Notes on Importance Sampling and Policy Gradient

Nan Jiang

October 25, 2018

【强化学习理论 66】StatisticalRL 9

张楚珩
清华大学 交叉信息院博士在读

这是UIUC姜楠老师开设的CS598统计强化学习（理论）课程的第六讲，这一讲的主要内容是Fitted Q-iteration。

原文传送门

CS598 Note6
nanjiang.cs.illinois.edu

一、Importance sampling

1. 估计期望

Consider the problem of estimating $\mathbb{E}_{x\sim p}[f(x)]$ for distribution $p \in \Delta(\mathcal{X})$ and function $f : \mathcal{X} \to \mathbb{R}$. If we can sample $x \sim p$, the standard Monte-Carlo estimate is $f(x)$, and averaging such estimates over multiple i.i.d. samples of x will give us an accurate estimate of $\mathbb{E}_{x\sim p}[f(x)]$. This is particularly useful if it is easy to sample from p but difficult to calculate the integral in $\mathbb{E}_{x\sim p}[f(x)]$.

Now what if we cannot sample from p, but have access to $x \sim q$ for some other distribution $q \in \Delta(\mathcal{X})$? It turns out that, if p is fully supported on q, that is, for all $x \in \mathcal{X}$ where $p(x) > 0$ we have $q(x) > 0$, then the following *importance weighted* estimator also gives an unbiased estimate of $\mathbb{E}_{x\sim p}[f(x)]$:

$$\frac{p(x)}{q(x)} f(x). \tag{1}$$

To verify unbiasedness:

$$\mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)} f(x)\right] = \sum_{x\in\mathcal{X}} q(x)\frac{p(x)}{q(x)} f(x) = \sum_{x\in\mathcal{X}} p(x) f(x) = \mathbb{E}_{x\sim p}[f(x)].$$

$p(x)/q(x)$ has many names: importance weight, importance ratio, or inverse propensity score (IPS). A useful property of importance ratio to keep in mind is that

$$\mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}\right] = 1.$$

当要估计 $\mathbb{E}_{x\sim p}[f(x)]$ 的时候，如果可以从分布 p 中采样，那么可以直接把样本上的平均函数值作为该期望的无偏估计；如果只能从另外一个分布 q 中采样，那么可以使用一个调整系数 $\frac{p(x)}{q(x)}$ 来使其仍然是一个无偏估计。这样做的要求是 p is fully supported on q（具体定义见截图）。这种方法叫做 importance sampling （IS）。

2. 考虑单步情形

构造一个无偏估计

考虑单步的情形，即contextual bandit问题。同时假设状态分布为 μ ，奖励范围在 $[0,1]$ 之间。假设有使用behavior policy $x\sim\mu, a\sim\pi_b(x)$ 采样到的的数据集 $\{(x,a,r)\}$ ，目标是使用该数据集估计target policy下的性能 $v^\pi := \mathbb{E}[r|a\sim\pi]$ 。

方式很简单，就是选用如下estimate

$$\rho\, r, \quad \text{where } \rho = \frac{\pi(a|x)}{\pi_b(a|x)}. \tag{3}$$

它是unbiased的，根据之前的结论，使用importance weight是unbiased，即

$$\mathbb{E}[r|a\sim\pi] = \mathbb{E}_{(x,a,r)\sim p}[r] = \mathbb{E}_{(x,a,r)\sim q}\left[\frac{p(x,a,r)}{q(x,a,r)}\, r\right].$$

而该importance weight可以仅使用behavior policy和target policy来计算出来

$$\frac{p(x,a,r)}{q(x,a,r)} = \frac{\mu(x)\pi(a|x)R(r|x,a)}{\mu(x)\pi_b(a|x)R(r|x,a)} = \frac{\pi(a|x)}{\pi_b(a|x)} = \rho.$$

方差分析

下面来分析这种方法的方差，为了便于分析，假设behavior policy是对于各个action均匀采样，而target policy是一个确定性策略，同时奖励是一个确定性的常数。那么其方差可以较为容易地推导出来

$$
\begin{aligned}
&\mathbb{V}[\rho r|a\sim U]=r^2\mathbb{V}[\rho|a\sim U]\\
&=r^2(\mathbb{E}[\rho^2|a\sim U]-(\mathbb{E}[\rho|a\sim U])^2)\\
&=r^2(\mathbb{E}[\rho^2|a\sim U]-1) && \text{(the mean of }\rho\text{ is always 1)}\\
&=r^2\left(\mathbb{E}\left[\frac{\mathbb{I}[a=\pi(x)]}{1/K^2}\,\middle|\,a\sim U\right]-1\right)\\
&=r^2(K-1).
\end{aligned}
$$

