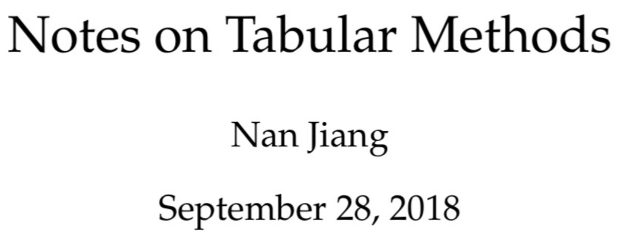


Notes on Tabular Methods

Nan Jiang

September 28, 2018

【强化学习理论 60】StatisticalRL 4

张楚珩
清华大学 交叉信息院博士在读

3 人赞同了该文章

这是UIUC姜楠老师开设的CS598统计强化学习（理论）课程的第三讲。

原文传送门

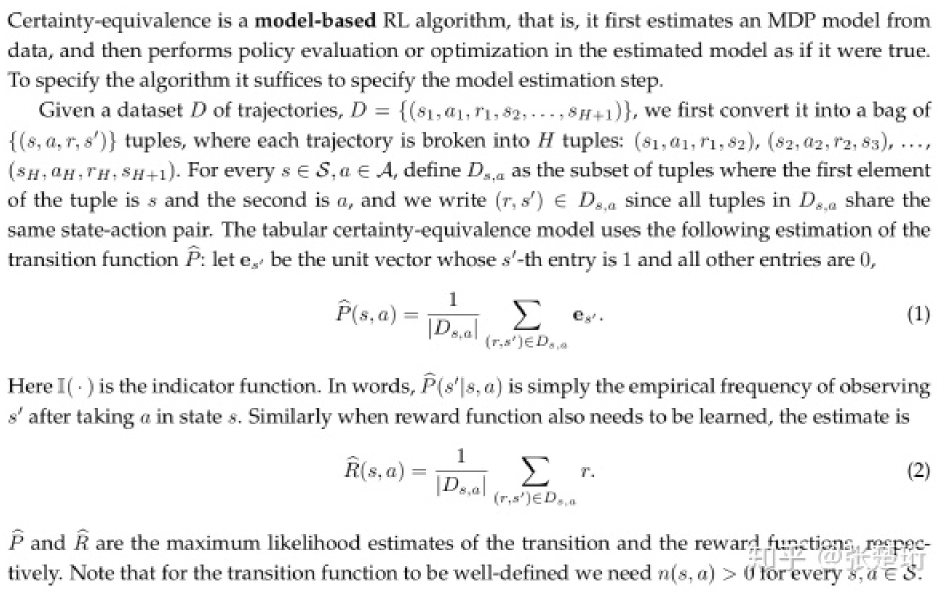
CS598 Note3
nanjiang.cs.illinois.edu

摘要

这一讲主要讲的是certainty-equivalence及其理论分析，这个名字我也是第一次听说，也没太懂这两个词和所讲的东西是什么关系。

MDP模型（主要是dynamics P 和reward R ）并不是直接知道的，我们需要先估计，再求解。那么估计过程中产生的误差会对最后求解到的最优策略性能产生多大的损失呢？考虑一个MDP，从每个 (s,a) 出发都采集 n 个样本，这样就能够对于MDP进行一个估计，如果要求估计到的MDP $\hat{M}$ 上最优策略性能不太差，那么至少需要多少采样？这就是本讲讨论的问题。

一、Certainty-equivalence的定义



Certainty-equivalence is a **model-based** RL algorithm, that is, it first estimates an MDP model from data, and then performs policy evaluation or optimization in the estimated model as if it were true. To specify the algorithm it suffices to specify the model estimation step.

