



人工智慧作業報告簡報

信息工程學院 2021 級 干皓丞 2101212850

DAGs with No Fears: A Closer Look at Continuous Optimization for Learning

Bayesian Networks

無所畏懼的 DAG：深入了解學習貝葉斯網絡的持續優化

<https://arxiv.org/abs/2010.09133>

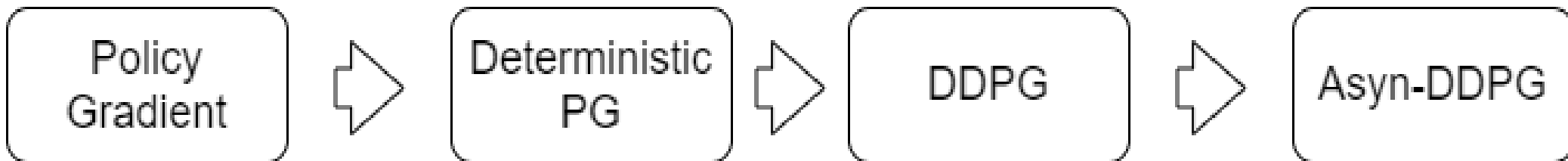
Neural Information Processing Systems (NeurIPS 2020)




Subjects: Machine Learning (cs.LG); Machine Learning (stat.ML)





Motivation 动机



arXiv  Cornell University  

Computer Science > Machine Learning

arXiv:1509.02971 (cs)

[Submitted on 9 Sep 2015 (v1), last revised 5 Jul 2019 (this version, v6)]

Continuous control with deep reinforcement learning

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, Daan Wierstra

Download PDF

We adapt the ideas underlying the success of Deep Q-Learning to the continuous action domain. We present an actor-critic, model-free algorithm based on the deterministic policy gradient that can operate over continuous action spaces. Using the same learning algorithm, network architecture and hyper-parameters, our algorithm robustly solves more than 20 simulated physics tasks, including classic problems such as cartpole swing-up, dexterous manipulation, legged locomotion and car driving. Our algorithm is able to find policies whose performance is competitive with those found by a planning algorithm with full access to the dynamics of the domain and its derivatives. We further demonstrate that for many of the tasks the algorithm can learn policies end-to-end: directly from raw pixel inputs.

Comments: 10 pages + supplementary

Subjects: **Machine Learning (cs.LG)**; Machine Learning (stat.ML)

Cite as: [arXiv:1509.02971](https://arxiv.org/abs/1509.02971) [cs.LG]
(or [arXiv:1509.02971v6](https://arxiv.org/abs/1509.02971v6) [cs.LG] for this version)



