BEHAVIORAL ENTROPY-GUIDED DATASET GENERATION FOR OFFLINE REINFORCEMENT LEARNING

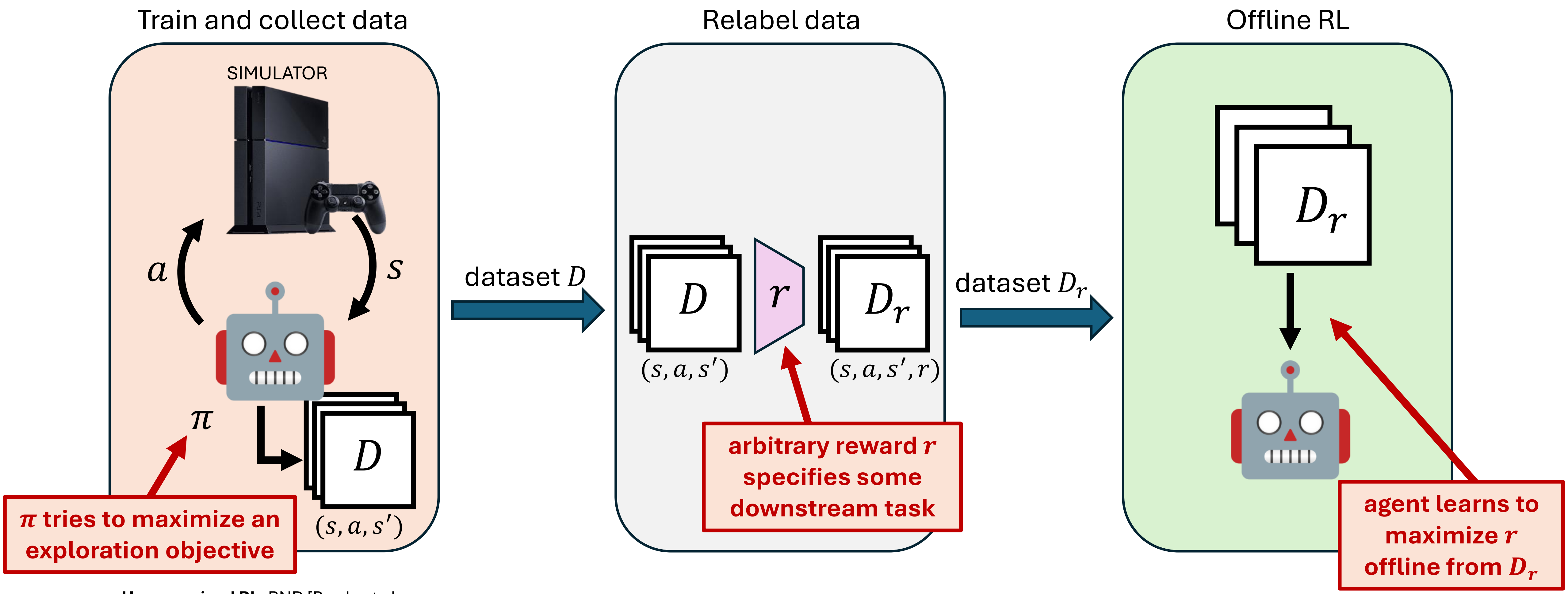
Wesley A. Suttle*, **Aamodh Suresh*** , Carlos Nieto-Granda