Given a dataset D of trajectories, $D = \{(s_1, a_1, r_1, s_2, \ldots, s_{H+1})\}$, we first convert it into a bag of $\{(s, a, r, s')\}$ tuples, where each trajectory is broken into H tuples: (s_1, a_1, r_1, s_2), (s_2, a_2, r_2, s_3), $\ldots$, (s_H, a_H, r_H, s_{H+1}). For every $s \in \mathcal{S}, a \in \mathcal{A}$, define $D_{s,a}$ as the subset of tuples where the first element of the tuple is s and the second is a, and we write $(r, s') \in D_{s,a}$ since all tuples in $D_{s,a}$ share the same state-action pair. The tabular certainty-equivalence model uses the following estimation of the transition function $\widehat{P}$: let $\mathbf{e}_{s'}$ be the unit vector whose s'-th entry is 1 and all other entries are 0,

$$\widehat{P}(s,a) = \frac{1}{|D_{s,a}|} \sum_{(r,s') \in D_{s,a}} \mathbf{e}_{s'}. \tag{1}$$

Here $\mathbb{I}(\cdot)$ is the indicator function. In words, $\widehat{P}(s'|s,a)$ is simply the empirical frequency of observing s' after taking a in state s. Similarly when reward function also needs to be learned, the estimate is

$$\widehat{R}(s,a) = \frac{1}{|D_{s,a}|} \sum_{(r,s') \in D_{s,a}} r. \tag{2}$$

$\widehat{P}$ and $\widehat{R}$ are the maximum likelihood estimates of the transition and the reward functions, respectively. Note that for the transition function to be well-defined we need $n(s,a) > 0$ for every $s, a \in \mathcal{S}$.

即从每个 (s,a) 出发都采集 $|D_{s,a}|$ 个样本，并由此估计dynamics P 和reward R 。

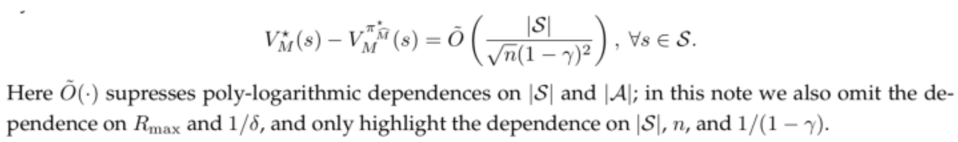
这里后面还比较了该方法和我们熟悉的value-based methods（比如Q-learning、SARSA等）。该方法空间复杂度更大 $O(|\mathcal{S}|^2|\mathcal{A}|)$ （需要存储每个 (s,a) 出发到每个 s' 的计数），而value-based methods空间复杂度更小 $O(|\mathcal{S}||\mathcal{A}|)$ （只需要存储每个 (s,a) 的价值函数）。但是该方法更加sample efficient（即本讲下面要说明的内容）。

二、Certainty-equivalence的分析

分析主要包括了三个由弱到强的结论，讲的就是如果要求估计到的MDP $\hat{M}$ 上最优策略性能不太差，那么至少需要多少采样 n 。

2.1. 结论一

我们关心的是估计到的MDP $\hat{M}$ 上的最优策略性能比真正MDP上最优性能差多少。先放第一个结论。

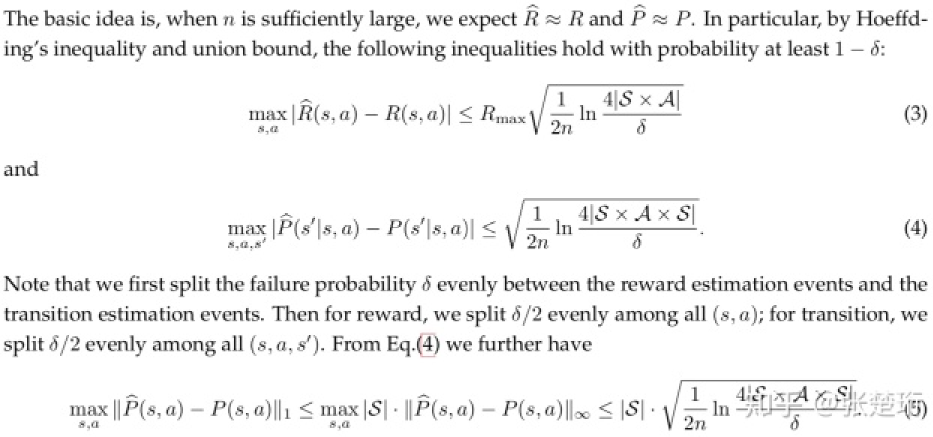


$$V_M^\star(s) - V_M^{\pi^\star_{\widehat{M}}}(s) = \tilde{O}\left(\frac{|\mathcal{S}|}{\sqrt{n}(1-\gamma)^2}\right), \ \forall s \in \mathcal{S}.$$