Intuition 直觉

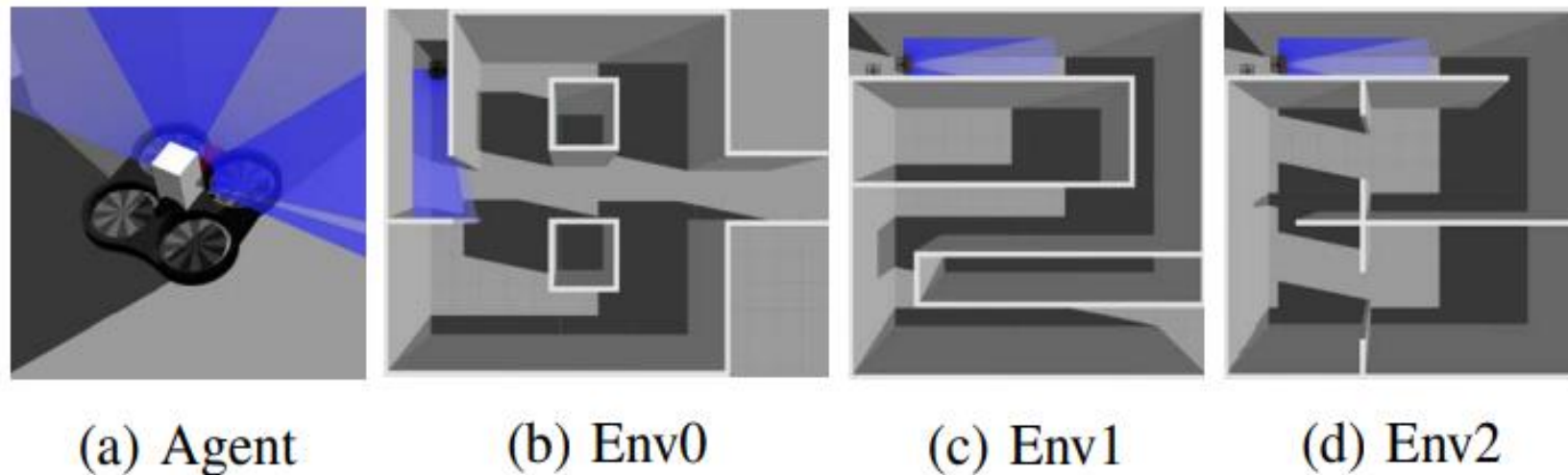


Fig. 3: The local patterns used for training. For each local pattern, a circle is a initial location, and a star is a goal.

在研究者的過往研究經驗中發現稀疏策略梯度可以顯著減少衝突任務之間的干擾，並使多任務學習有著更加穩定的樣本效率。為了確保為每個任務評估的梯度的稀疏性，**Asyn-DDPG** 將演員和評論家函數都表示為深度神經網絡，並使用 **Dropout** 對其進行正則化。在訓練期間，工作(**agents**)共享**actor** 和**critic** 函數，並使用特定於任務的梯度異步優化它們。另外，該策略可以很好地泛化以處理訓練期間看不見的環境。



Justification 理由

1) 評估的梯度是稀疏的，
將梯度異步應用到共享策略和 q 函數的策略。

2) 複製機制，在訓練期間，每個工作代理可以將他們在不同任務中學到的知識轉移到其他agents。

這允許每個工作代理根據其他 agent 完成的工作改進共享策略，這對於代理將其行為合成為一致的meta policy (i.e. the shared policy)至關重要。

由於梯度評估不依賴於任務特定信息，所有工作代理共同學習的策略可以推廣到不熟悉和可能更複雜的任務。這也是研究者所說的該策略可以很好地泛化以處理訓練期間看不見的環境。

Algorithm 1 A robot applies gradients to shared networks.

```
1: function ASYNUPDATE()
2:   copy  $w^*$  and  $\theta^*$  to  $w$  and  $\theta$ 
3:   while the shared policy hasn't converge do
4:     deploy the robot at a fixed initial location
5:     while an episode is not terminated do
6:       store experience to replay buffer
7:       sample a batch of experience
8:       evaluate policy gradient  $\delta_w$ 
9:       evaluate temporal-difference gradients  $\delta_\theta$ 
10:      apply  $\delta_w$  and  $\delta_\theta$  to  $w^*$  and  $\theta^*$ 
11:      copy  $w^*$  and  $\theta^*$  to  $w$  and  $\theta$ 
12:    end while
13:  end while
14: end function
```

可以從該研究中的演算法 1，看到這裡所總結的所有agent遵循的異步更新程序。 $\mu(\theta_t^i)$ 和 $Q(w_t^i)$ 是代理 i 的共享策略和 q 函數的本地副本。在時間 t ，工作agent i 評估確定性策略梯度和時間差異梯度。



