VARIANCE REDUCTION FOR POLICY GRADIENT WITH ACTION-DEPENDENT FACTORIZED BASELINES

Cathy Wu[1], **Aravind Rajeswaran**[2], **Yan Duan**[13], **Vikash Kumar**[23], **Alexandre M Bayen**[14], **Sham Kakade**[2], **Igor Mordatch**[3], **Pieter Abbeel**[13]
cathywu@eecs.berkeley.edu, aravraj@cs.washington.edu, rockyduan@eecs.berkeley.edu, vikash@cs.washington.edu, bayen@berkeley.edu, sham@cs.washington.edu, igor.mordatch@gmail.com, pabbeel@cs.berkeley.edu
[1] Department of EECS, UC Berkeley
[2] Department of CSE, University of Washington
[3] OpenAI
[4] Institute for Transportation Studies, UC Berkeley

【强化学习 46】Action-dependent Baseline

张楚珩
清华大学 交叉信息院博士在读

原文传送门

Wu, Cathy, et al. "Variance reduction for policy gradient with action-dependent factorized baselines." arXiv preprint arXiv:1803.07246 (2018). （ICLR 2018）

特色

Baseline 是 policy gradient 类方法的一个重要的减小方差的手段，这里针对行动可以拆分为若干个条件独立部分的情形，提出了更进一步减小方差的方法。该方法仍然可以保持bias-free。我主要是想随便找篇文章看看 policy gradient 类方法应该如何分析方差。

过程

1. baseline不改变策略梯度的期望

我们先来看只依赖于状态的baseline（通常我们所说的baseline）

$$\nabla_\theta \mathbb{E}_x[f(x)] = \nabla_\theta \int p_\theta(x) f(x) dx = \int p_\theta(x) \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} f(x) dx$$
$$= \int p_\theta(x) \nabla_\theta \log p_\theta(x) f(x) dx = \mathbb{E}_x \left[\nabla_\theta \log p_\theta(x) f(x)\right]. \quad (1)$$

如果 $f(x)$ 与 x 无关，那么等式左边为零，这就是baseline不改变策略梯度期望的原因。

$$\mathbb{E}_{a_t}\left[\nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)\right] = \nabla_\theta \mathbb{E}_{a_t}\left[b(s_t)\right] = 0 \quad (5)$$

2. 最优baseline

baseline不改变策略梯度的期望，但是能够改变它的方差，我们期望能够找到一个最优的baseline

使得方差最小。

策略梯度写为

$$\nabla_\theta \eta(\pi_\theta) := \mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\left(\hat{Q}(s_t,a_t) - b(s_t)\right)\right] \tag{19}$$

把期望里面的梯度看做一个随机变量

$$g := \nabla_\theta \log \pi_\theta(a_t|s_t)\left(\hat{Q}(s_t,a_t) - b(s_t)\right), \quad a_t \sim \pi_\theta(a_t|s_t), s_t \sim \rho_\pi(s_t) \tag{20}$$

令 $g = g_1 - g_2$ ，分别是含 $\hat{Q}$ 和含 的两项。

策略梯度的方差可以写为（文中写漏了一项，还需要加上 $Var(g_1)$ ，即本身加baseline之前的方差，不过后面求导就没了，不影响；另外期望下标 ρ_π 代表的是 $s_t \sim \rho_\pi$ ，因此 $b(s_t)$ 应该写在期望里面）

$$\begin{aligned}
\mathrm{Var}(g) &= \mathbb{E}_{\rho_\pi,\pi}\left[(g - \mathbb{E}_{\rho_\pi,\pi}[g])^T(g - \mathbb{E}_{\rho_\pi,\pi}[g])\right] && (21)\\
&= \mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\right] b(s_t)^2 && (22)\\
&\quad - 2\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t,a_t)\right] b(s_t) && (23)
\end{aligned}$$

令方差的导数为零，可以求导最优的baseline。（上面的方差是对于所有 s_t 求的期望，每一项是一个二次型，因此对于任意一个 s_t ，都要使得该二次型最小，即导数为零；由此，下面公式中期望下标 ρ_π 应该去掉）

$$\begin{aligned}
&\frac{\partial}{\partial b}\left[\mathrm{Var}(g)\right] = 0 && (24)\\
&\quad = 2\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\right] b(s_t) && (25)\\
&\quad - 2\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t,a_t)\right] && (26)
\end{aligned}$$

$$\implies b^*(s_t) = \frac{\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t,a_t)\right]}{\mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)^T \nabla_\theta \log \pi_\theta(a_t|s_t)\right]} \tag{27}$$