Here $\tilde{O}(\cdot)$ supresses poly-logarithmic dependences on $|\mathcal{S}|$ and $|\mathcal{A}|$; in this note we also omit the dependence on $R_{\max}$ and $1/\delta$, and only highlight the dependence on $|\mathcal{S}|$, n, and $1/(1-\gamma)$.

证明过程如下。

步骤一：先利用Hoeffding不等式得到对于dynamics P 和reward R 的误差bound。（确定模型估计误差）



The basic idea is, when n is sufficiently large, we expect $\widehat{R} \approx R$ and $\widehat{P} \approx P$. In particular, by Hoeffding's inequality and union bound, the following inequalities hold with probability at least $1-\delta$:

$$\max_{s,a} |\widehat{R}(s,a) - R(s,a)| \le R_{\max}\sqrt{\frac{1}{2n}\ln\frac{4|\mathcal{S}\times\mathcal{A}|}{\delta}} \tag{3}$$

and

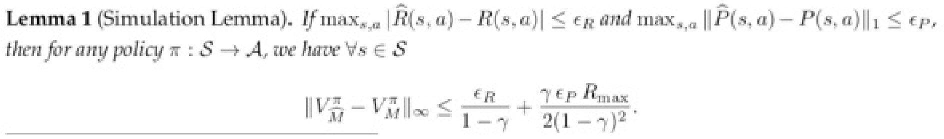
$$\max_{s,a,s'} |\widehat{P}(s'|s,a) - P(s'|s,a)| \le \sqrt{\frac{1}{2n}\ln\frac{4|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}{\delta}}. \tag{4}$$

Note that we first split the failure probability δ evenly between the reward estimation events and the transition estimation events. Then for reward, we split $\delta/2$ evenly among all (s,a); for transition, we split $\delta/2$ evenly among all (s,a,s'). From Eq.(4) we further have

$$\max_{s,a} \|\widehat{P}(s,a) - P(s,a)\|_1 \le \max_{s,a} |\mathcal{S}| \cdot \|\widehat{P}(s,a) - P(s,a)\|_\infty \le |\mathcal{S}| \cdot \sqrt{\frac{1}{2n}\ln\frac{4|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}{\delta}}. \tag{5}$$

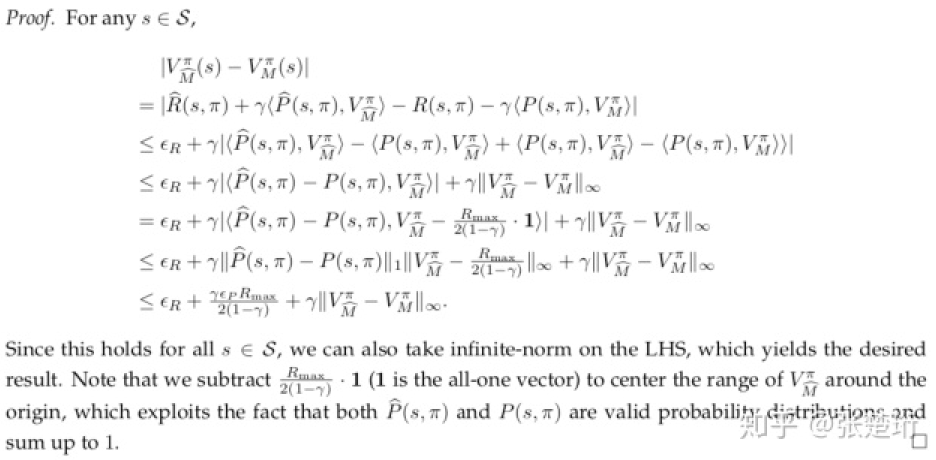
这里就利用上一讲里面红色框里面的Hoeffding不等式即可得到（虽然常数项我推导了几遍都跟这里写的对不上= =）。 $\|x\|_1 = \sum_{i=1}^{n}|x_i| \le \sum_{i=1}^{n}\max_j |x_j| = n\|x\|_\infty$ 是一个常见的bound。