其中 $K:=|\mathcal{A}|$ 是可能动作的个数。观察到，当 behavior policy = target policy 时，方差应该为零，而加上了 IS 之后产生了更大的方差。可以想象成只有采样到的样本选择的动作和 target policy 选择的动作相同的时候，该样本才能用上，因此会有了一个产生与 K 有关的方差。

当奖励不是确定性的常数而是在 $[0,1]$ 之间的随机变量时，方差上界为

$$\mathbb{V}[\rho r|a\sim U]\leq\mathbb{E}[\rho^2r^2|a\sim U]\leq\mathbb{E}[\rho^2|a\sim U]=K.$$

有限样本分析

下面考虑使用有限数目的样本能够使用 ρr 估计 v^π 到何种精度。使用Bernstein不等式相比于Hoeffding不等式能够得到更紧的上界。考虑 $r\in[0,1],\rho r\in[0,K],\mathbb{V}[\rho r]\leq K$ ，可以得到上界为

$$\sqrt{\frac{2K}{n}\ln\frac{2}{\delta}}+\frac{2K}{3n}\ln\frac{2}{\delta}.$$

其中使用到的Bernstein不等式可以表述为

Theorem *Let $x_1,x_2,\ldots,x_n$ be independent bounded random variables such that $Ex_i=0$ and $|x_i|\leq\varsigma$ with probability 1 and let $\sigma^2=\frac{1}{n}\sum_{i=1}^n Var\{x_i\}$ Then for any $a>0$ we have*

$$P\left(\frac{1}{n}\sum_{i=1}^n x_i\geq\epsilon\right)\leq e^{-\frac{n\epsilon^2}{2\sigma^2+2\varsigma\epsilon/3}}$$

Weighted importance sampling

回到奖励是确定性常数以及target policy为确定性策略的情形，在这种情况下，如果只使用和target policy一致的样本（subsamples），那么应该该估计的方差应该为零，但是前面介绍的 IS 方法得到的方差却不为零。IS方法的估计为

$$\frac{1}{n}\sum_{i=1}^n\frac{\mathbb{I}[a_i=\pi(x_i)]}{1/K}r_i=\frac{1}{n/K}\sum_{i:a_i=\pi(x_i)}r_i.$$

如果分母部分为 $|\{i:a_i=\pi(x_i)\}|$ ，那么就相当于在subsample上做估计，得到的估计方差为零；但是现

在这里分母却不是 $|\{i : a_i = \pi(x_i)\}|$ ，而是它的期望 n/K ，这导致了估计的方差大于零。weighted importance sampling（WIS）则另外构造了下面这样的估计方法，可以在此情况下方差为零，起到了减小方差的目的。

$$\frac{1}{\sum_{i=1}^{n} \rho_i} \sum_{i=1}^{n} \rho_i \, r_i. \tag{4}$$

WIS的缺点是它是biased的，即样本有限时，期望不等于真实数值；不过它是consistent的，即当样本数目趋向于无穷的时候，其分布会趋向于真实分布。

进一步减小方差

根据前面的结果，方差和奖励的平方有关，由此想到可以先把奖励减去某个常数 c，然后求导相应的估计之后，再加上这个常数，由此能够得到更小的方差 $(r-c)^2(K-1)$ 。

进一步推广可以得到 doubly robust (DR) estimate，即使用一个估计来的 $\hat{Q}(x,a)$ 来替代前面根据先验知识设置的 c ，并且期望 $\hat{Q}(x,a) \approx \mathbb{E}_{a'\sim\pi}[r|x,a]$ 。

$$\mathbb{E}_{a'\sim\pi}[\hat{Q}(x,a')] + \rho\left(r - \hat{Q}(x,a)\right). \tag{5}$$

