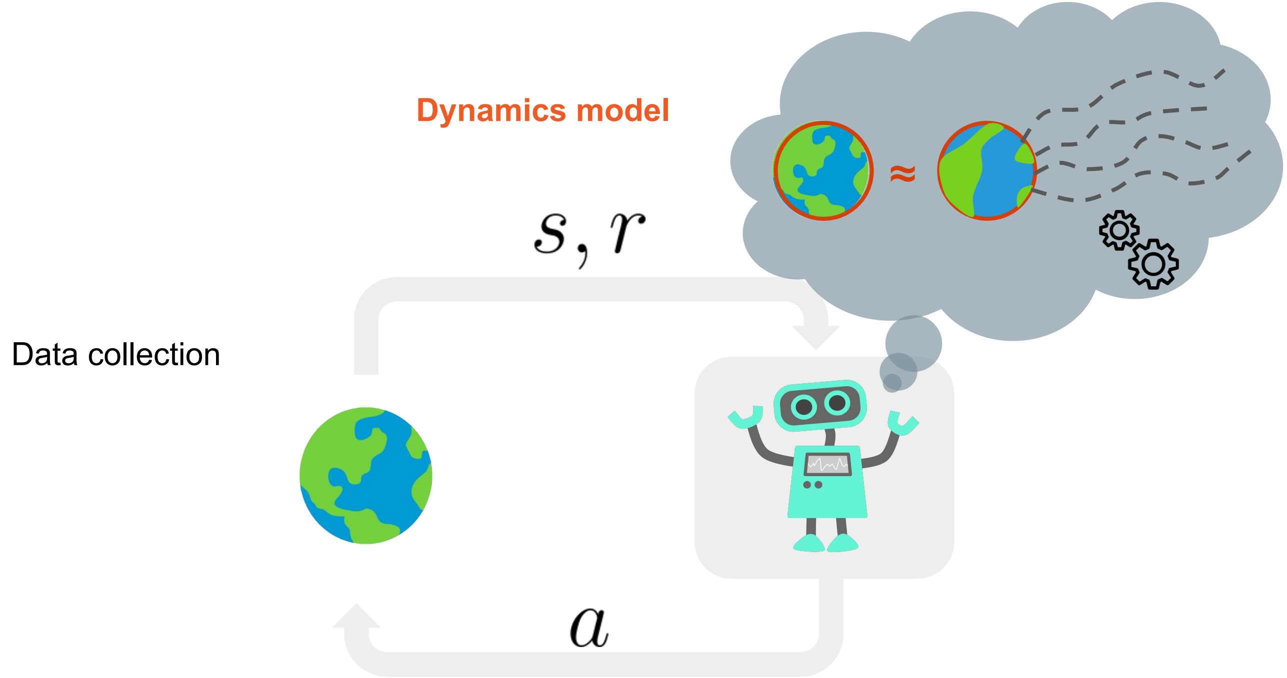
Goal: maximize the expected sum of rewards (r)

