Distributional Policy Optimization: An Alternative Approach for Continuous Control

Chen Tessler*, Guy Tennenholtz* and Shie Mannor

Technion Institute of Technology, Israel

Continuous Control

Setting:

- MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{A}$ is continuous.

Objective:

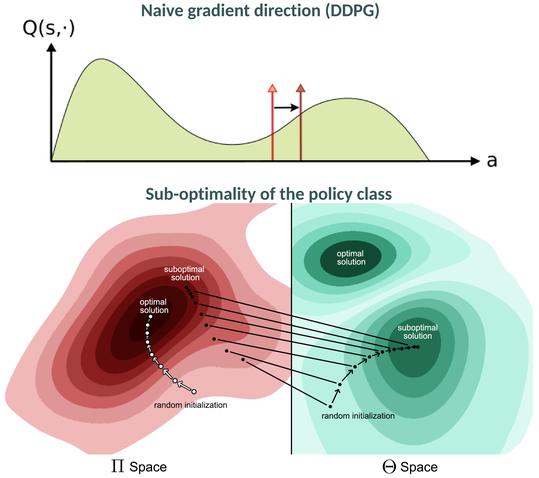
Find π^* which satisfies

$$\pi^*(s) \in \arg\max_{\pi \in \Pi} Q^\pi(s, \pi(s)) = \arg\max_{\pi \in \Pi} \mathbb{E}^\pi \left[\gamma^t r_t | s_0 = s\right]$$

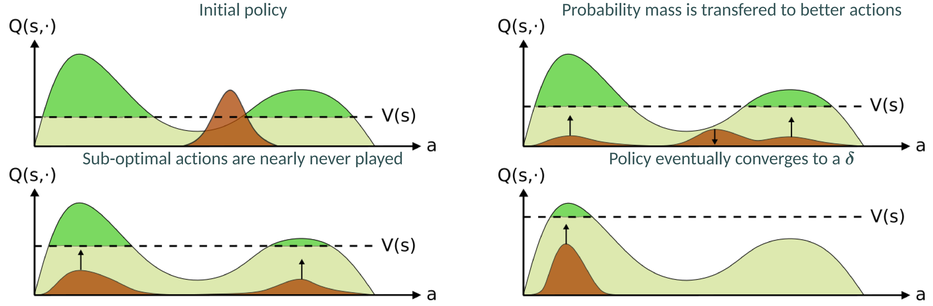
Problem: $Q(s, a)$ is *not necessarily convex*!

In this work we provide a framework for tackling continuous control problems with non-convex returns

Pitfalls of Current Approaches

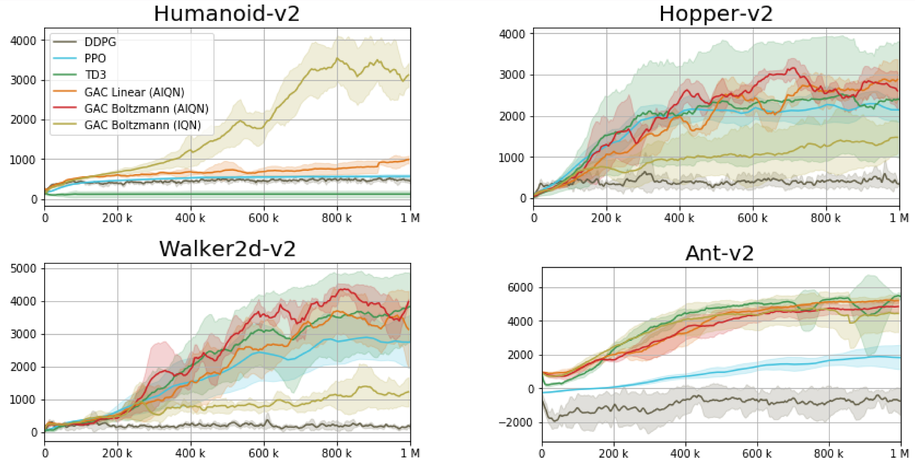


Distributional Policy Optimization (Illustration)



DPO moves probability mass from sub-optimal actions towards better ones, while considering the entire action space

Experiments



Environment	Humanoid-v2	Walker2d-v2	Hopper-v2	HalfCheetah-v2	Ant-v2	Swimmer-v2
Relative Result	$+2968\ (+512\%)$	$+533\ (+14\%)$	$+394\ (+14\%)$	$-381\ (-4\%)$	$-272\ (-5\%)$	$-85\ (-65\%)$

Relative best GAC results compared to the best policy gradient baseline

Distributional Policy Optimization (Algorithm)

1: Input: learning rates $\alpha_k \gg \beta_k \gg \delta_k$
2: $\pi_{k+1} = \Gamma\left(\pi_k - \alpha_k \nabla_\pi d(\mathcal{D}^{\pi'_k}_{I^{\pi'_k}}, \pi) \mid_{\pi=\pi_k}\right)$
3: $Q^{\pi'}_{k+1}(s,a) = Q^{\pi'}_k(s,a) + \beta_k \left(r(s,a) + \gamma v^{\pi'}_k(s) - Q^{\pi'}_k(s,a)\right)$
4: $v^{\pi'}_{k+1}(s) = v^{\pi'}_k + \beta_k \int_{\mathcal{A}} \left(Q^{\pi'}_k(s,a) - v^{\pi'}_k(s)\right)$
5: $\pi'_{k+1} = \pi'_k + \delta_k(\pi_k - \pi'_k)$

Summary

1. We proposed the Distributional Policy Optimization (DPO) framework for tackling non-convex returns.
2. DPO requires the ability to represent arbitrary policies and optimize by minimizing the distance to a target distribution.
3. We achieve this by modeling the actor (policy) using an Autoregressive Implicit Quantile Network, a generative model.
4. Empirical tests attain results competitive to policy gradient methods, while remaining as sample efficient.

Future Work and Possible Extensions

1. In this work we modeled the policy using an Autoregressive Implicit Quantile Network, however it can also be modeled using other methods such as GANs, VAEs and Normalizing Flows.
2. While the approach is efficient in the number of samples (environment interactions) it is not as computational efficient as alternative methods.
3. As the optimization considers a target probability distribution, this can be leveraged for improving exploration and sample efficiency using UCB-like methods.

Additional Details

1. Email: `chen.tessler@campus.technion.ac.il`, `guytenn@gmail.com`
2. Code: `github.com/tesslerc/GAC`