***Equal contribution**

ICLR 2025 Singapore, ***Thu 24 Apr 3 p.m***

DEVCOM U.S. Army Research Laboratory
Adelphi, MD, 20783, USA

Background: Exploratory data generation for offline RL



π tries to maximize an exploration objective

arbitrary reward r specifies some downstream task

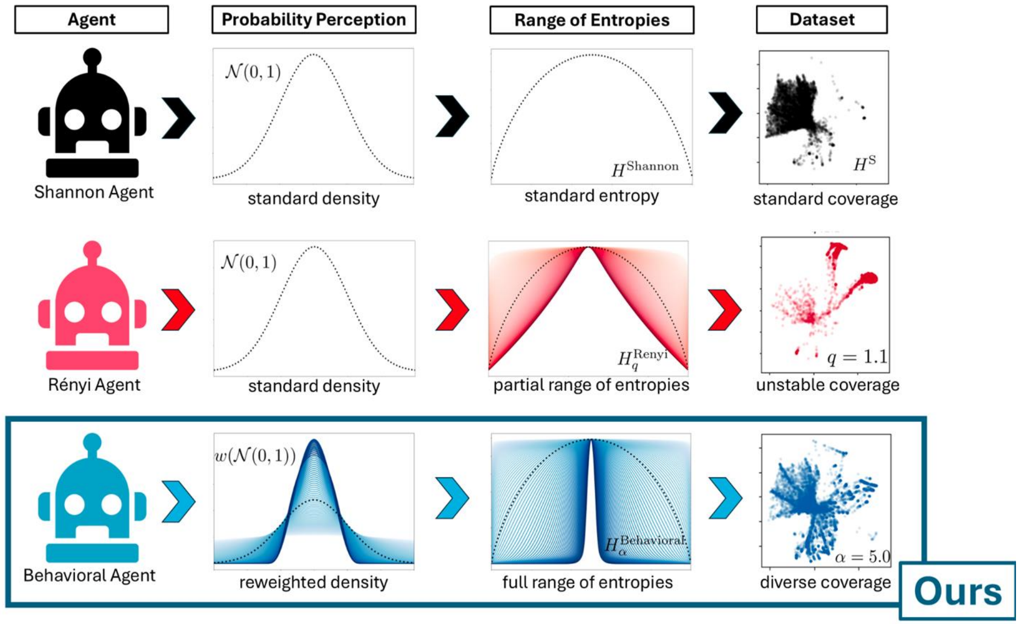
agent learns to maximize r offline from D_r

Unsupervised RL: RND [Burda et al., ICLR 2019], SMM [Lee et al., 2019], APT [Liu & Abbeel, NeurIPS 2021]

ExORL framework: [Yarats et al., ICLR 2022 Workshop on Generalizable Policy Learning in the Physical World]

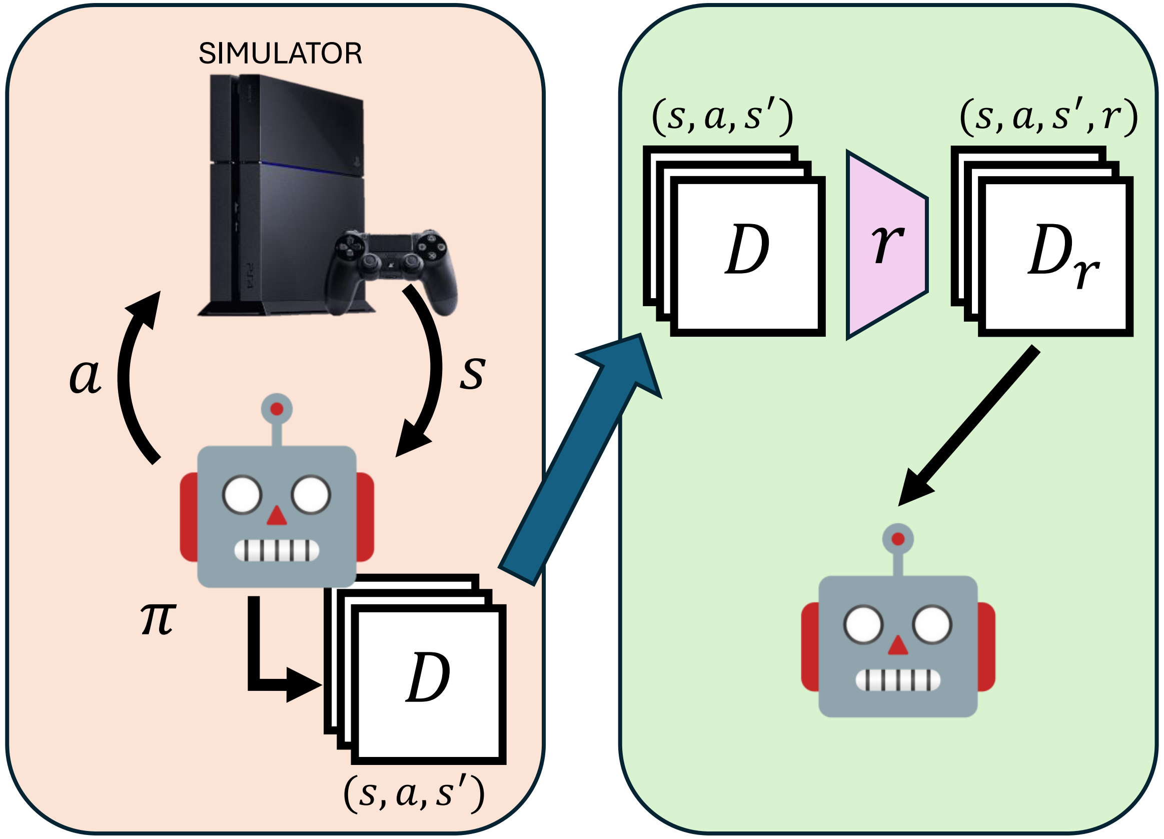
Offline RL: very well-developed; see surveys [Levine et al., 2020], [Prudencio et al., 2023]