Model based RL: Stage one

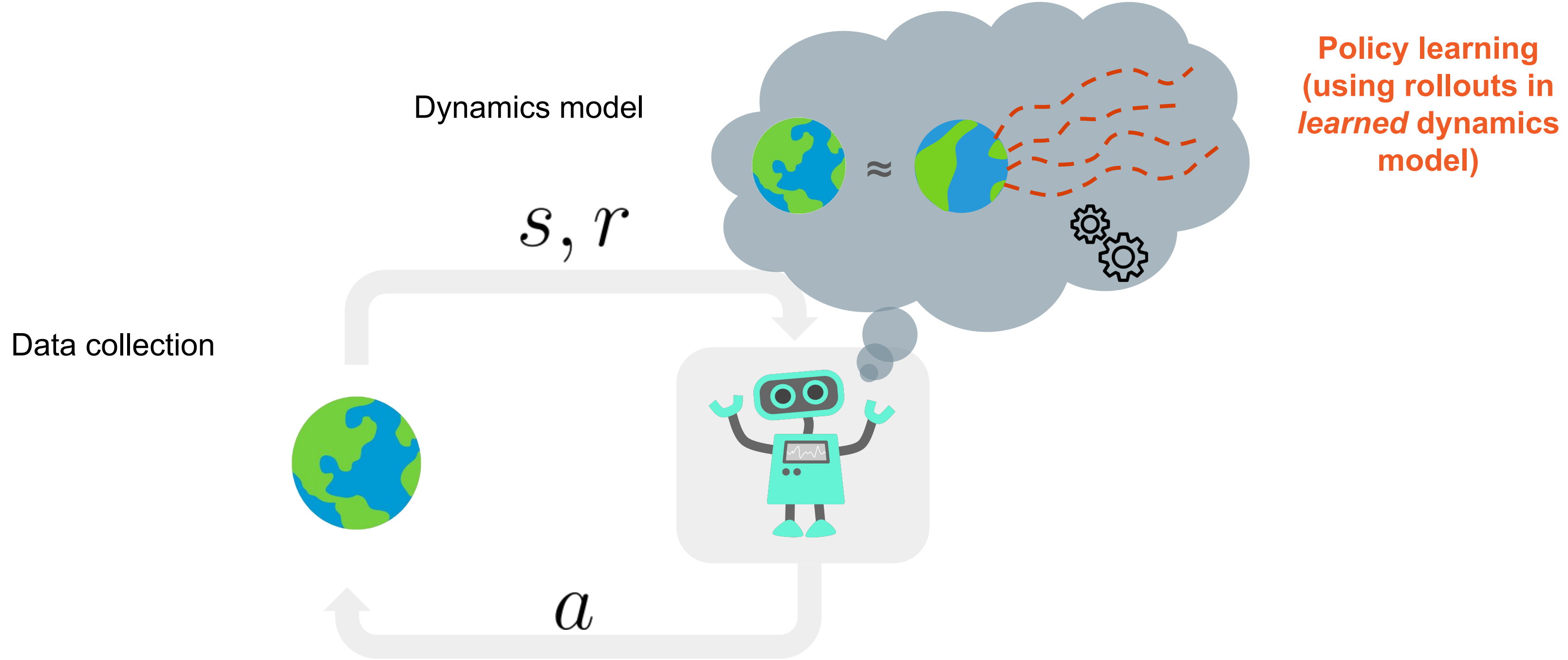
Data collection
(s, a -> s', r)