步骤二：对于任意策略，估计到的MDP $\hat{M}$ 上的价值函数相比于真实 MDP M 上价值函数误差可以被模型估计误差 bound。（模型估计误差到价值函数估计误差）



Lemma 1 (Simulation Lemma). *If* $\max_{s,a} |\widehat{R}(s,a) - R(s,a)| \le \epsilon_R$ *and* $\max_{s,a} \|\widehat{P}(s,a) - P(s,a)\|_1 \le \epsilon_P$, *then for any policy* $\pi : \mathcal{S} \to \mathcal{A}$, *we have* $\forall s \in \mathcal{S}$

$$\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty \le \frac{\epsilon_R}{1-\gamma} + \frac{\gamma\,\epsilon_P\,R_{\max}}{2(1-\gamma)^2}.$$



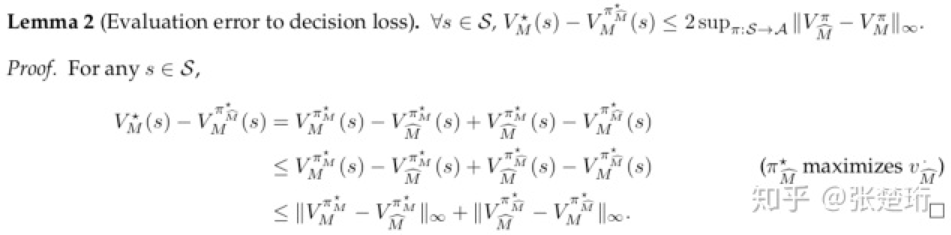
Proof. For any $s \in \mathcal{S}$,

$$\begin{aligned}
&|V^\pi_{\widehat{M}}(s) - V^\pi_M(s)| \\
&= |\widehat{R}(s,\pi) + \gamma\langle \widehat{P}(s,\pi), V^\pi_{\widehat{M}}\rangle - R(s,\pi) - \gamma\langle P(s,\pi), V^\pi_M\rangle| \\
&\le \epsilon_R + \gamma|\langle \widehat{P}(s,\pi), V^\pi_{\widehat{M}}\rangle - \langle P(s,\pi), V^\pi_{\widehat{M}}\rangle + \langle P(s,\pi), V^\pi_{\widehat{M}}\rangle - \langle P(s,\pi), V^\pi_M\rangle\rangle| \\
&\le \epsilon_R + \gamma|\langle \widehat{P}(s,\pi) - P(s,\pi), V^\pi_{\widehat{M}}\rangle| + \gamma\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty \\
&= \epsilon_R + \gamma|\langle \widehat{P}(s,\pi) - P(s,\pi), V^\pi_{\widehat{M}} - \tfrac{R_{\max}}{2(1-\gamma)} \cdot \mathbf{1}\rangle| + \gamma\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty \\
&\le \epsilon_R + \gamma\|\widehat{P}(s,\pi) - P(s,\pi)\|_1 \|V^\pi_{\widehat{M}} - \tfrac{R_{\max}}{2(1-\gamma)}\|_\infty + \gamma\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty \\
&\le \epsilon_R + \tfrac{\gamma\epsilon_P R_{\max}}{2(1-\gamma)} + \gamma\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty.
\end{aligned}$$

Since this holds for all $s \in \mathcal{S}$, we can also take infinite-norm on the LHS, which yields the desired result. Note that we subtract $\frac{R_{\max}}{2(1-\gamma)} \cdot \mathbf{1}$ ($\mathbf{1}$ is the all-one vector) to center the range of $V^\pi_{\widehat{M}}$ around the origin, which exploits the fact that both $\widehat{P}(s,\pi)$ and $P(s,\pi)$ are valid probability distributions and sum up to 1. □