Our work: new exploration objectives



Main Idea:

- Reweight probabilities using Behavioral economics certified functions
- Devolop most general entropy to evaluate coverage
- Wider range of exploration policies
- Better coverage and eventual Offline RL performance



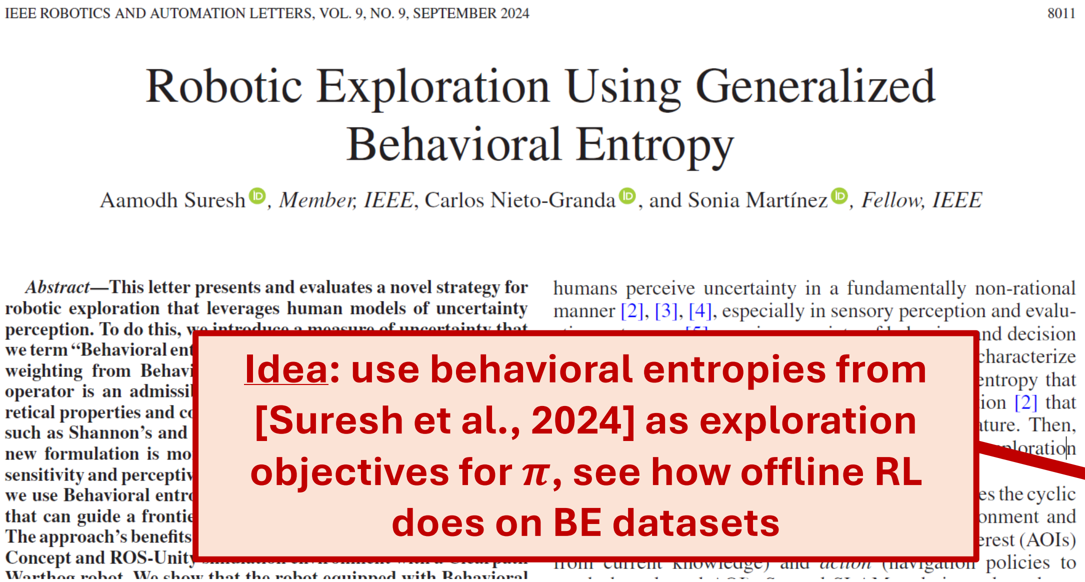
Train and collect data

Offline RL

Unsupervised RL: RND [Burda et al., ICLR 2019], SMM [Lee et al., 2019], APT [Liu & Abbeel, NeurIPS 2021]

Offline RL: see surveys [Levine et al., 2020], [Prudencio et al., 2023]

Our work: new exploration objectives



IEEE ROBOTICS AND AUTOMATION LETTERS, VOL. 9, NO. 9, SEPTEMBER 2024 8011

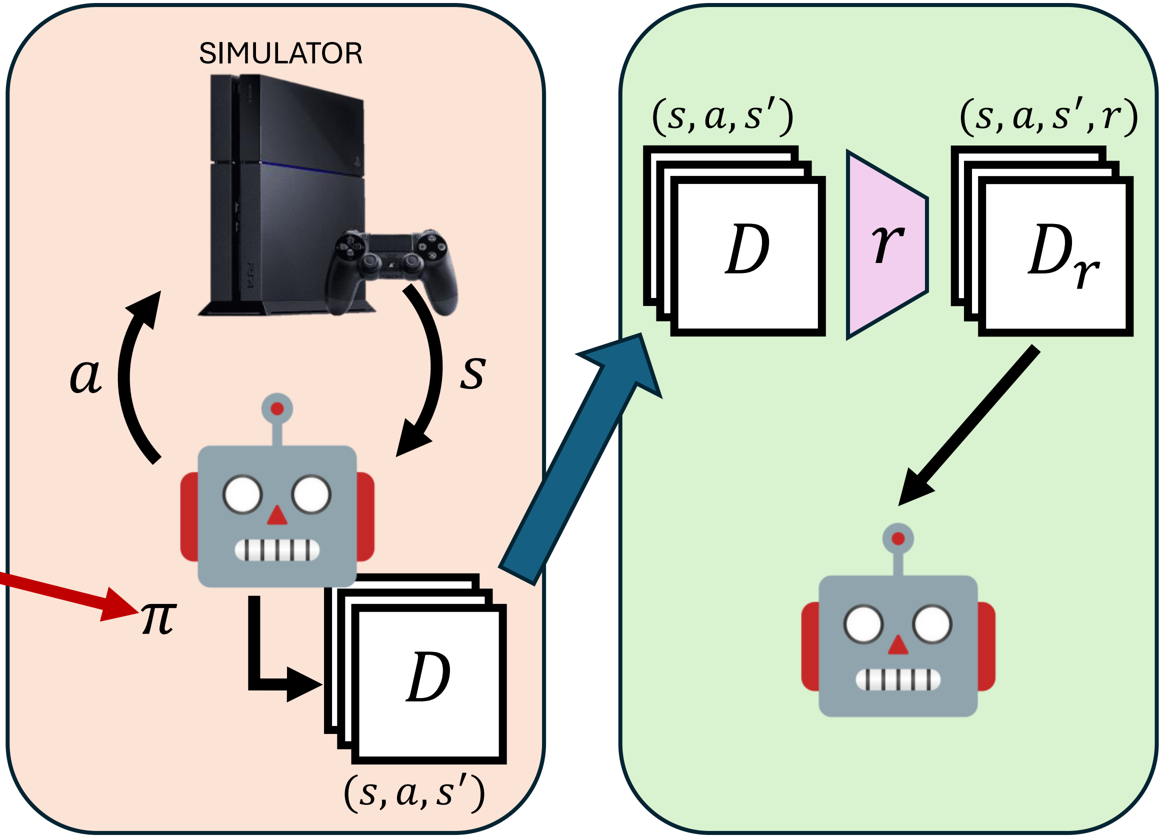
Robotic Exploration Using Generalized Behavioral Entropy