3. 考虑多步情形

无偏估计

类似地，对于多步情形（即标准的RL设定），可以得到无偏的估计

$$\begin{aligned}
v^\pi &= \mathbb{E}\left[\sum_{h=1}^{H} \gamma^{h-1} r_h \,\middle|\, a_{1:H} \sim \pi\right] = \mathbb{E}_{\tau\sim p}\left[\sum_{h=1}^{H} \gamma^{h-1} r_h\right] = \mathbb{E}_{\tau\sim q}\left[\frac{p(\tau)}{q(\tau)} \sum_{h=1}^{H} r_h\right] \\
&= \mathbb{E}_{\tau\sim q}\left[\frac{\mu(s_1)\pi(a_1|s_1)R(r_1|s_1,a_1)P(s_2|s_1,a_1)\cdots\pi(a_H|s_H)R(r_H|s_H,a_H)}{\mu(s_1)\pi_b(a_1|s_1)R(r_1|s_1,a_1)P(s_2|s_1,a_1)\cdots\pi_b(a_H|s_H)R(r_H|s_H,a_H)} \sum_{h=1}^{H} \gamma^{h-1} r_h\right] \\
&= \mathbb{E}_{\tau\sim q}\left[\frac{\pi(a_1|s_1)\cdots\pi(a_H|s_H)}{\pi_b(a_1|s_1)\cdots\pi_b(a_H|s_H)} \sum_{h=1}^{H} \gamma^{h-1} r_h\right] = \mathbb{E}\left[\frac{\pi(a_1|s_1)\cdots\pi(a_H|s_H)}{\pi_b(a_1|s_1)\cdots\pi_b(a_H|s_H)} \sum_{h=1}^{H} \gamma^{h-1} r_h\right]
\end{aligned}$$

由此可以得到**per-trajectory IS estimator**

So the expression in the bracket is an unbiased estimate of v^π. Let $\rho_h := \pi(a_h|s_h)/\pi_b(a_h|s_h)$ and $\rho_{1:h}$ be a shorthand for $\prod_{h'=1}^{h} \rho_{h'}$, the **per-trajectory** IS estimator is [2, 3]:

$$\rho_{1:H} \sum_{h=1}^{H} \gamma^{h-1} r_h. \tag{6}$$

观察到第h步之后的样本不会影响到第h步的奖励，因此，对于第h步来说，计算IS的时候，可以把h步之后都去掉。得到**per-step IS estimator**

$$\sum_{h=1}^{H} \gamma^{h-1} \rho_{1:h} r_h. \tag{7}$$

它还可以写成递归的形式，即

$$v_{H-h+1} := \rho_h (r_h + \gamma v_{H-h}). \tag{8}$$

类似地，可以得到**DR estimator**

$$v_{H-h+1}^{DR} := \mathbb{E}_{a\sim\pi}[\hat{Q}^{\pi}(s_h, a)] + \rho_h \left(r_h + \gamma v_{H-h}^{DR} - \hat{Q}^{\pi}(s_h, a_h) \right) \tag{9}$$

通过递归的形式，可以顺着递归地证明上述estimator是都是无偏的。

方差分析

下面分析DR estimator的方差

Variance of per-step IS The variance of Eq.(7) also satisifies an interesting recursion, which has important implications outside off-policy evaluation. Let $\mathbb{V}_h[\cdot]$ and $\mathbb{E}_h[\cdot]$ denote conditional variance and expectation, respectively, conditioned on $s_1, a_1, r_1, \ldots, s_{h-1}, a_{h-1}, r_{h-1}$. For simplicity assume reward is a deterministic function of state and action, then