这里面比较关键的是如下观察 1） $P(s,a)$ 向量乘以常数向量恒等于 1，因此第五行中减去一个常向量不影响结果；2） $x^Ty \le ||x||_1||y||_\infty$ 。

步骤三：对于 $\hat{M}$ 上的最优策略，其性能相比于真实 MDP M 上最优策略的性能差距可以被价值函数估计误差 bound。（价值函数估计误差到最优策略损失）



Lemma 2 (Evaluation error to decision loss). $\forall s \in \mathcal{S}$, $V^\star_M(s) - V^{\pi^\star_{\widehat{M}}}_M(s) \le 2\sup_{\pi:\mathcal{S}\to\mathcal{A}} \|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty$.

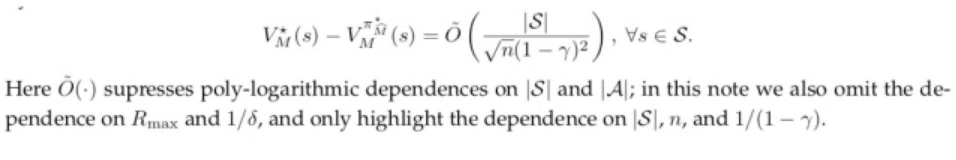
Proof. For any $s \in \mathcal{S}$,

$$\begin{aligned}
V^\star_M(s) - V^{\pi^\star_{\widehat{M}}}_M(s) &= V^{\pi^\star_M}_M(s) - V^{\pi^\star_M}_{\widehat{M}}(s) + V^{\pi^\star_M}_{\widehat{M}}(s) - V^{\pi^\star_{\widehat{M}}}_M(s) \\
&\le V^{\pi^\star_M}_M(s) - V^{\pi^\star_M}_{\widehat{M}}(s) + V^{\pi^\star_{\widehat{M}}}_{\widehat{M}}(s) - V^{\pi^\star_{\widehat{M}}}_M(s) && (\pi^\star_{\widehat{M}} \text{ maximizes } v_{\widehat{M}}) \\
&\le \|V^{\pi^\star_M}_M - V^{\pi^\star_M}_{\widehat{M}}\|_\infty + \|V^{\pi^\star_{\widehat{M}}}_{\widehat{M}} - V^{\pi^\star_{\widehat{M}}}_M\|_\infty.
\end{aligned}$$

□

拼起来：模型估计误差 -> 价值函数估计误差 -> 最优策略损失

把上面三个步骤得到的结论拼起来就可以得到最后的结论



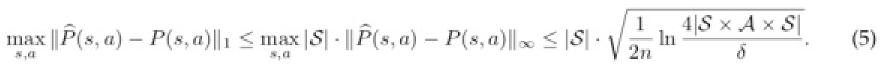
$$V^\star_M(s) - V^{\pi^\star_{\widehat{M}}}_M(s) = \tilde{O}\left(\frac{|\mathcal{S}|}{\sqrt{n}(1-\gamma)^2}\right), \ \forall s \in \mathcal{S}.$$

Here $\tilde{O}(\cdot)$ supresses poly-logarithmic dependences on $|\mathcal{S}|$ and $|\mathcal{A}|$; in this note we also omit the dependence on $R_{\max}$ and $1/\delta$, and only highlight the dependence on $|\mathcal{S}|$, n, and $1/(1-\gamma)$.