Aamodh Suresh, *Member, IEEE*, Carlos Nieto-Granda, and Sonia Martínez, *Fellow, IEEE*

Abstract—This letter presents and evaluates a novel strategy for robotic exploration that leverages human models of uncertainty perception. To do this, w
we term “Behavioral ent
weighting from Behavi
operator is an admissi
retical properties and co
such as Shannon’s and
new formulation is mo
sensitivity and perceptiv
we use Behavioral entro
that can guide a frontie
The approach’s benefits
Concept and ROS-Unity

humans perceive uncertainty in a fundamentally non-rational manner [2], [3], [4], especially in sensory perception and evalu-
nd decision
characterize
entropy that
ion [2] that
ture. Then,
ploration
es the cyclic
onment and
erest (AOIs)
from current knowledge) and *action* (navigation policies to

Idea: use behavioral entropies from [Suresh et al., 2024] as exploration objectives for π, see how offline RL does on BE datasets



Train and collect data

Offline RL

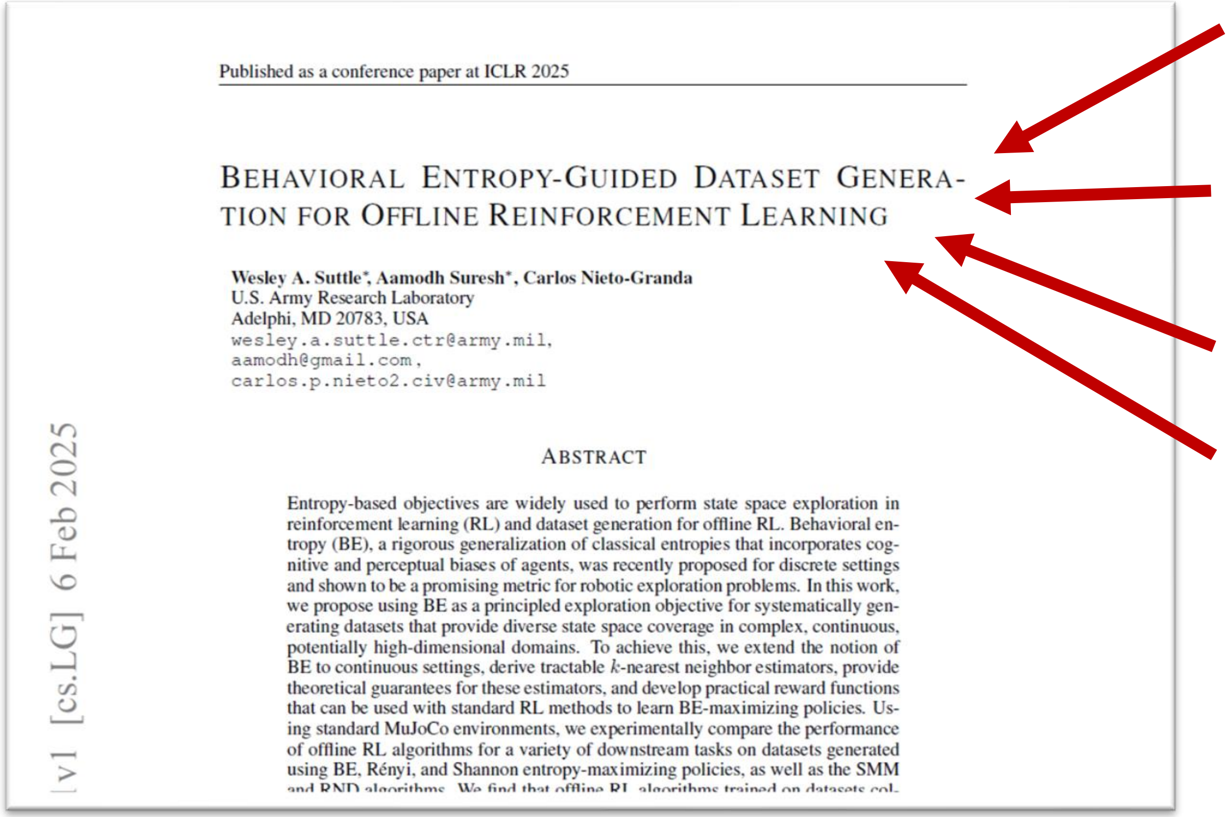
Unsupervised RL: RND [Burda et al., ICLR 2019], SMM [Lee et al., 2019], APT [Liu & Abbeel, NeurIPS 2021]