$$\begin{aligned}
& VV_h[v_{H-h+1}] = \mathbb{E}_h[v_{H-h+1}^2] - (\mathbb{E}_h[v_{H-h+1}])^2 \\
&= \mathbb{E}_h[v_{H-h+1}^2] - (\mathbb{E}_h[V^{\pi}(s_h)])^2 \qquad (V^{\pi}(s_h) = \mathbb{E}_h[v_{H-h+1} \mid s_h]) \\
&= \mathbb{E}_h[(\rho_h Q^{\pi}(s_h, a_h) + \rho_h (r_h + \gamma v_{H-h} - Q^{\pi}(s_h, a_h)))^2] - (\mathbb{E}_h[V^{\pi}(s_h)])^2 \\
&= \mathbb{E}_h[(\rho_h Q^{\pi}(s_h, a_h))^2] + \mathbb{E}_h[\rho_h^2 (r_h + \gamma v_{H-h} - Q^{\pi}(s_h, a_h)))^2] - (\mathbb{E}_h[V^{\pi}(s_h)])^2 \\
&= \mathbb{E}_h[(V^{\pi}(s_h) + \rho_h Q^{\pi}(s_h, a_h) - V^{\pi}(s_h))^2] + \gamma^2 \mathbb{E}_h[\rho_h^2 (v_{H-h} - V^{\pi}(s_{h+1}))^2] - (\mathbb{E}_h[V^{\pi}(s_h)])^2 \\
&= \mathbb{E}_h[V^{\pi}(s_h))^2] + \mathbb{E}_h[\mathbb{V}_h[\rho_h Q^{\pi}(s_h, a_h) \mid s_h]] + \gamma^2 \mathbb{E}_h[\rho_h^2 \mathbb{V}_{h+1}[v_{H-h}]] - (\mathbb{E}_h[V^{\pi}(s_h)])^2 \\
&= \mathbb{V}_h[V^{\pi}(s_h)] + \mathbb{E}_h[\mathbb{V}_h[\rho_h Q^{\pi}(s_h, a_h) \mid s_h]] + \gamma^2 \mathbb{E}_h[\rho_h^2 \mathbb{V}_{h+1}[v_{H-h}]].
\end{aligned}$$

考虑on-policy并且测量是确定性的情形，上式可以化简为如下形式，即Bellman equation for variance。

$$\mathbb{V}_h[v_{H-h+1}] = \mathbb{V}_h[V^{\pi}(s_h)] + \gamma^2 \mathbb{E}_h[\mathbb{V}_{h+1}[v_{H-h}]].$$

二、策略梯度

略。主要讲了策略梯度的推导和使用baseline来减小variance，本专栏的入门系列有讲，同时贴一下个人之前总结的笔记。

Define value functions and optimization objective as usual.

Definition 1.1 (State value function).

$$V^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty}\sum_{s'} p(s_t = s'|s_0 = s, \pi_\theta)\gamma^t r_t] \tag{1}$$

Definition 1.2 (Action value function).

$$Q^\pi(s,a) = \mathbb{E}[\sum_{t=0}^{\infty}\sum_{s'} p(s_t = s'|s_0 = s, a_0 = a, \pi_\theta)\gamma^t r_t] \tag{2}$$

Definition 1.3 (Optimization objective). *The ultimate goal of model-free reinforcement learning is to maximize*

$$\eta(\pi) := \mathbb{E}_{s_0}[V^\pi(s_0)] \tag{3}$$

1.1 Policy gradient

Theorem 1.1 (Policy gradient).

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}_{s\sim\rho_\pi, a\sim\pi}[\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s,a)]$$

where $\rho_\pi(s) = \mathbb{E}_{s_0}[\sum_{t=0}^{\infty}\gamma^t p(s_t = s|s_0,\pi)]$ *is the state visitation frequency.*

Proof. Take derivative and iteratively unroll.

$$\begin{aligned}
\nabla_\theta V^\pi(s_0) &= \nabla_\theta[\sum_{a_0}\pi_\theta(a_0|s_0)Q^\pi(s_0,a_0)] \\
&= \sum_{a_0}[\nabla_\theta\pi_\theta(a_0|s_0)Q^\pi(s_0,a_0) + \pi_\theta(a_0|s_0)\nabla_\theta Q^\pi(s_0,a_0)] \\
&= \sum_{a_0}[\nabla_\theta\pi_\theta(a_0|s_0)Q^\pi(s_0,a_0) + \pi_\theta(a_0|s_0)\nabla_\theta\sum_{s_1,r_0}p(s_1,r_0|s_0,a_0)(r+\gamma V^\pi(s_1))] \\
&= \sum_{a_0}[\nabla_\theta\pi_\theta(a_0|s_0)Q^\pi(s_0,a_0) + \pi_\theta(a_0|s_0)\sum_{s_1}\gamma p(s_1|s_0,a_0)\nabla_\theta V^\pi(s_1)] \\
&= \sum_{s\in\mathcal{S}}\sum_{t=0}^{\infty}\gamma^t p(s_t = s|s_0,\pi_\theta)\mathbb{E}_{a\sim\pi}[\nabla_\theta \log\pi_\theta(a|s)Q^\pi(s,a)]
\end{aligned}$$

$$\nabla_\theta \eta(\pi_\theta) = \sum_{s_0} \sum_{s \in \mathcal{S}} \sum_{t=0}^{\infty} p(s_0) \gamma^t p(s_t = s | s_0, \pi_\theta) \mathbb{E}_{a \sim \pi} [\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s,a)]$$
$$= \mathbb{E}_{s \sim \rho_\pi, a \sim \pi} [\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s,a)]$$

