A Natural Policy Gradient

Sham Kakade
Gatsby Computational Neuroscience Unit
17 Queen Square, London, UK WC1N 3AR
http://www.gatsby.ucl.ac.uk
sham@gatsby.ucl.ac.uk

【强化学习 53】Natural PG

张楚珩
清华大学 交叉信息院博士在读

2002年的老文章了，natural policy gradient。

原文传送门

Kakade, Sham M. "A natural policy gradient." Advances in neural information processing systems. 2002.

特色

Natural gradient用到强化学习策略梯度方法上，搭配TRPO食用效果更好。

过程

1. 自然梯度

普通的policy gradient可以写成

$$\nabla\eta(\theta) = \sum_{s,a} \rho^{\pi}(s) \nabla\pi(a; s, \theta) Q^{\pi}(s, a)$$

natural policy gradient的意思是约束策略的行动概率变化小于一定数值下，使得目标函数变化最大的方向（steepest ascent）。自然梯度方向可以表示为 $d\theta$ ，它在 $|d\theta|^2 := d\theta^T G(\theta) d\theta \leq \delta$ 约束下，使得 $\eta(\theta + d\theta)$ 最大。在普通的policy gradient中即认为了 $G(\theta) = I$ 。

但其实策略参数 θ 带给策略概率密度 $\pi(a|s,\theta)$ 并不是均匀的，更好的方法是使用Fisher information matrix来作为 $G(\theta)$ 。

$$F_s(\theta) \equiv E_{\pi(a;s,\theta)}[\frac{\partial \log \pi(a; s, \theta)}{\partial \theta_i} \frac{\partial \log \pi(a; s, \theta)}{\partial \theta_j}],$$

相应的策略梯度可以定义为

$$\tilde{\nabla}\eta(\theta) \equiv F(\theta)^{-1}\nabla\eta(\theta)\,.$$

其中

$$F(\theta) \equiv E_{\rho^\pi(s)}[F_s(\theta)]$$

2. 自然梯度的优势

普通的策略梯度每一步迭代的策略参数变化方向是选择一个比当前策略好一些的策略；而自然梯度的更新方向是greedy policy。文章对此进行了理论说明。

考虑一个compatible function approximation $f^\pi(s,a|\omega)$ ，其定义如下（回顾一下，compatible代表的含义是用这个估计的价值函数替换策略梯度中的真实价值函数，即使有parametrization带来的误差，得到的策略梯度也是准确的）

$$\psi(s,a)^\pi = \nabla \log \pi(a;s,\theta), \quad f^\pi(s,a;\omega) = \omega^T\psi^\pi(s,a)$$

定义 ω 为价值函数网络的最优参数

$\tilde{\omega}$ minimize the squared error $\epsilon(\omega,\pi) \equiv \sum_{s,a}\rho^\pi(s)\pi(a;s,\theta)(f^\pi(s,a;\omega)-Q^\pi(s,a))^2$

这样如果规定一个特殊的策略表示，可以看到，***如果往自然梯度方向一直更新，最终会得到相对于 $f^\pi(s,a|\omega)$ 的greedy policy***。

Theorem 2. *For $\pi(a;s,\theta) \propto \exp(\theta^T\phi_{sa})$, assume that $\tilde{\nabla}\eta(\theta)$ is non-zero and that $\tilde{\omega}$ minimizes the approximation error. Let $\pi_\infty(a;s) = \lim_{\alpha\to\infty}\pi(a;s,\theta+\alpha\tilde{\nabla}\eta(\theta))$. Then $\pi_\infty(a;s) \neq 0$ if and only if $a \in argmax_{a'} f^\pi(s,a';\tilde{\omega})$.*

更一般地，如果不考虑这样特殊的策略表示，***自然梯度方向也是把策略变为greedy policy的方向。***

Theorem 3. *Assume that $\tilde{\omega}$ minimizes the approximation error and let the update to the parameter be $\theta' = \theta + \alpha\tilde{\nabla}\eta(\theta)$. Then*

$$\pi(a;s,\theta') = \pi(a;s,\theta)(1+\alpha f^\pi(s,a;\tilde{\omega})) + O(\alpha^2)$$

这些理论如果不关心的话，实际观察到的有什么优势呢？

- 收敛速度更快，因为直接朝着相对所估计到价值函数greedy policy的方向更新，不走弯路；
- 更加满足[1]里面的policy improvement theorem对于策略变化不要太大的要求，它要求每一步策略更新策略空间中变化不要太大（注意不是参数空间）；

- 不会因为初始策略离最优策略过远而收敛特别慢；

Fisher Information Matrix

一堆可观测变量 x，一个隐变量 θ，隐变量决定了可观测变量的概率分布 $f(X|\theta)$。Fisher information讲的就是观察到一些变量 X 的时候，对于确定 θ 提供了多大的信息量。假设 θ 变化一点点，对应可观测变量的概率密度变化特别多（即可观测变量 x 的分布很“窄”），那么一个观测样本带给我们的信息量就很大，因为我们可以很确信 θ 的值。

由此我们想到，可以使用loglikelihood的导数来定义观测样本带给我们的信息量。如果定义Fisher information matrix，

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2 \middle| \theta\right] = \int \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)\,dx,$$

那么 $d\theta^T\mathcal{I}(\theta)d\theta$ 可以理解为在 θ 附近，参数对于样本的敏感程度；如果限定 $d\theta^T\mathcal{I}(\theta)d\theta \leq \delta$ 则代表希望在 θ 附近更新的一小步不会特别大地影响样本上的loglikelihood。

参考文献

[1] Kakade, Sham, and John Langford. "Approximately optimal approximate reinforcement learning."*ICML*. Vol. 2. 2002.

发布于 2019-04-10

强化学习 (Reinforcement Learning)