Offline RL: see surveys [Levine et al., 2020], [Prudencio et al., 2023]

Challenges:

- extension of BE to continuous spaces
- Continuous BE estimators
- RL algorithm development

Contributions



Published as a conference paper at ICLR 2025

BEHAVIORAL ENTROPY-GUIDED DATASET GENERATION FOR OFFLINE REINFORCEMENT LEARNING

Wesley A. Suttle*, Aamodh Suresh*, Carlos Nieto-Granda
U.S. Army Research Laboratory
Adelphi, MD 20783, USA
wesley.a.suttle.ctr@army.mil,
aamodh@gmail.com,
carlos.p.nieto2.civ@army.mil

ABSTRACT

Entropy-based objectives are widely used to perform state space exploration in reinforcement learning (RL) and dataset generation for offline RL. Behavioral entropy (BE), a rigorous generalization of classical entropies that incorporates cognitive and perceptual biases of agents, was recently proposed for discrete settings and shown to be a promising metric for robotic exploration problems. In this work, we propose using BE as a principled exploration objective for systematically generating datasets that provide diverse state space coverage in complex, continuous, potentially high-dimensional domains. To achieve this, we extend the notion of BE to continuous settings, derive tractable k-nearest neighbor estimators, provide theoretical guarantees for these estimators, and develop practical reward functions that can be used with standard RL methods to learn BE-maximizing policies. Using standard MuJoCo environments, we experimentally compare the performance of offline RL algorithms for a variety of downstream tasks on datasets generated using BE, Rényi, and Shannon entropy-maximizing policies, as well as the SMM and RND algorithms. We find that offline RL algorithms trained on datasets col-

- Extension of BE from [Suresh et al., 2024] to continuous spaces
- Developed and analyzed k-nearest neighbor (k-NN) BE estimators
- k-NN-based RL reward for BE
- Experiments demonstrating promising performance on BE-generated datasets

Behavioral entropy for continuous spaces

probability density

$w(f(x))$

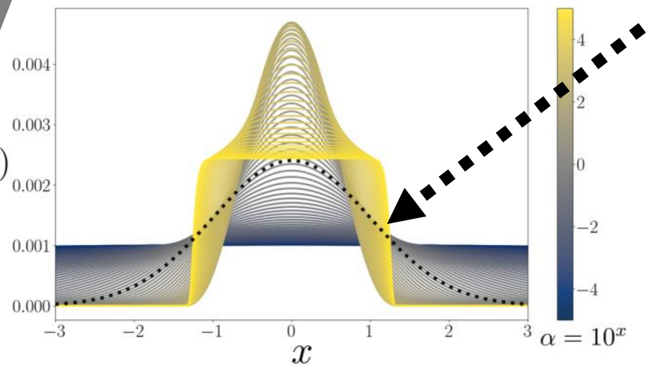


Fig. 1: Probability weightings transform probability densities

SHANNON ENTROPY (discrete)

$$H^S(X) = -\sum_{i=1}^{M} \log(p_i)p_i$$

RÉNYI ENTROPY (discrete)

$$H_q^R(X) = \frac{1}{1-q}\log\sum_{i=1}^{M} p_i^q$$

BEHAVIORAL ENTROPY (discrete)

