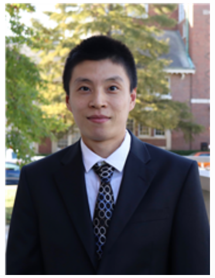
Harnessing Distribution Ratio Estimators for Learning Agents with Quality and Diversity

Tanmay Gangwani, Jian Peng, Yuan Zhou





ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Conference on Robot Learning (CoRL), 2020

Introduction and Motivation

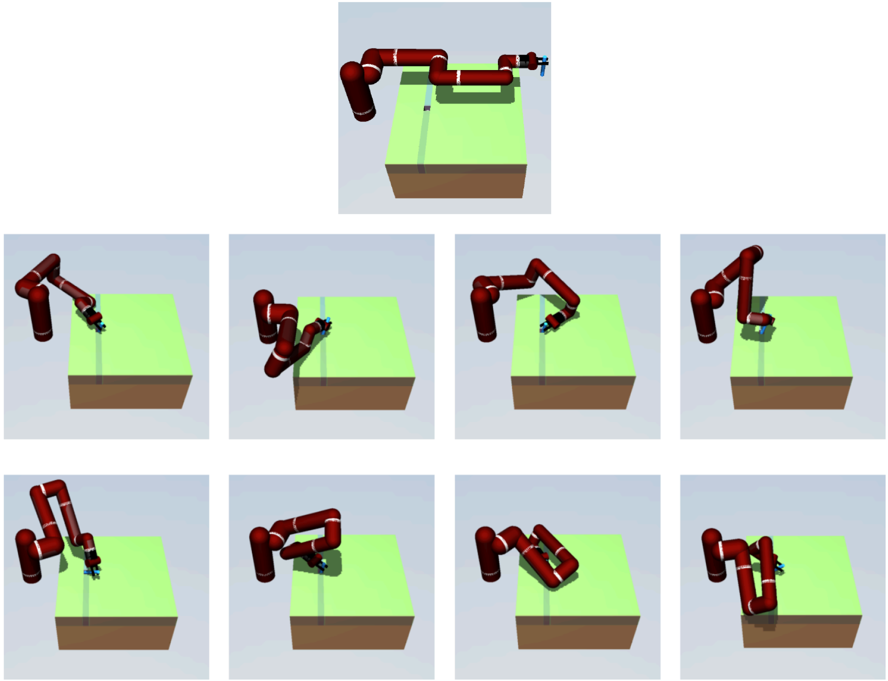
Objective: Learn agents that achieve **high task-returns** and are **behaviorally diverse**

Inspired by Quality-Diversity (QD) algorithms from the Neuroevolution literature

Benefits

- Efficient exploration
- Accelerated downstream tasks via skill-composition
- Transfer learning policy to mismatched environments
- Dynamically adaptive agents

Image from *Robots that can adapt like animals, Cully et al.*



Different ways to complete the peg-insertion task in a MuJoCo model of a 7 DOF arm, based on the Sawyer robot

Background: QD Training via Variational Inference

Learn a **high-entropy** distribution over policy parameters θ that **maximizes the expected returns**

$$\max_q \mathbb{E}_{\theta\sim q}[\eta(\theta)] + \mathcal{H}(q)\,; \qquad \mathcal{H}(q) = \mathbb{E}_{\theta\sim q}[-\log q(\theta)]$$

Stein-Variational Policy Gradient[1-3] (SVPG) provides a solution:

$$\theta_i \leftarrow \theta_i + \epsilon\Delta\theta_i, \qquad \Delta\theta_i = \frac{1}{n}\sum_{j=1}^{n}\Big[\underbrace{\nabla_{\theta_j}\eta(\pi_{\theta_j})k(\theta_j,\theta_i)}_{\text{Quality-enforcing}} + \underbrace{\nabla_{\theta_j}k(\theta_j,\theta_i)}_{\text{Diversity-enforcing}}\Big]$$

Ensemble of *n* *interacting* policies

[1] Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm, Liu and Wang
[2] Stein Variational Policy Gradient, Liu et al.
[3] Learning Self-Imitating Diverse Policies, Gangwani et al.

Recast Objective with Density Ratios

$$\Delta\theta_i = \frac{1}{n}\sum_{j=1}^{n}\Big[\underbrace{\nabla_{\theta_j}\eta(\pi_{\theta_j})k(\theta_j,\theta_i)}_{\text{Quality-enforcing}} + \underbrace{\nabla_{\theta_j}k(\theta_j,\theta_i)}_{\text{Diversity-enforcing}}\Big] \quad (1)$$

- $\boldsymbol{\rho}_\pi$ is the stationary discounted state-action visitation distribution of the policy π

$$\zeta_{ij}(s,a) = \frac{\rho_{\pi_i}(s,a)}{\rho_{\pi_j}(s,a)}$$

Density Ratio Estimation (DRE)