Model based RL: Stage two



Model based RL: Stage three



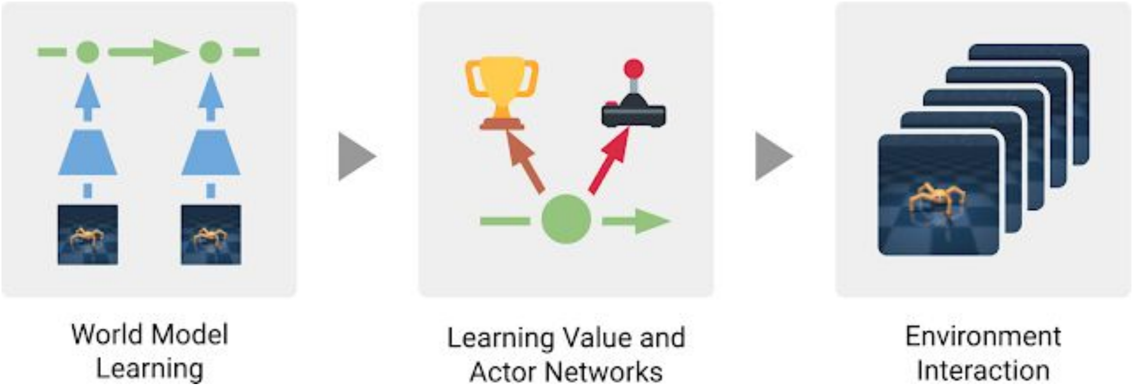
Dynamics model

≈

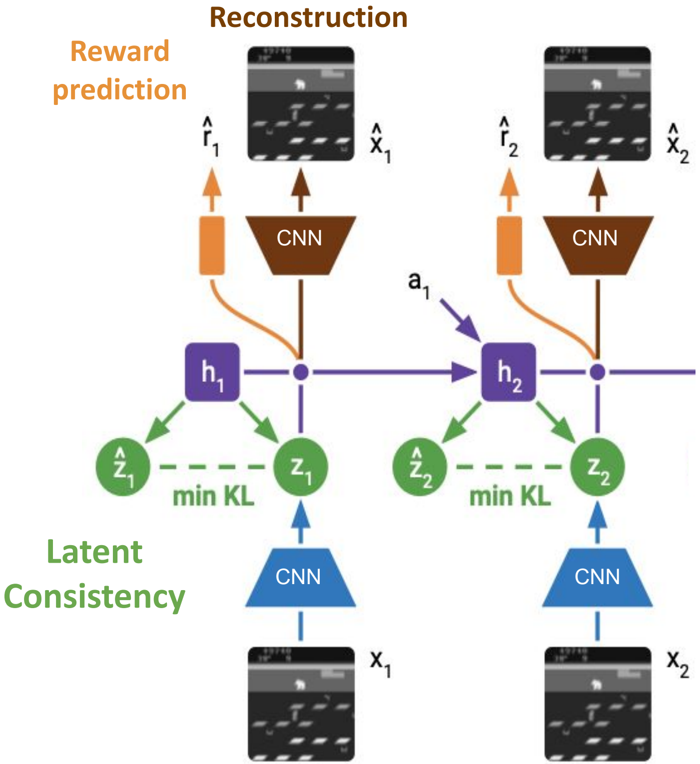
What should the dynamics model do?

- Reward function : `p(r | s)`
- Transition function : `p(s’ | s, a)`

Model based visual RL : Dreamer



Model based visual RL : Dreamer



World Model

Similar to sequential VAE

- **The posterior:** Learn the meaningful latent z from the observation and history by reconstructing the observation

$$z_t \propto p(z_t | x_t, h_t)$$

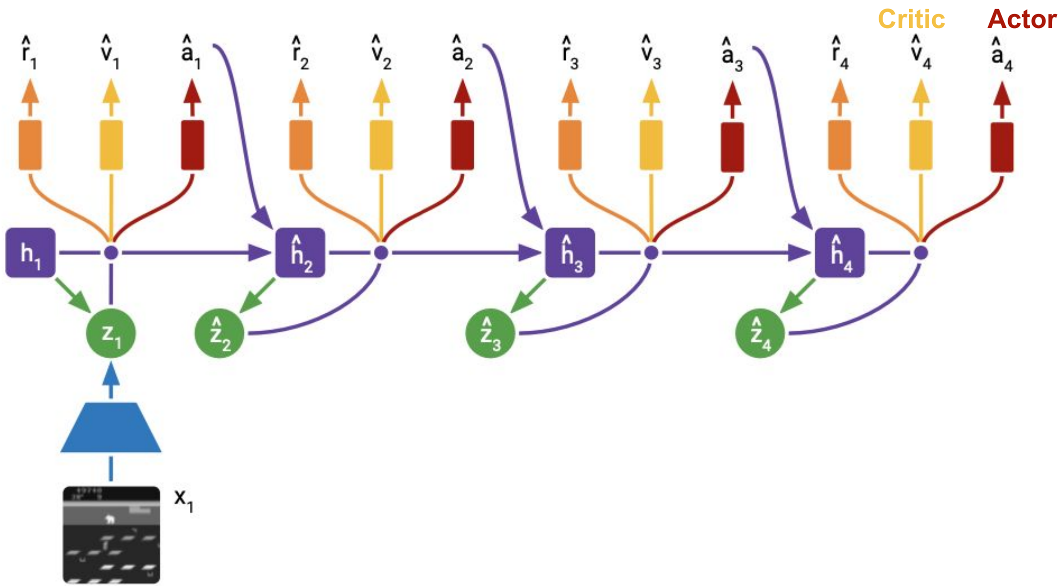
- **The prior:** Learn to predict z^ from the previous timestep (history) without the observation

$$\hat{z}_t \propto p(z_t | h_t)$$

- **Loss** = reconstruction loss + KL_loss(the posterior | the prior)

Mastering Atari with Discrete world models, Danijar Hafner et al. (ICLR 2021)

Model based visual RL : Dreamer



The trajectories start from posterior states computed during model training and predict forward by sampling actions from the actor network and the prior distribution.