1.2 Policy gradient with baseline

Theorem 1.2 (Policy gradient with baseline)**.** *Baseline $b(s)$ dose not change expected value of policy gradient.*

$$\nabla_\theta \eta(\theta) = \mathbb{E}_{s \sim \rho_\pi, a \sim \pi} \left[\nabla_\theta \log \pi_\theta(a|s) \left(Q^\pi(s,a) - b(s) \right) \right] \tag{4}$$

Moreover, when the baseline is chosen to be

$$b^*(s) = \frac{\mathbb{E}_{a \sim \pi} [\log \pi_\theta(a|s)^T \log \pi_\theta(a|s) Q^\pi(s,a)]}{\mathbb{E}_{a \sim \pi} [\log \pi_\theta(a|s)^T \log \pi_\theta(a|s)]} \tag{5}$$

the variance of policy gradient is minimized.

Proof. In terms of expected value, we only need to notice that

$$\mathbb{E}_a [\nabla_\theta \log \pi_\theta(a|s) b(s)] = \sum_a [\nabla_\theta \pi_\theta(a|s)] b(s) = \nabla_\theta (\sum_a \pi_\theta(a|s)) b(s) = 0$$

In terms of the variance, let $g = \nabla_\theta \log \pi_\theta(a|s) \left(Q^\pi(s,a) - b(s) \right)$ and $g_0 = \nabla_\theta \log \pi_\theta(a|s) Q^\pi(s,a)$.

$$\begin{aligned} Var(g) =& \mathbb{E}_{s \sim \rho_\pi, a \sim \pi} [(g - \mathbb{E}_{s \sim \rho_\pi, a \sim \pi}[g])^T (g - \mathbb{E}_{s \sim \rho_\pi, a \sim \pi}[g])] \\ =& Var(g_0) - 2\mathbb{E}_{s \sim \rho_\pi, a \sim \pi} [\log \pi_\theta(a|s)^T \log \pi_\theta(a|s) Q^\pi(s,a)] b(s) \\ &+ \mathbb{E}_{s \sim \rho_\pi, a \sim \pi} [\log \pi_\theta(a|s)^T \log \pi_\theta(a|s)] b^2(s) \end{aligned}$$

Optimal baseline can be found to minimize the variance

$$b^*(s) = \frac{\mathbb{E}_{a \sim \pi} [\log \pi_\theta(a|s)^T \log \pi_\theta(a|s) Q^\pi(s,a)]}{\mathbb{E}_{a \sim \pi} [\log \pi_\theta(a|s)^T \log \pi_\theta(a|s)]} \tag{6}$$

□

For convenience, a popular baseline is chosen to be

$$b(s) = \mathbb{E}_{a\sim\pi} Q^{\pi}(s,a) = V^{\pi}(s) \tag{7}$$

or function approximated state value function

$$b(s) = V_{\phi}^{\pi}(s) \tag{8}$$

This is an unbiased estimate of policy gradient, though function approximated state value function is used for baseline.

$$\nabla_{\theta}\eta(\theta) = \mathbb{E}_{s\sim\rho_{\pi},a\sim\pi}\left[\nabla_{\theta}\log\pi_{\theta}(a|s)\left(Q^{\pi}(s,a) - V_{\phi}^{\pi}(s)\right)\right] \tag{9}$$

When the state value function is unbiased (e.g. learn from Monte Carlo return samples), the following policy gradient estimate is also unbiased, where s' is the next state following s and a.

$$\nabla_{\theta}\eta(\theta) = \mathbb{E}_{s\sim\rho_{\pi},a\sim\pi}\left[\nabla_{\theta}\log\pi_{\theta}(a|s)\left(r + \gamma V_{\phi}^{\pi}(s') - V_{\phi}^{\pi}(s)\right)\right] \tag{10}$$

发布于 2019-06-03