EX^2: Exploration with Exemplar Models for Deep Reinforcement Learning

Justin Fu*, John D. Co-Reyes*, Sergey Levine

University of California, Berkeley

* equal contribution

Introduction

Problem: Efficient exploration in high-dimensional spaces. Most methods require building a generative model over the state, such as dynamics ($P(s'|a, s)$) or counting ($P(s, a)$ or $P(s)$).

Approach:

- Train discriminators to classify new states against previously seen states. Easily classifiable states are "novel".
- Augment the reward with a novelty bonus to encourage the policy to visit new states.

Key Insights:

- Our algorithm can be interpreted as approximating count/density-based exploration.
- Our algorithm also resembles a GAN, except the generator (policy) and discriminator are cooperative.

Discriminators and Density Estimation

- We consider classifying an "exemplar" x^* as a positive against negatives $x' \sim P(x)$. Letting $Q(x) = \delta_{x^*}(x)$ denote a delta function around x^*, we optimize a discriminator $D : \mathcal{X} \to [0, 1]$ via a standard cross-entropy loss:

$$D^* = \mathrm{argmax}_D\{E_{x\sim Q(x)}[\log D(x)] - E_{x\sim P(x)}[\log 1 - D(x)]\}$$

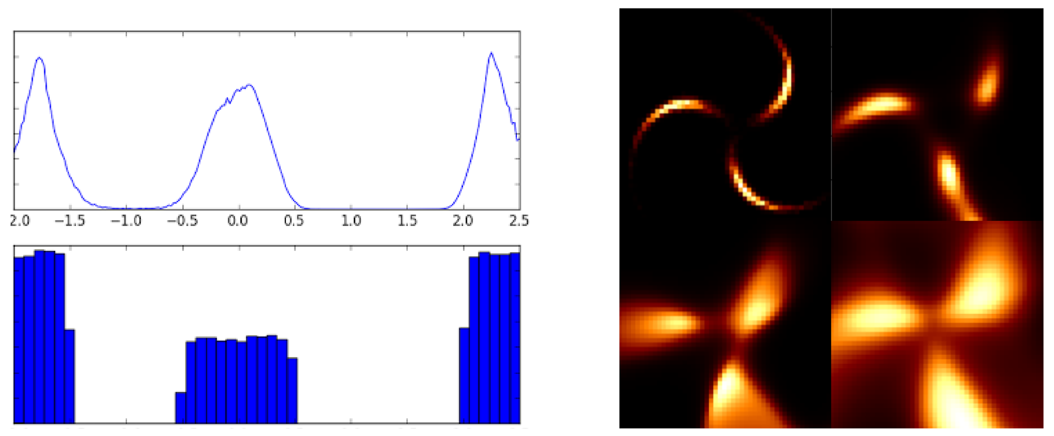
- Since the distribution A^* is known, we can show that when evaluated at $x^* \in X^*$,

$$D^*(x^*) = \frac{1}{1 + P(x^*)}$$

Thus, we can recover density estimates $P(x)$ for x in the positive set $\mathcal{X}^*$ as:

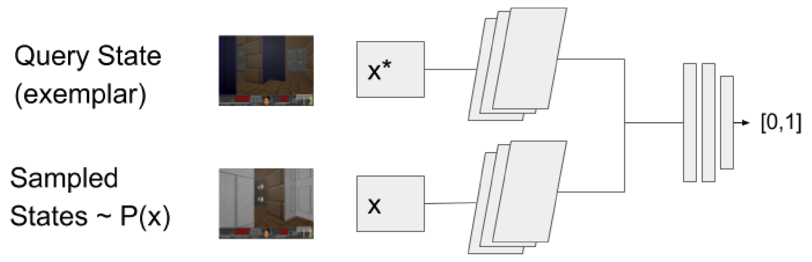
$$P(x) = \frac{1 - D^*(x)}{D^*(x)}$$

- **Continuous distributions**: By adding noise to the exemplar distribution $Q(x)$, we can show an analogous result in the continuous case.
- These results hold for optimal discriminators. We also find that a slightly suboptimal discriminators found in practice, along with injecting noise to the data distribution $P(x)$ will **generalize and smooth density estimates**.
  - Injecting Gaussian noise into the negative distribution $P(x)$ results in a method similar to KDE with Gaussian kernels.