Policy Learning

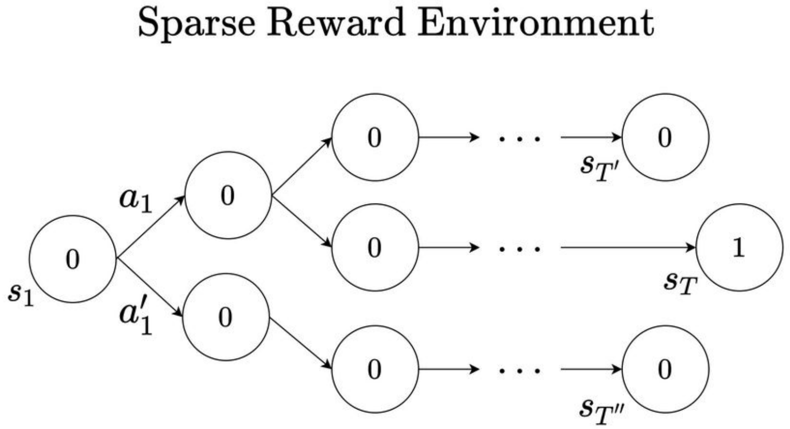
Mastering Atari with Discrete world models, Danijar Hafner et al. (ICLR 2021)

Motivation

Agents need to explore states to find the successful paths leading to the target states

In sparse-reward setting, random exploration takes *long* time to find important states

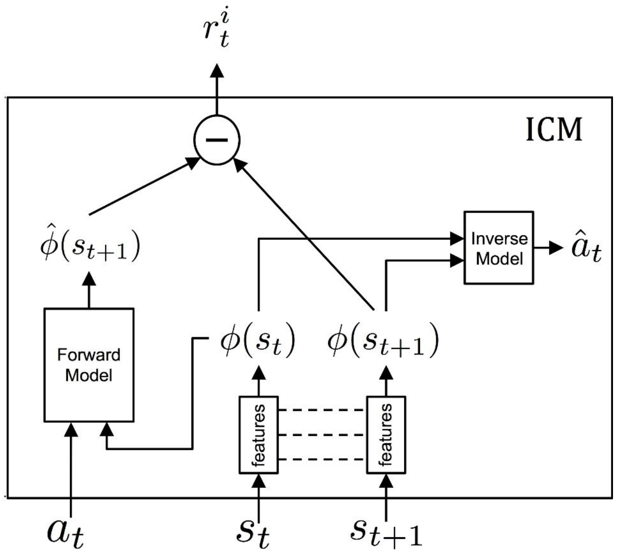
Agents need to explore *intelligently* in the world model



Adaptive Discount Factor for Deep Reinforcement Learning in Continuing Tasks with Uncertainty

Knowledge based exploration – Curiosity

Idea: Quantify exploration reward using the prediction error of the dynamics model

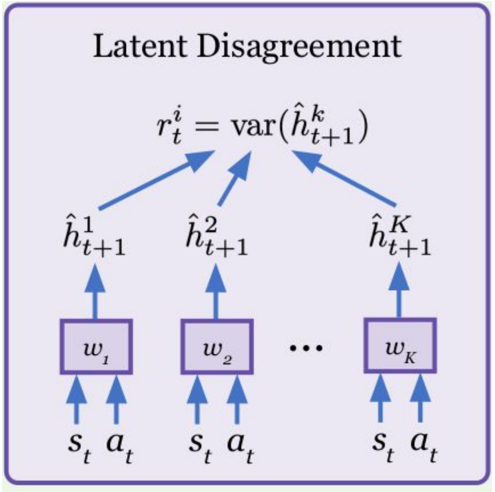


Curiosity-driven Exploration by Self-supervised Prediction, Deepak Pathak et al, ICML 2017

Knowledge based exploration – Plan2Explore

Idea: Compute intrinsic reward as *disagreement* between ensembles

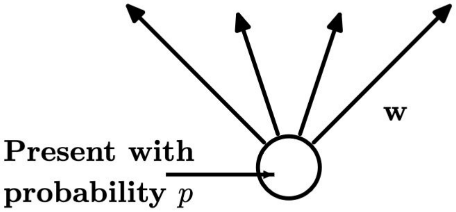
Encourages agents to explore more uncertain regions