Framework 框架

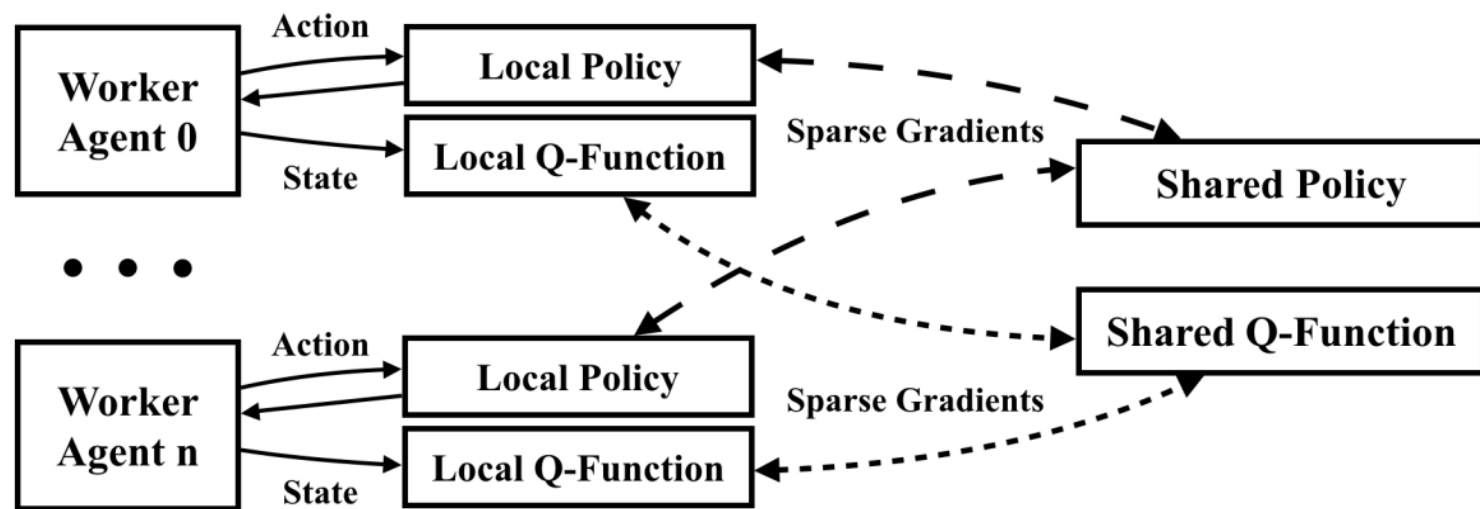


Fig. 1: The overview of Asyn-DDPG

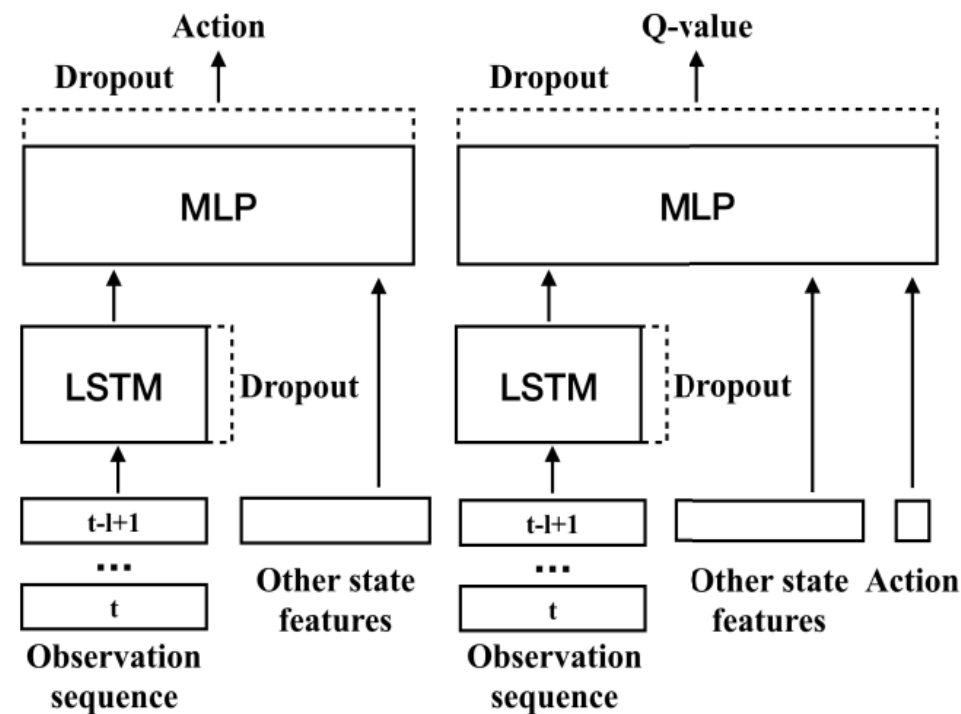


Fig. 2: The shared policy (left) and q-function (right).

可以看到該研究的異步 DDPG 可以將不同的 ”退出率” 配置應用於其本地策略和 q 函數，且研究者發現讓所有工作 agent 採用相同的退出率配置已經可以讓 Asyn-DDPG 收斂到一個健壯的解決方案。