在平常使用中，我们并不使用这个计算比较困难的baseline，而是使用 $\mathbb{E}_\pi[\hat{Q}(s_t,a_t)] = V(s_t)$ 。这做了 $\nabla_\theta \log \pi_\theta(a_t|s_t)$ 与 $\hat{Q}(s_t,a_t)$ 相互独立的假设（注意到它们都含有随机变量 a_t ）。

3. Action-dependent baseline

首先观察到很多情况下，action存在这样的结构

$$\pi_\theta(a_t|s_t) = \prod_{i=1}^m \pi_\theta(a_t^i|s_t)$$

- 连续控制的时候，常使用协方差矩阵仅有对角项的多元高斯分布，这时候action的各个维度就可以写为这样的形式；
- 多智能体强化学习里，如果各个智能体的执行是非中心化的，那么各个智能体之间的行动也可以写成这样的形式；

在这种结构下，可以把梯度拆成多个部分的求和

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{Q}(s_t,a_t)\right] = \mathbb{E}_{\rho_\pi,\pi}\left[\sum_{i=1}^{m} \nabla_\theta \log \pi_\theta(a_t^i|s_t)\hat{Q}(s_t,a_t)\right] \quad (6)$$

各个部分使用不同的baseline，规定每个部分的baseline只需要不和action对应的这部分有关就可以，即规定 $b_i(s_t, a_t^{-i})$ ，其中 a_t^{-i} 表示行动中不含 a_t^i 的其他部分。这是因为

$$\mathbb{E}_{a_t}\left[\nabla_\theta \log \pi_\theta(a_t^i|s_t) b_i(s_t, a_t^{-i})\right] = \mathbb{E}_{a_t^{-i}}\left[\nabla_\theta \mathbb{E}_{a_t^i}\left[b_i(s_t, a_t^{-i})\right]\right] = 0 \quad (7)$$

由此，我们可以得到含有action-dependent baseline的策略梯度

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{\rho_\pi,\pi}\left[\sum_{i=1}^{m} \nabla_\theta \log \pi_\theta(a_t^i|s_t)\left(\hat{Q}(s_t,a_t) - b_i(s_t, a_t^{-i})\right)\right] \quad (8)$$

4. Optimal action-dependent baseline

最优的意思就是方差最小，分析方式和前面类似，不过为了抵消掉交叉项，需要做如下假设

$$\nabla_\theta \log \pi_\theta(a_t^i|s_t)^T \nabla_\theta \log \pi_\theta(a_t^j|s_t) \equiv z_i^T z_j = 0, \quad \forall i \neq j \quad (11)$$

该假设说明行动中的不同部分被不同部分的策略参数影响。

策略梯度可以看做如下若干部分的梯度求和

$$\nabla \eta_i(\pi_\theta) := \mathbb{E}_{\rho_\pi,\pi}\left[\nabla_\theta \log \pi_\theta(a_t^i|s_t)\left(\hat{Q}(s_t,a_t) - b_i(s_t, a_t^{-i})\right)\right]. \quad (28)$$

各个部分的随机梯度向量

$$g_i := \nabla_\theta \log \pi_\theta(a_t^i|s_t)\left(\hat{Q}(s_t,a_t) - b_i(s_t, a_t^{-i})\right), \quad a_t \sim \pi_\theta(a_t|s_t), s_t \sim \rho_\pi(s_t), \quad (29)$$

把方差写出来

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{m} g_i\right) &= \sum_i \mathrm{Var}(g_i) + \sum_i \sum_{j\neq i} \mathrm{Cov}(g_i, g_j) && (32)\\
&= \sum_i \mathrm{Var}(g_i) + \sum_i \sum_{j\neq i} \mathbb{E}_{\rho_\pi,\pi}\left[g_i^T g_j\right] - \mathbb{E}_{\rho_\pi,\pi}\left[g_i\right]^T \mathbb{E}_{\rho_\pi,\pi}\left[g_j\right] && (33)\\
&= \sum_i \mathrm{Var}(g_i) + 0 - \sum_i \sum_{j\neq i} \mathbb{E}_{\rho_\pi,\pi}\left[g_i\right]^T \mathbb{E}_{\rho_\pi,\pi}\left[g_j\right] \quad \text{(by Equation (31))} && (34)\\
&= \sum_i \mathrm{Var}(g_i) - \sum_i \sum_{j\neq i} M_{ij} \quad \text{(by score function estimator)} && (35)
\end{aligned}$$