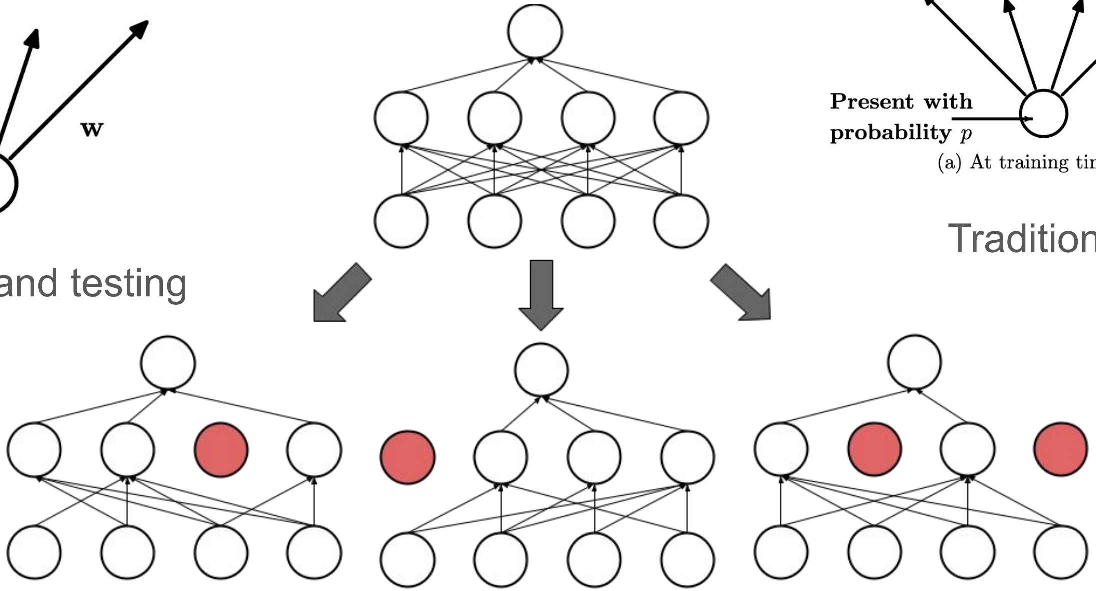


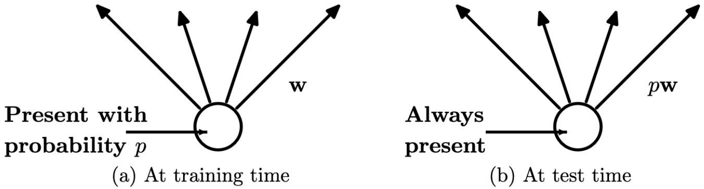
Planning to Explore via Self-Supervised World Models, Ramanan Sekar, Oleh Rybkin* et al., ICML 2020*

Knowledge based exploration – Monte Carlo Dropout

Idea: Dropout during training time with randomly generated ensembles on-the-fly.







Traditional Dropout

Both in training and testing
for MC Dropout

Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, Yarin Gal & Zoubin Ghahramani, ICML 2016

Environments: Point Maze

Task: move the green ball to reach the target red goal in a closed maze

State space: (x, y, vel_x, vel_y)

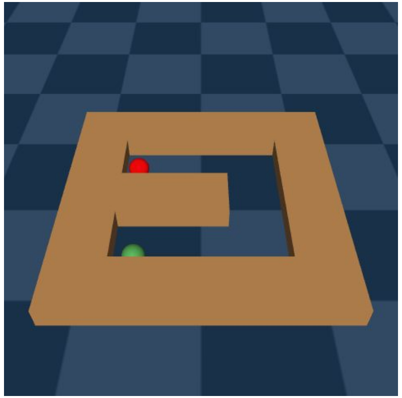
(x, y): coordinate of the green ball

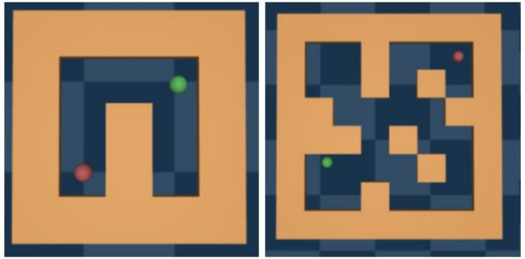
(vel_x, vel_y): linear velocities for each axis

Action space: ($motor_x$, $motor_y$): linear forces exerted on the ball

Reward:

- Sparse: 0 if the ball hasn't reached the target and 1 if it has reached the target (within 0.5m)
- Dense: negative Euclidean distance between the current position and the goal position