$$H^B(X) = -\sum_{i=1}^{M} w(p_i)\log(w(p_i))$$

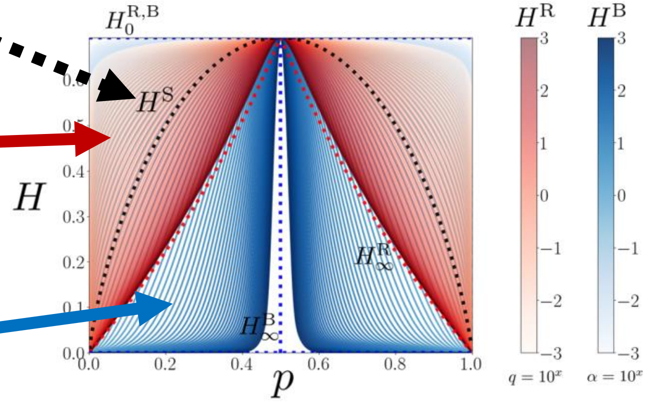


Fig. 2: Entropies achievable by Shannon, Rényi, behavioral entropies on Bernoulli probability density

$$w(x) = e^{-\beta(-\log x)^\alpha}, \quad \alpha, \beta > 0$$

Prelec **probability weighting function** (Prelec, 1998) modeling human uncertainty perception

Continuous-spaces behavioral entropy

general weighting *w*

$$H^{B,w}(f) = -\int_{\mathcal{X}} \log(w(f(x)))w(f(x))dx$$

Prelec weighting *w*

$$H^{B,\alpha,\beta}(f) = \beta\int_{\mathcal{X}} e^{-\beta(-\log(f(x)))^\alpha}(-\log f(x))^\alpha dx$$

k-nearest neighbor estimators, RL reward

k-NN entropy estimator formulation

$$X_1, X_2, \ldots, X_n \sim f(\cdot)$$

n i.i.d. samples

$$R_{k,n}(x) = \|x - NN_k(x)\|_2$$

distance from kth NN

$$\hat{f}(x) = \frac{k\Gamma(d/2+1)}{n\pi^{d/2}R^d_{k,n}(x)}$$

density estimator

behavioral entropy estimator for general w

$$\boxed{\widehat{H}^{B,w}_{k,n}(f) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\hat{f}(X_i)}w(\hat{f}(X_i))\log w(\hat{f}(X_i))}$$

k-NN estimator analysis

Theorem 1. Under suitable conditions on k, n, w, and f, we have $\widehat{H}^{B,w}_{k,n}(f) \to H^{B,w}(f)$ both uniformly and in probability.

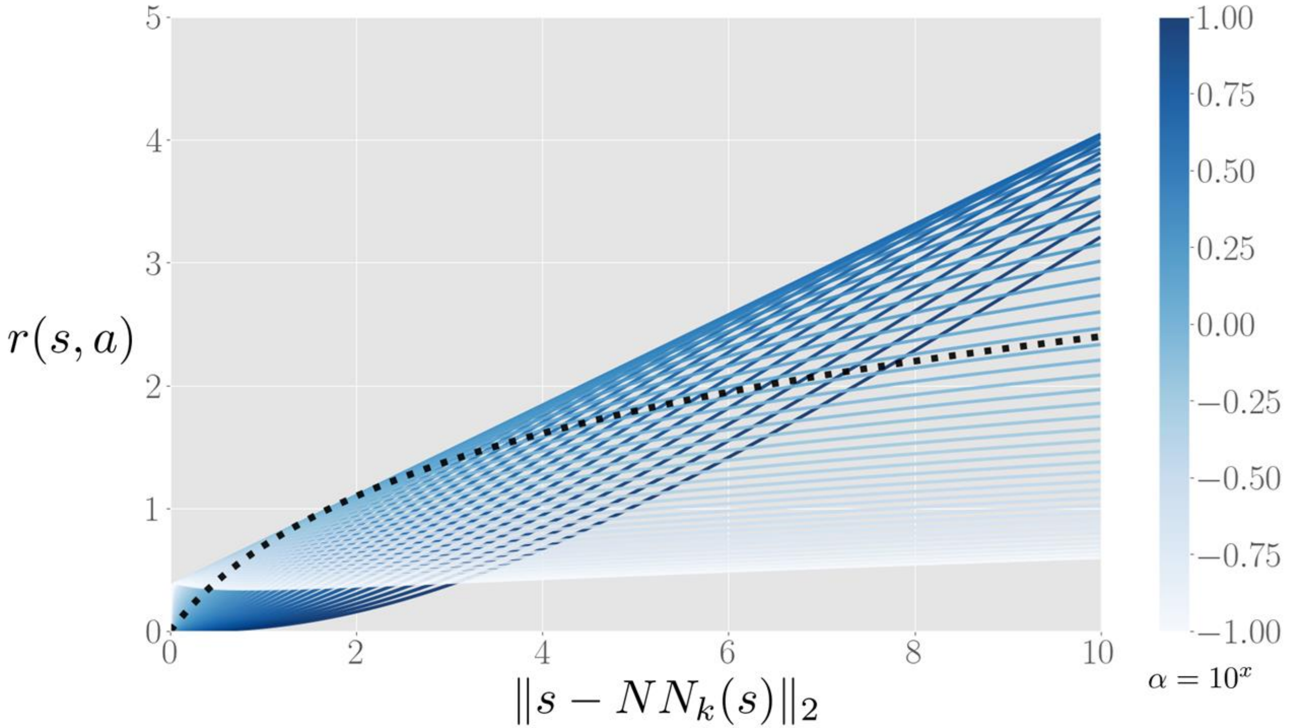
Theorem 2. Under suitable conditions on the density f and for fixed k, k-NN estimators of density functionals approximate their target functionals up to