Amortized Exemplar Model

- In practice, training a single discriminator for every state is prohibitively expensive. We instead condition the discriminator on the exemplar x^*, which we refer to as the *amortized exemplar model*.



- This architecture has the appearance of an similarity function (in a "reference equality" sense) - it is trained to output 0 when $x \neq x^*$ and 1 when $x = x^*$. ($=$ denoting "reference" rather than "value" equality)

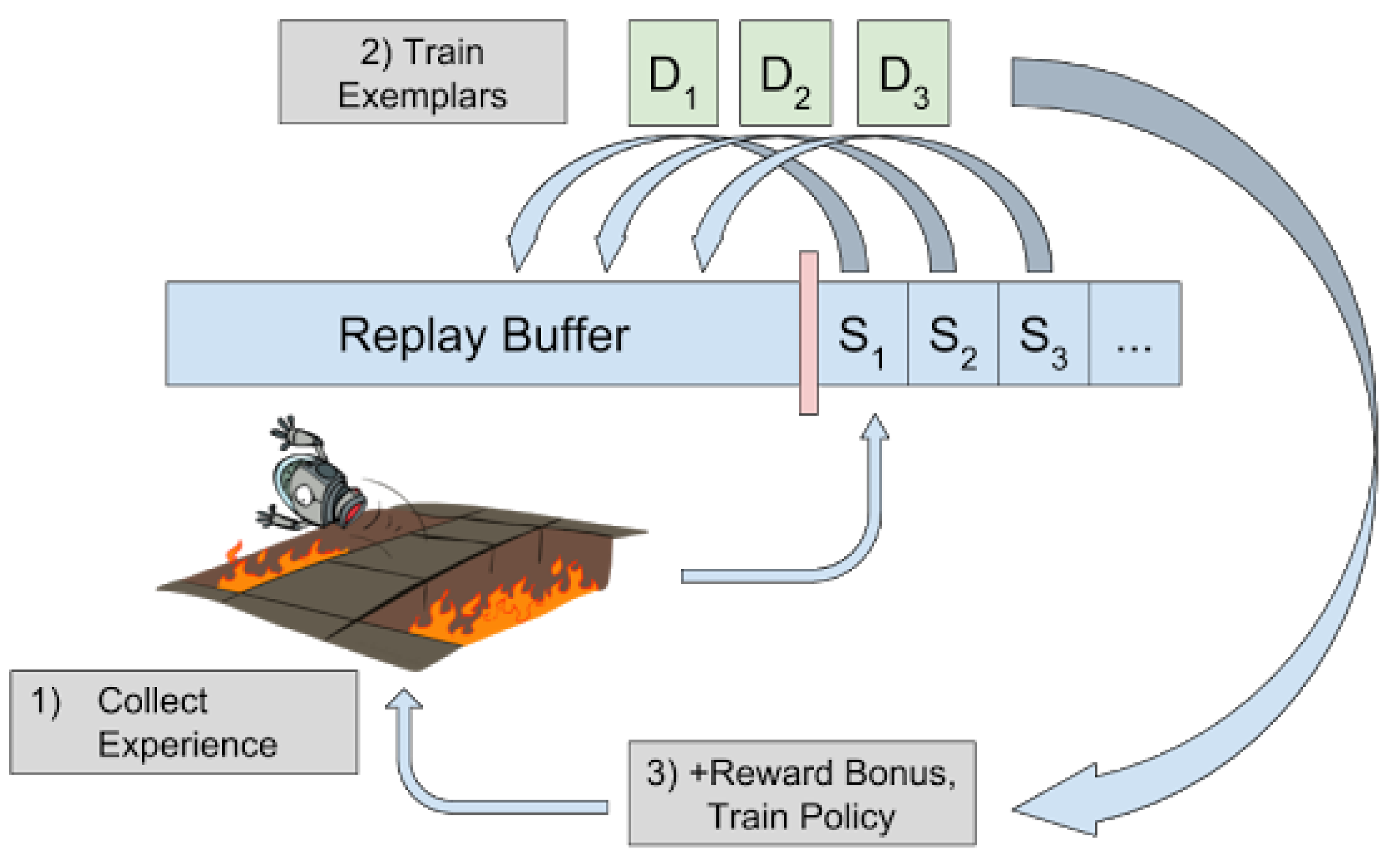
Exploration with Exemplar Models

- We consider the reinforcement learning problem of finding a policy that maximizes expected returns:

$$\pi^* = \mathrm{argmax}_\pi\{E_{\tau\sim\pi}[\sum_{t=0}^{T} R(s_t, a_t)]\}$$

- We adopt the count-based exploration paradigm, and add a reward bonus to states with low $P(s)$, where $P(s)$ is the distribution over all states visited by the algorithm during training.