- Kernels based on *f*-divergence between visitation distributions

$$k_f(\theta_j,\theta_i) = \exp(-D_f(\rho_{\pi_{\theta_j}}, \rho_{\pi_{\theta_i}})/T)$$

- For Jenson-Shannon, KL, Symmetric-KL, (1) can be written as a function of ζ_{ij}

Harnessing DRE Methods

- Noise Contrastive Estimation (NCE)
  - A binary classification objective directly estimates $\boldsymbol{\rho}_\pi$ given samples from π and a noise distribution
  - Compute ζ_{ij} explicitly using output of two networks
  - Requires on-policy samples

$$\zeta_{ij}(s,a) = \frac{\rho_{\pi_i}(s,a)}{\rho_{\pi_j}(s,a)}$$

- **DI**stribution **C**orrection **E**stimation (DICE)
  - Train neural network to directly estimate ζ_{ij}
  - Only *(s,a,s')* samples from π_j are required for ζ_{ij} (no samples from π_i)
  - DualDICE[1], ValueDICE[2], GenDICE[3], …

[1] DualDICE: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections, Nachum et al.
[2] Imitation Learning via Off-Policy Distribution Matching, Kostrikov et al.
[3] GenDICE: Generalized Offline Estimation of Stationary Values, Zhang et al.

Algorithm Sketch

$$\Delta\theta_i = \frac{1}{n}\sum_{j=1}^{n}\Big[\underbrace{\nabla_{\theta_j}\eta(\pi_{\theta_j})k(\theta_j,\theta_i)}_{\text{Quality-enforcing}} + \underbrace{\nabla_{\theta_j}k(\theta_j,\theta_i)}_{\text{Diversity-enforcing}}\Big] \quad (1)$$

Initialize ensemble of n policies $\{\pi_i\}_1^n$

Initialize density ratio estimation networks ζ_{ij}^{ϕ} *// parameterization depends on the DRE method*

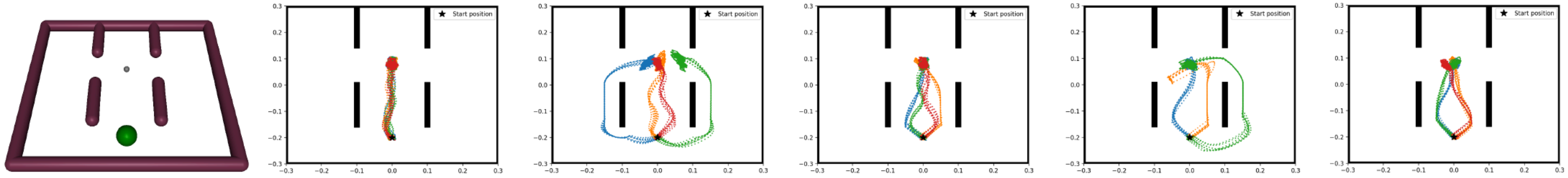
Sample trajectories using all the policies $\{\pi_i\}_1^n$

Update all ζ_{ij}^{ϕ} networks *// loss function depends on the DRE method*

Use ζ_{ij}^{ϕ} to compute the kernel-value and kernel-gradient (required in (1))

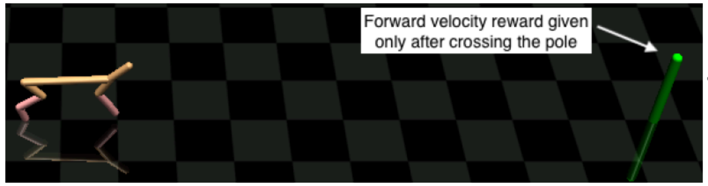
Update all policies with the SVPG gradient (1)

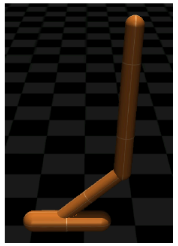
Experiments

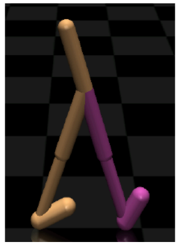


(a) (b) Standard RL (c) QD-NCE-JS (d) QD-NCE-KLS (e) QD-*GenDICE*-JS (f) QD-*GenDICE*-KLS

- Qualitatively diverse behaviors in 2D-Maze navigation
- Multi-modal locomotion with deceptive rewards
- Emergence of skills in absence of environmental rewards
- Quantitative comparison of the NCE and DICE-based estimators using a metric correlated with behavioral diversity







Takeaways

- Learning diverse and high-return policies
- Extend SVPG with kernels based on *f*-divergence between the stationary distributions of policies
- For kernels based on {JS, KL, Symmetric-KL}, reduce the problem to efficient DRE
- Harness NCE and DICE-based algorithms for DRE

Paper + Code: *https://github.com/tgangwani/QDAgents*

Thank you!☺