(a) Small Maze (b) Medium Maze

(c) Large Maze

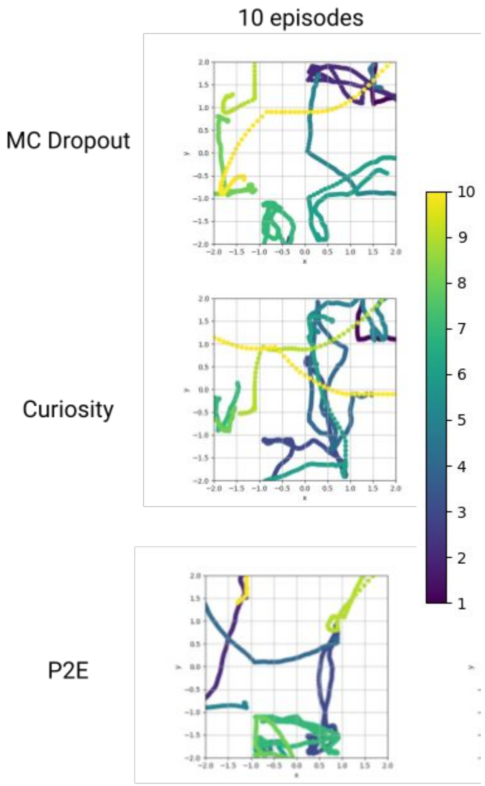
D4RL: Datasets for Deep Data-Driven Reinforcement Learning (Fu et al. 2021)

Experiments

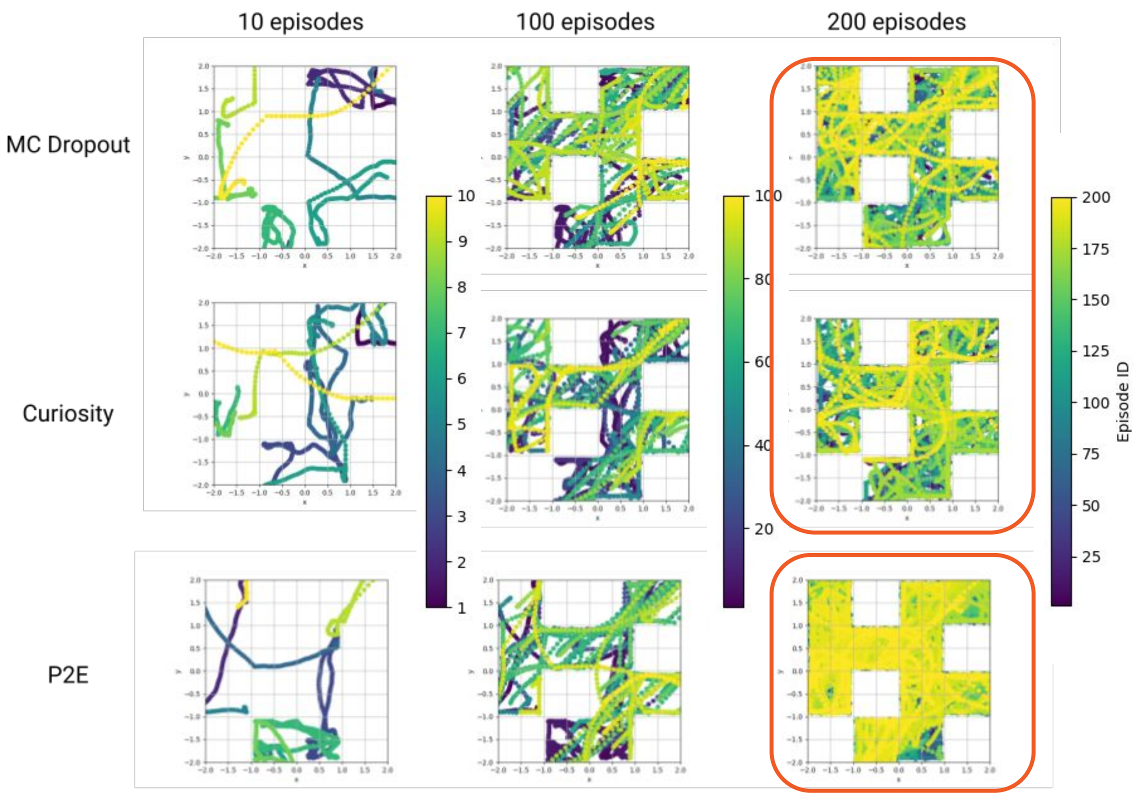
Methods:

1. DreamerV3 + Curiosity
2. DreamerV3 + Plan2Explore
3. DreamerV3 + MC-Dropout

Experiments: Qualitative Results



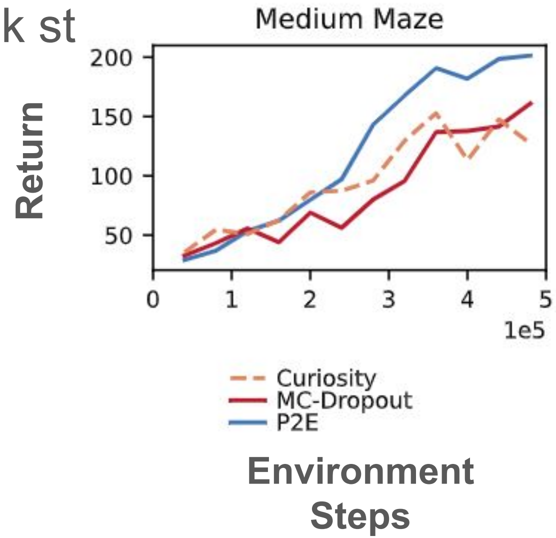
Experiments: Qualitative Results



Experiments: Quantitative Results

Setting

1. First explore for 200k steps (~1000-1200 episodes)
2. Adding task specific reward at 200k mark until 500k st



Observations

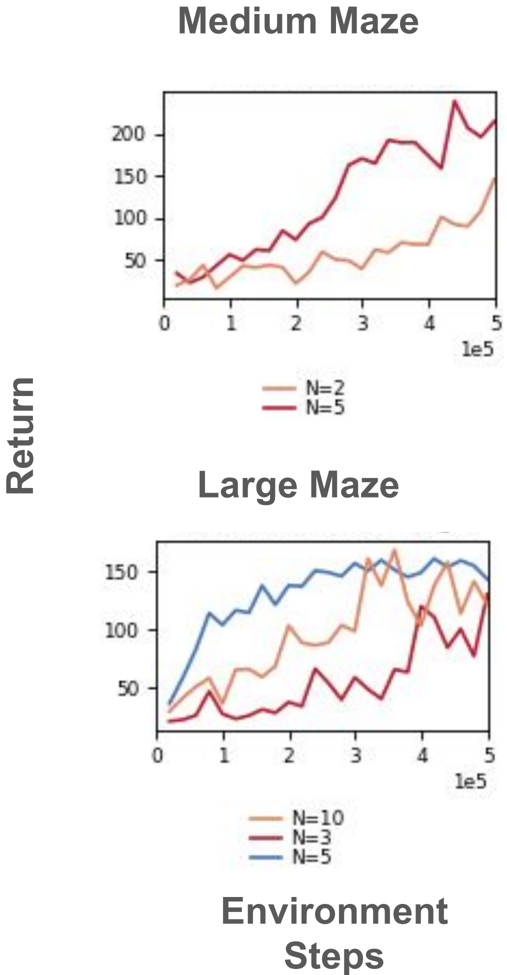
1. Similar to the qualitative results, **P2E performs the best** and is more sample efficient.
2. Interestingly, MC-Dropout and Curiosity based exploration perform very similar!

Experiments: Ablation Study

We ablate the number of ensembles in Plan2Explore.

Observations:

1. We find that typically higher the ensembles the better the exploration is.
    a. Potentially because higher ensembles can better capture the uncertainty in the state.
2. We did not observe any significant differences with increase in number of ensembles > 5.



Limitations

1. We did not consider *competence* based methods such as DIAYN (Eysenbach et al. 2018) and other mutual information methods based on DIAYN formulation and leave that as a future work

2. Do experiments on more complicated environments tasks such as DM Control (Tassa et al. 2018) which have much larger state and action spaces

Summary

1. We investigated different knowledge based exploration methods – namely Plan2Explore, Curiosity and MC-Dropout.

2. We found that Plan2Explore performs the best on PointMaze tasks and interestingly MC-Dropout is on par with Curiosity based exploration

3. Finally, we performed ablation study of the effect of number of ensembles in P2E and showed that there is a “sweet-spot” for the number of ensemble (N).

Questions