但是需要注意到的是，这个误差里面主导的部分是来自于对于 dynamics P 的估计所产生的误差，即 ϵ_P 。下面我们将得到更好的关于 ϵ_P 的 bound，从而改善这个最后的 bound。

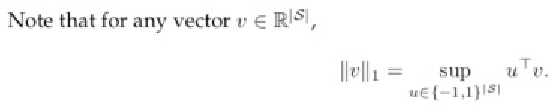
2.2. 结论二

注意到前面推导中的下面这一步的 bound 是很松的，原因是这里只认为 $\hat{P}(s,a)-P(s,a)$ 是一个普通的向量，但其实这个向量有更为特殊的性质，它的每一项之间还有约束关系即 $P(s,a)$ 或者 $\hat{P}(s,a)$ 的每一项都非负并且加起来为 1。我们下面就利用这个性质来推导更紧的 bound。



$$\max_{s,a}\|\widehat{P}(s,a)-P(s,a)\|_1 \le \max_{s,a}|\mathcal{S}|\cdot\|\widehat{P}(s,a)-P(s,a)\|_\infty \le |\mathcal{S}|\cdot\sqrt{\frac{1}{2n}\ln\frac{4|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}{\delta}}. \qquad (5)$$

注意到如下性质



Note that for any vector $v\in\mathbb{R}^{|\mathcal{S}|}$,

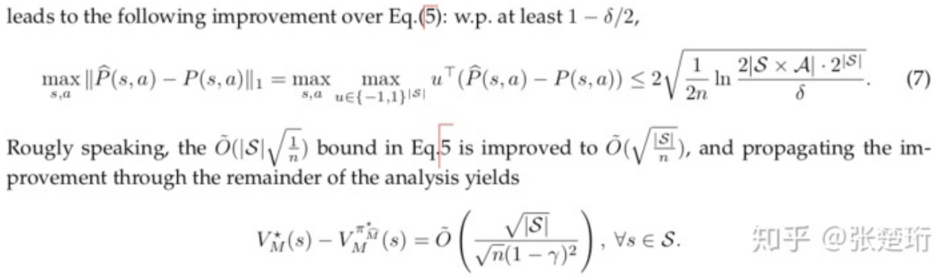
$$\|v\|_1=\sup_{u\in\{-1,1\}^{|\mathcal{S}|}} u^\top v.$$

多解释一下，即， $||v||_1=\sum_{i=1}^{n}|v_i|\le\sup_{u\in\{-1,1\}^{|S|}}\sum_{i=1}^{n}v_iu_i=\sup_{u\in\{-1,1\}^{|S|}}u^Tv$ 。

直接利用 union bound + Hoeffding 不等式。

$$\begin{aligned}
&Pr[\max_{s,a}||\hat{P}(s,a)-P(s,a)||_1\ge t]\\
=&Pr[\max_{s,a}\sup_{u}u^T(\hat{P}(s,a)-P(s,a))\ge t]\\
\le&|\mathcal{S}\times\mathcal{A}|2^{|\mathcal{S}|}Pr[u^T(\hat{P}(s,a)-P(s,a))\ge t]\quad\text{(union bound)}\\
\le&|\mathcal{S}\times\mathcal{A}|2^{|\mathcal{S}|}2e^{-2nt^2}\quad\text{(Hoeffding)}
\end{aligned}$$

即可得到最后的结论（仍然差一些常数项 = =）



leads to the following improvement over Eq.(5): w.p. at least $1-\delta/2$,

$$\max_{s,a}\|\widehat{P}(s,a)-P(s,a)\|_1=\max_{s,a}\max_{u\in\{-1,1\}^{|\mathcal{S}|}}u^\top(\widehat{P}(s,a)-P(s,a))\le 2\sqrt{\frac{1}{2n}\ln\frac{2|\mathcal{S}\times\mathcal{A}|\cdot 2^{|\mathcal{S}|}}{\delta}}. \qquad (7)$$

Rougly speaking, the $\tilde{O}(|\mathcal{S}|\sqrt{\frac{1}{n}})$ bound in Eq.5 is improved to $\tilde{O}(\sqrt{\frac{|\mathcal{S}|}{n}})$, and propagating the improvement through the remainder of the analysis yields

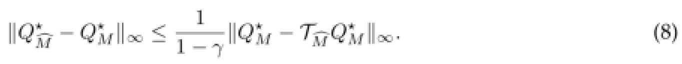
$$V_M^\star(s)-V_M^{\pi_{\widehat{M}}^\star}(s)=\tilde{O}\left(\frac{\sqrt{|\mathcal{S}|}}{\sqrt{n}(1-\gamma)^2}\right),\ \forall s\in\mathcal{S}.$$