Result 结果

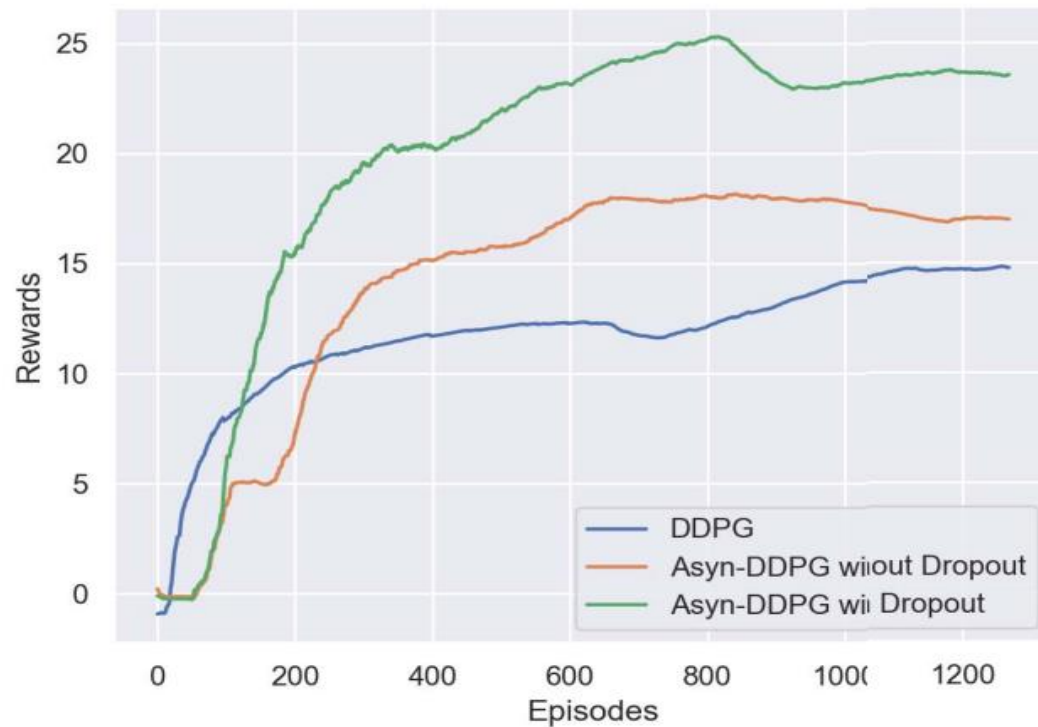


Fig. 4: Learning performance of DDPG and Asyn-DDPG in terms of episodic rewards. Both methods are evaluated in Env3.

從在研究文獻中的圖 4，名為 DDPG 和 Asyn-DDPG 在情節獎勵方面的學習性能比較中，可以看到研究者在這兩種方法都在 Env3 中進行了評估，可以看到研究者所提出的方法性能高於 DDPG。