$$O\left(\left(\frac{k}{n}\right)^{\frac{1}{d}} + \frac{1}{\sqrt{k}}\right).$$

k-NN-based RL reward

$$r(s,a) = \|s - NN_k(s)\|_2\, e^{-\beta(\log(\|s-NN_k(s)\|_2+c))^{\alpha}}\,(\log(\|s - NN_k(s)\|_2 + c))^{\alpha}$$

k-nearest neighbor estimators, RL reward



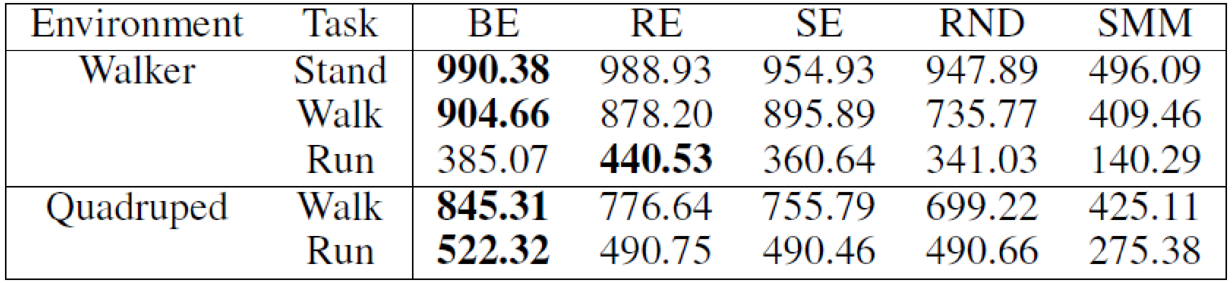
k-NN-based RL reward

$$r(s,a) = \|s - NN_k(s)\|_2 \, e^{-\beta(\log(\|s-NN_k(s)\|_2+c))^{\alpha}} \left(\log(\|s - NN_k(s)\|_2 + c)\right)^{\alpha}$$

Experimental setup and summary

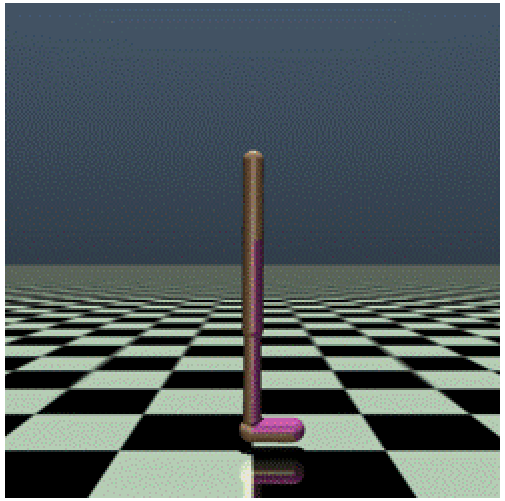
- Domains: Walker, Quadruped
- Tasks: Stand (on Walker only), Walk (both), Run (both)
- Data generation algorithms:
  - ICM-APT (Shannon), RND, SMM
  - ICM-APT (Rényi) for range of q
  - ICM-APT (BE) for range of α
- Offline RL evaluation methods:
  - TD3, CQL, CRR
- 500k data generation steps
- 100k offline RL training steps
- Evaluation every 10k steps

