# HIGH-DIMENSIONAL CONTINUOUS CONTROL USING GENERALIZED ADVANTAGE ESTIMATION

**John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan and Pieter Abbeel**
Department of Electrical Engineering and Computer Science
University of California, Berkeley
{joschu,pcmoritz,levine,jordan,pabbeel}@eecs.berkeley.edu

# 【强化学习技术 28】GAE

张楚珩
清华大学 交叉信息院博士在读



全称是generalized advantage estimator，几乎所有最先进的policy gradient算法实现里面都使用了该技术。

## 原文传送门

Schulman, John, et al. "High-dimensional continuous control using generalized advantage estimation." arXiv preprint arXiv:1506.02438 (2015).

## 特色

这篇文章介绍了一种能够广泛适用的advantage的估计方法，所估计的advantage应用在策略梯度类方法里面能够有效减小梯度估计的方差，从而降低训练所需要的样本。该方法一经发明之后广泛地被应用到各种最先进的强化学习算法实现中。

## 过程

### 1. 策略梯度的估计

策略梯度的估计有多种不同的形式，下面列出的这些形式都是无偏估计（注意到Q、V、A都是准确的），但有着不同的方差，其中advantage（第4种或者第6种）几乎有最小的方差。

$$g = \mathbb{E}\left[\sum_{t=0}^{\infty} \Psi_t \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right], \tag{1}$$

where $\Psi_t$ may be one of the following:

1. $\sum_{t=0}^{\infty} r_t$: total reward of the trajectory.
2. $\sum_{t'=t}^{\infty} r_{t'}$: reward following action $a_t$.
3. $\sum_{t'=t}^{\infty} r_{t'} - b(s_t)$: baselined version of previous formula.
4. $Q^\pi(s_t, a_t)$: state-action value function.
5. $A^\pi(s_t, a_t)$: advantage function.
6. $r_t + V^\pi(s_{t+1}) - V^\pi(s_t)$: TD residual.

$$V^{\pi}(s_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t:\infty}}}\left[\sum_{l=0}^{\infty} r_{t+l}\right] \qquad Q^{\pi}(s_t, a_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t+1:\infty}}}\left[\sum_{l=0}^{\infty} r_{t+l}\right] \tag{2}$$

$$A^{\pi}(s_t, a_t) := Q^{\pi}(s_t, a_t) - V^{\pi}(s_t), \quad \text{(Advantage function)}. \tag{3}$$

**2. 引入参数 $\gamma$ 的策略梯度估计**

其实这里引入的参数 $\gamma$ 的形式和discount rate一样，只不过这里把它当做一种参数。

$$g^{\gamma} := \mathbb{E}_{\substack{s_{0:\infty}\\ a_{0:\infty}}}\left[\sum_{t=0}^{\infty} A^{\pi,\gamma}(s_t, a_t)\nabla_{\theta} \log \pi_{\theta}(a_t \mid s_t)\right]. \tag{6}$$

$$V^{\pi,\gamma}(s_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t:\infty}}}\left[\sum_{l=0}^{\infty} \gamma^l r_{t+l}\right] \qquad Q^{\pi,\gamma}(s_t, a_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t+1:\infty}}}\left[\sum_{l=0}^{\infty} \gamma^l r_{t+l}\right] \tag{4}$$

$$A^{\pi,\gamma}(s_t, a_t) := Q^{\pi,\gamma}(s_t, a_t) - V^{\pi,\gamma}(s_t). \tag{5}$$

注意到这里的 $g^{\gamma}$ 相对于前面的 $g$ 是有偏的。不过在大多数的带有discount rate的强化学习问题里面，实际上也是以discounted cumulative reward为目标的，相应的策略梯度估计就是这里的这种。

接下来文中给出了 $\gamma$ -*just* 的定义，其实就是说找到 $A^{\pi,\gamma}$ 的一个估计 $\hat{A}_t$ ，使得用这个估计来计算得到的梯度估计期望不变。

**Definition 1.** *The estimator $\hat{A}_t$ is $\gamma$-just if*

$$\mathbb{E}_{\substack{s_{0:\infty}\\ a_{0:\infty}}}\left[\hat{A}_t(s_{0:\infty}, a_{0:\infty})\nabla_{\theta} \log \pi_{\theta}(a_t \mid s_t)\right] = \mathbb{E}_{\substack{s_{0:\infty}\\ a_{0:\infty}}}\left[A^{\pi,\gamma}(s_t, a_t)\nabla_{\theta} \log \pi_{\theta}(a_t \mid s_t)\right]. \tag{7}$$

如下的这些估计都是 $\gamma$ -*just*的

- $\sum_{l=0}^{\infty} \gamma^l r_{t+l}$
- $Q^{\pi,\gamma}(s_t, a_t)$
- $A^{\pi,\gamma}(s_t, a_t)$
- $r_t + \gamma V^{\pi,\gamma}(s_{t+1}) - V^{\pi,\gamma}(s_t)$.

**3. GAE**

文章提出一种generalized advantage estimator，定义如下

$$\hat{A}_t^{GAE(\gamma,\lambda)} = \sum_{l=1}^{\infty}(\gamma\lambda)^l\delta_{t+l}^V = \sum_{l=1}^{\infty}(\gamma\lambda)^l(r_t + \gamma V(s_{t+l+1}) - V(s_{t+l}))$$

它具有如下性质：

- $GAE(\gamma,1)$ ： $\hat{A}_t = \sum_{l=1}^{\infty}\gamma^l r_{t+l} - V(s_t)$ ，右边第一项的期望就是 $Q^\pi$ （无偏地），后面相当于是个baseline，因此它不管 $V(s_t)$ 估计的准不准，都是 $\gamma$ -*just*的。
- $GAE(\gamma,0)$ ： $\hat{A}_t = \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$ ，仍然可以把最后一项看做baseline，但是当 $V(s_t)$ 估计地不准的时候，前面两项的期望就和 $Q^\pi$ 不一致了；因此仅当 $V = V^\pi$ 时，它才是 $\gamma$ -*just*的。
- 个人理解，要想方差更小，就需要 $\gamma$ 和 $\lambda$ 较小，因为它们较小的时候会更多地考虑较近的奖励而降低了很远的奖励的影响。但是较小的 $\gamma$ 和 $\lambda$ 都会引入额外的偏差；其中 $\gamma$ 控制了 $g$ 到 $g^\gamma$ 之间的偏差， $\lambda$ 控制了 $\hat{A}_t^{GAE(\gamma,\lambda)}$ 到 $A^{\pi,\gamma}$ 之间的偏差。

**4. 与reward shaping的关系**

在专栏前面的文章里面讲了potential-based reward shaping

$$\tilde{r}(s,a,s') = r(s,a,s') + \gamma\Phi(s') - \Phi(s),$$

一个自然的想法就是使用估计到的 $V(s)$ 来作为这个势能，同时为了避免很远的奖励带来的噪声，加上一个更快的衰减，自然就得到了前面定义的GAE

$$\sum_{l=0}^{\infty}(\gamma\lambda)^l\tilde{r}(s_{t+l},a_t,s_{t+l+1}) = \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}^V = \hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)}.$$

由此可以看出，GAE是就是一种reward shaping的应用，用来估计一个更instructive的value function。

**5. 其他**

后面的实验作者就是把GAE用在了TRPO上面，事实上GAE版本的TRPO和PPO已经是baselines里面的标准版本了。

---

Ps. 在草稿箱里面存了好长时间，终于写出来了。

编辑于 2019-03-14

机器学习 技术 强化学习 (Reinforcement Learning)



**文章被以下专栏收录**