2.3. 结论三

注意到前面的结论，最后的误差来自于两个部分 ϵ_R 和 ϵ_P ，但是最后主导的项来自于后者。现在我们考虑不要把它们拆开分别计算误差界，而是直接对它们整体 bound，这样能够得到更好的上界。

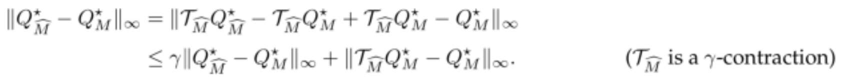
主要证明遵循如下的过程： MDP 估计误差 -(1)-> $\|Q_M^*-\mathcal{T}_{\hat{M}}Q_M^*\|_\infty$ -(2)->价值函数估计误差 -(3)-> 策略性能损失。

步骤一：上述(2)



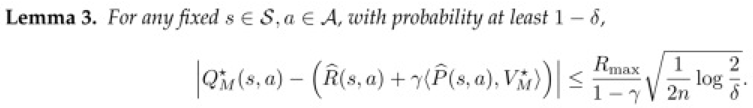
$$\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty \le \frac{1}{1-\gamma}\|Q^\star_M - \mathcal{T}_{\widehat{M}} Q^\star_M\|_\infty. \tag{8}$$

推导利用 Bellman operator 的 contraction。



$$\begin{aligned}\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty &= \|\mathcal{T}_{\widehat{M}} Q^\star_{\widehat{M}} - \mathcal{T}_{\widehat{M}} Q^\star_M + \mathcal{T}_{\widehat{M}} Q^\star_M - Q^\star_M\|_\infty \\ &\le \gamma\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty + \|\mathcal{T}_{\widehat{M}} Q^\star_M - Q^\star_M\|_\infty. \qquad (\mathcal{T}_{\widehat{M}} \text{ is a } \gamma\text{-contraction})\end{aligned}$$

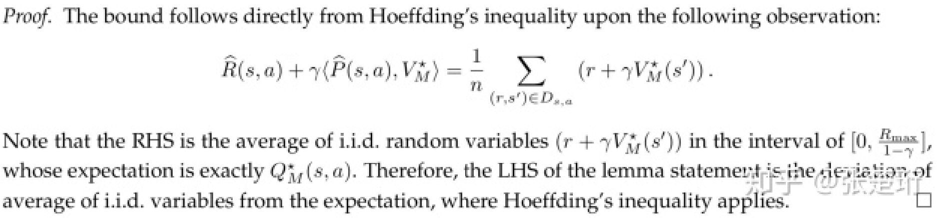
步骤二：上述(1)



Lemma 3. *For any fixed $s \in \mathcal{S}, a \in \mathcal{A}$, with probability at least $1-\delta$,*

$$\left|Q^\star_M(s,a) - \left(\widehat{R}(s,a) + \gamma\langle \widehat{P}(s,a), V^\star_M\rangle\right)\right| \le \frac{R_{\max}}{1-\gamma}\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}.$$

证明过程即直接对于一个隐含 R 和 P 的随机变量直接用 Hoeffding 不等式。



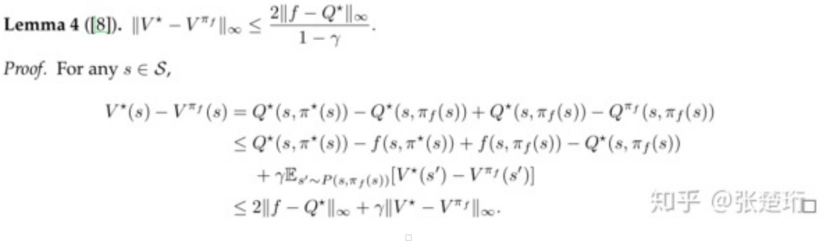
Proof. The bound follows directly from Hoeffding's inequality upon the following observation:

$$\widehat{R}(s,a) + \gamma\langle \widehat{P}(s,a), V^\star_M\rangle = \frac{1}{n}\sum_{(r,s')\in D_{s,a}} (r + \gamma V^\star_M(s')).$$