Table 1: Max performance over all offline RL algorithms and all trials

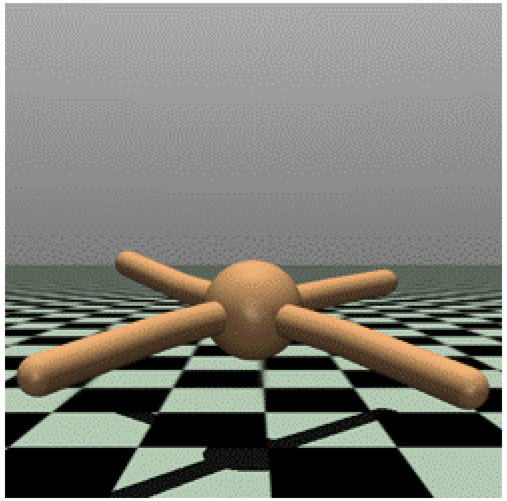


Environment	Task	BE	RE	SE	RND	SMM
Walker	Stand	**990.38**	988.93	954.93	947.89	496.09
	Walk	**904.66**	878.20	895.89	735.77	409.46
	Run	385.07	**440.53**	360.64	341.03	140.29
Quadruped	Walk	**845.31**	776.64	755.79	699.22	425.11
	Run	**522.32**	490.75	490.46	490.66	275.38

Walker

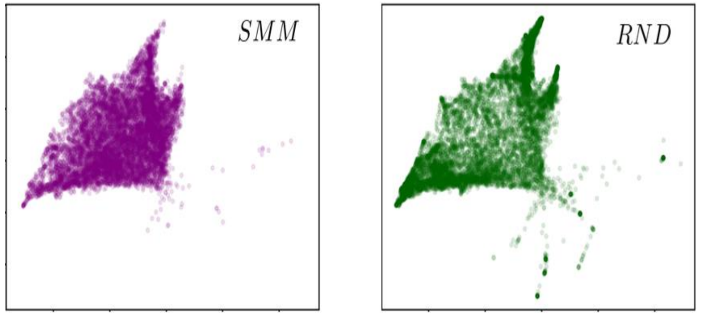


Quadruped



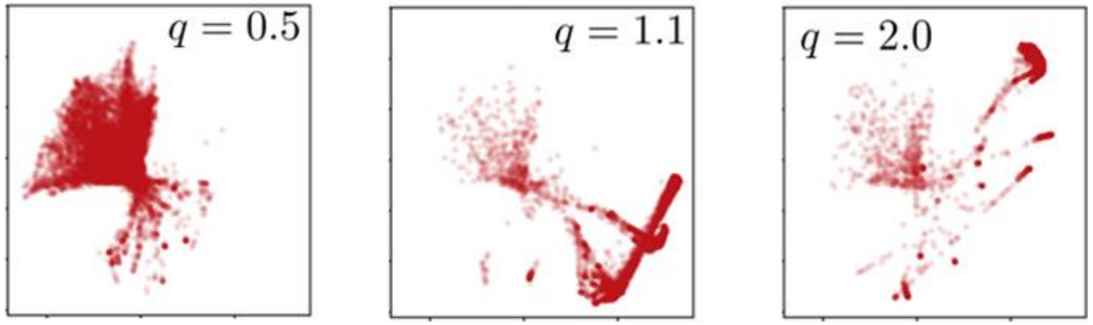
State Coverage using PHATE for Walker Domain

Traditional Objectives



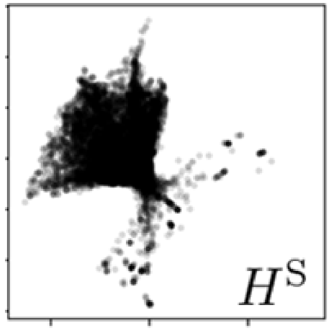
LIMITED COVERAGE

Renyi Entropic Objectives



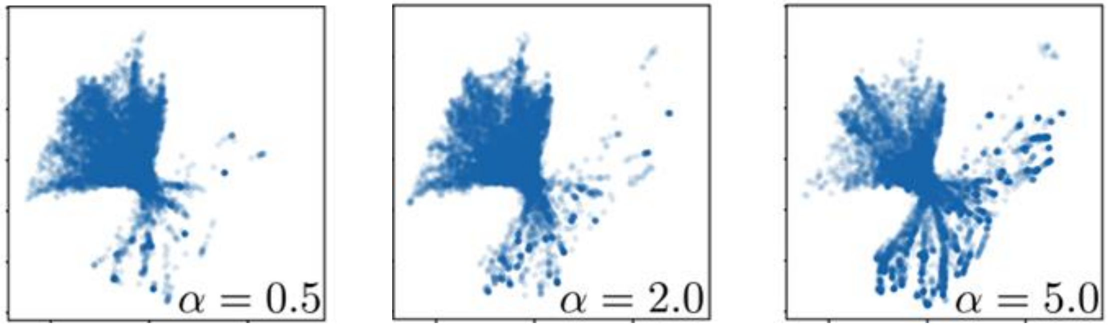
UNSTABLE COVERAGE

Shannon Entropic Objectives



INFLEXIBLE COVERAGE

Behavioral Entropic Objectives



EXTENSIVE, FLEXIBLE AND STABLE COVERAGE