其中

$$M_{ij} := \mathbb{E}_{\rho_\pi,\pi}\left[z_i\hat{Q}(s_t,a_t)\right]^T \mathbb{E}_{\rho_\pi,\pi}\left[z_j\hat{Q}(s_t,a_t)\right] \quad M = \sum_i \sum_j M_{ij}$$

(33)里面第二项由于前面的假设消掉了， M 和baseline无关，后面我们要对于baseline求导数，因此这一项也没啥用。由此，我们可以看出根据action分成若干部分之后，其方差也是若干部分简单加和。

我们来看单一 $Var(g_i)$ 一项等于什么。

$$\begin{aligned}
\operatorname{Var}(g_i) &= \mathbb{E}_{\rho_\pi,\pi}\left[z_i^T z_i\left(\hat{Q}(s_t,a_t) - b_i(s_t,a_t^{-i})\right)^2\right] && (36)\\
&\quad - \mathbb{E}_{\rho_\pi,\pi}\left[z_i\left(\hat{Q}(s_t,a_t) - b_i(s_t,a_t^{-i})\right)\right]^T \mathbb{E}_{\rho_\pi,\pi}\left[z_i\left(\hat{Q}(s_t,a_t) - b_i(s_t,a_t^{-i})\right)\right]\\
&= \mathbb{E}_{\rho_\pi,\pi}\left[z_i^T z_i\left(\hat{Q}(s_t,a_t)^2 - 2b_i(s_t,a_t^{-i})\hat{Q}(s_t,a_t) + b_i(s_t,a_t^{-i})\right)^2\right] && (37)\\
&\quad - \mathbb{E}_{\rho_\pi,\pi}\left[z_i\left(\hat{Q}(s_t,a_t)\right)\right]^T \mathbb{E}_{\rho_\pi,\pi}\left[z_i\left(\hat{Q}(s_t,a_t)\right)\right]\\
&= \mathbb{E}_{\rho_\pi,\pi}\left[z_i^T z_i\hat{Q}(s_t,a_t)^2\right] && (38)\\
&\quad + \mathbb{E}_{\rho_\pi,a_t^{-i}}\left[-2b_i(s_t,a_t^{-i})\mathbb{E}_{a_t^i}\left[z_i^T z_i\hat{Q}(s_t,a_t)\right] + b_i(s_t,a_t^{-i})^2\mathbb{E}_{a_t^i}\left[z_i^T z_i\right]\right] - M_{ii}
\end{aligned}$$

对其求导为零，可以得到最优的baseline。

$$b_i^*(s_t,a_t^{-i}) = \frac{\mathbb{E}_{a_t^i}\left[\nabla_\theta \log\pi_\theta(a_t^i|s_t)^T\nabla_\theta\log\pi_\theta(a_t^i|s_t)\hat{Q}(s_t,a_t)\right]}{\mathbb{E}_{a_t^i}\left[\nabla_\theta\log\pi_\theta(a_t^i|s_t)^T\nabla_\theta\log\pi_\theta(a_t^i|s_t)\right]} \quad (41)$$

5. 选择一个方便计算的action-dependent baseline

和state-dependent baseline类似，我们推导到最优的baseline，但是出于practical的考虑，一般使用 $\mathbb{E}_\pi[\hat{Q}(s_t,a_t)] = V(s_t)$ 作baseline。文章提出了几种practical的action-dependent baselien。

- **Marginalized Q baseline**：即使用 $\mathbb{E}_{a^i}[\hat{Q}(s_t,a_t)]$ 作为baseline，其好处是它对于方差的减小程度近乎最优的action-dependent baseline。实际中可以估计一个 $Q(s_t,a_t)$ 然后再做marginalization，得到相应的 m 个baseline。

- **Monte Carlo Q baseline**：估计到一个 $Q(s_t, a_t)$ 之后，要做marginalization需要通过蒙特卡洛采样得到，即

$$b_i(s_t, a_t^{-i}) = \frac{1}{M}\sum_{j=0}^{M} Q_{\pi_\theta}(s_t, (a_t^{-i}, \alpha_j)) \tag{17}$$

- **Mean marginalized Q baseline**：如果估计到的 $Q(s_t, a_t)$ 是一个神经网络表示的话，蒙特卡洛采样要求需要把该神经网络向前传播若干次，这比较耗费计算量；这里把它用均值代替，这样只需要正向计算一次了

$$b_i(s_t, a_t^{-i}) = Q_{\pi_\theta}(s_t, (a_t^{-i}, \bar{a}_t^i)) \tag{18}$$

where $\bar{a}_t^i = \mathbb{E}_{\pi_\theta}\left[a_t^i\right]$ is the average action for coordinate i.