Algorithm Diagram

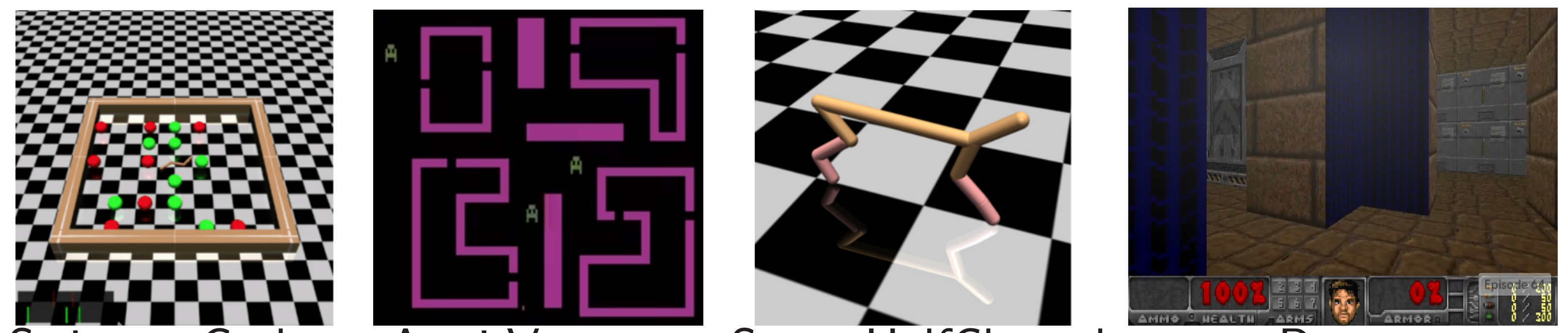


References

[1] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. D. Turck, and P. Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[2] H. Tang, R. Houthooft, D. Foote, A. Stooke, X. Chen, Y. Duan, J. Schulman, F. D. Turck, and P. Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Experimental Results



SwimmerGather Atari-Venture SparseHalfCheetah Doom+

Tasks:

- **SwimmerGather**: A hierarchical task that requires moving a 3-link robot to collect pellets for reward.
- **SparseHalfCheetah**: 6-DoF cheetah needs to move past a specified distance threshold.
- **DoomMyWayHome+**: A sparse, goal based visual navigation task inside a maze.
- **Atari**: Three Atari games (Freeway, Frostbite, Venture) requiring exploration.

