# Towards Generalization and Simplicity in Continuous Control

Aravind Rajeswaran* Kendall Lowrey* Emanuel Todorov Sham Kakade
University of Washington Seattle
{ aravraj, klowrey, todorov, sham } @ cs.washington.edu

# 【强化学习 39】Linear/RBF Representation

张楚珩
清华大学 交叉信息院博士在读



这篇文章并没有提出什么新算法，只是考察了一下不用神经网络而是使用一些简单的特征表示（linear or RBF representation）是不是也能在benchmark上有较好的效果。

## 原文传送门

Rajeswaran, Aravind, et al. "Towards generalization and simplicity in continuous control." Advances in Neural Information Processing Systems. 2017.

## 特色

本专栏前面也介绍了一些使用简单表示（representation）而训练获得较好效果的算法，这里也是仅仅使用线性和RBF表示却在Mujoco任务上取得了很好的效果。同时，通过随机化初始分布来进行训练，获得了一定的push recovery的能力。最开始吸引我读这篇文章的主要原因是它在Mujoco任务上的分数较高，同时有很好的sample complexity。

## 过程

### 1. 使用更为简单的特征表示

这篇文章认为一个完整的强化学习算法主要包括以下三个部分：特征表示（representation）、优化方法（optimization method）和任务设计与建模（task design and modeling）。这里主要想搞明白之前那些成功的算法中，神经网络的特征表示是否有必要，这里作者选用简单的线性表示和RBF表示来作为特征表示，也获得了很好的实验结果。其中RBF特征表示构造如下

$$y_t^{(i)} = \sin\left(\frac{\sum_j P_{ij} s_t^{(j)}}{\nu} + \phi^{(i)}\right),$$

其中 $s_t$ 为原来的特征； $P_{ij}$ 是从标准正态分布中取出来的随机数； $\nu$ 是一个固定的数，带宽；相位 $\phi^{(i)}$ 采样自均匀分布 $U[-\pi,\pi)$ 。RBF特征再经过一个线性策略的映射即得到RBF策略的输出（行动的均值）。

本文使用的优化方法是NPG（natural policy gradient），是一个二阶优化方法。

**2. 获得更好的鲁棒性**

为了获得更好的鲁棒性，本文做了采取了如下措施：

- 去掉了Mujoco环境下的终止条件（应该是Gym wrapper中规定的终止条件）
- 使用随机化的初始化状态来进行训练
- 使用了与扰动相关的奖励函数来鼓励智能体探索如何对抗某种特定的扰动

这使得训练得到的智能体能够更好地应对中途的一些扰动。这这篇文章中主要试验了这三种扰动：Walker和Hopper任务中突然戳小人一下，或者Swimmer任务中旋转一下小人的前进方向。

## 结果

作者用的是Mujoco-v1的环境，不知道和Mujoco-v2差距有多大。以个人的经验，如果是v2环境的话，其末端的渐进得分还是挺高的。另外，其采样效率高，这一点不是很奇怪，不仅仅因为其策略参数少，而且也因为它使用了二阶优化方法NPG。

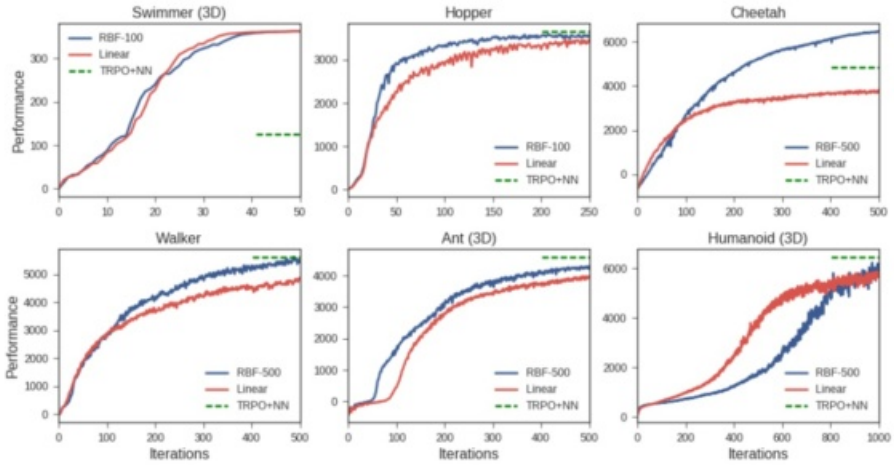


Figure 1: Learning curves for the Linear and RBF policy architectures. The green line corresponding to the reward achieved by neural network policies on the OpenAI Gym website, as of 02/24/2017 (trained with TRPO). It is observed that for all the tasks, linear and RBF parameterizations are competitive with state of the art results. The learning curves depicted are for the stochastic policies, where the actions are sampled as $a_t \sim \pi_\theta(s_t)$. The learning curves have been averaged across three runs with different random seeds.

实验效果看起来不错

在智能体鲁棒性的测试上面，作者对Swimmer、Walker、Hopper施加扰动，然后测试它们各自在扰动后走的距离，以此来判定小人是否具有抗干扰的能力。

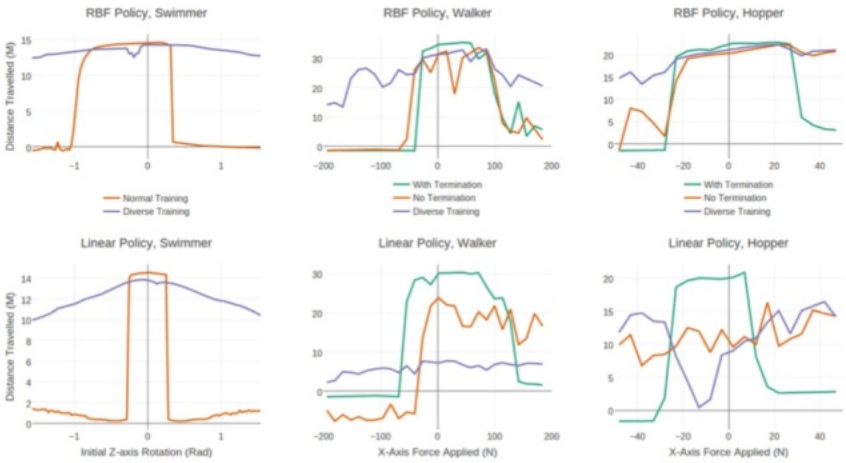


Figure 4: We test policy robustness by measuring distanced traveled in the swimmer, walker, and hopper tasks for three training configurations: (a) with termination conditions; (b) no termination, and peaked initial state distribution; and (c) with diverse initialization. Swimmer does not have a termination option, so we consider only two configurations. For the case of swimmer, the perturbation is changing the heading angle between $-\pi/2.0$ and $\pi/2.0$, and in the case of walker and hopper, an external force for 0.5 seconds along its axis of movement. All agents are initialized with the same positions and velocities.



可以把紫线看做实验组，绿线看做对照组

---

**这篇工作做的不是很完善啊**

1. 别人的算法也没跑出来对比，只是画了跟线表示了一下TRPO算法的渐进性能
2. 如果仅仅通过随机化初始状态来获得push recovery的能力还可以，但是又用了特定的奖励函数，这样看起来就不是那么有意思了

发布于 2019-01-29

强化学习 (Reinforcement Learning)



文章被以下专栏收录

强化学习前沿
读呀读paper