Task	K-Ex.(ours)	Amor.(ours)	VIME [1]	TRPO	Hashing [2]	KDE	Histogram
2D Maze	-104.2	-132.2	-135.5	-175.6	-	-117.5	-69.6
SparseHalfCheetah	3.56	173.2	98.0	0	0.5	0	-
SwimmerGather	0.228	0.240	0.196	0	0.258	0.098	-
Freeway (Atari)	-	33.3	-	16.5	33.5	-	-
Frostbite (Atari)	-	4901	-	2869	5214	-	-
Venture (Atari)	-	900	-	121	445	-	-
DoomMyWayHome	0.740	0.788	0.443	0.250	0.331	0.195	-

Table: Median (and mean in parentheses) scores

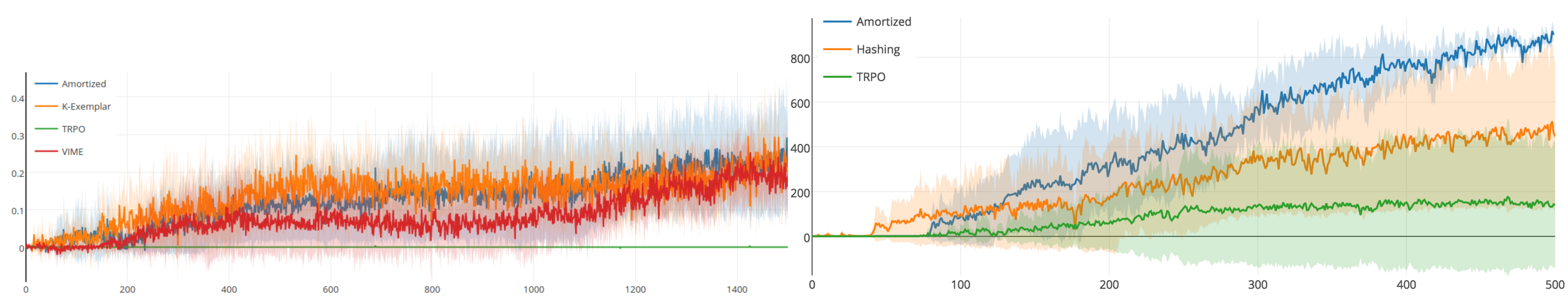


Figure: SwimmerGather (left) and Venture (right)

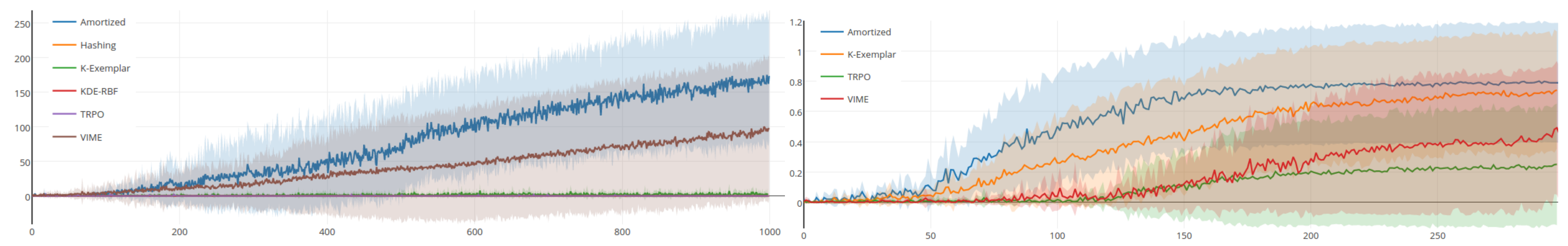


Figure: SparseHalfCheetah (left) and DoomMyWayHome+ (right)

Video results online:
`https://sites.google.com/view/ex2exploration/home`

Conclusions

We have presented:

- A method to obtain point density estimates using discriminators.
- An exploration method based on training only discriminative models that is scalable to high-dimensional observations such as images.

{justinjfu,jcoreyes,svlevine}@eecs.berkeley.edu