Note that the RHS is the average of i.i.d. random variables $(r + \gamma V^\star_M(s'))$ in the interval of $[0, \frac{R_{\max}}{1-\gamma}]$, whose expectation is exactly $Q^\star_M(s,a)$. Therefore, the LHS of the lemma statement is the deviation of average of i.i.d. variables from the expectation, where Hoeffding's inequality applies. □

步骤三：上述(3)

前面证明的如下 lemma。



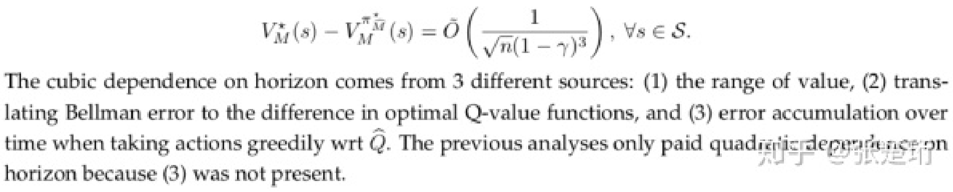
Lemma 4 ([8]). $\|V^\star - V^{\pi_f}\|_\infty \le \frac{2\|f - Q^\star\|_\infty}{1-\gamma}$.

Proof. For any $s \in \mathcal{S}$,

$$\begin{aligned}V^\star(s) - V^{\pi_f}(s) &= Q^\star(s,\pi^\star(s)) - Q^\star(s,\pi_f(s)) + Q^\star(s,\pi_f(s)) - Q^{\pi_f}(s,\pi_f(s)) \\ &\le Q^\star(s,\pi^\star(s)) - f(s,\pi^\star(s)) + f(s,\pi_f(s)) - Q^\star(s,\pi_f(s)) \\ &\quad + \gamma\mathbb{E}_{s'\sim P(s,\pi_f(s))}[V^\star(s') - V^{\pi_f}(s')] \\ &\le 2\|f - Q^\star\|_\infty + \gamma\|V^\star - V^{\pi_f}\|_\infty.\end{aligned}$$

□

串起来即得到最后的结论

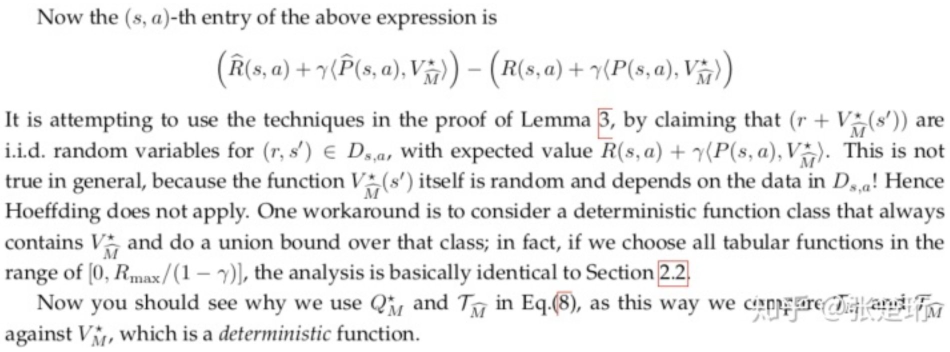


$$V^\star_M(s) - V^{\pi^\star_{\widehat{M}}}_M(s) = \tilde{O}\left(\frac{1}{\sqrt{n}(1-\gamma)^3}\right), \ \forall s \in \mathcal{S}.$$

The cubic dependence on horizon comes from 3 different sources: (1) the range of value, (2) translating Bellman error to the difference in optimal Q-value functions, and (3) error accumulation over time when taking actions greedily wrt $\widehat{Q}$. The previous analyses only paid quadratic dependence on horizon because (3) was not present.

少了一个 $|\mathcal{S}|$ 项，代价是多了一个 $\frac{1}{1-\gamma}$ 。上面提到这三个 $\frac{1}{1-\gamma}$ 的不同来源。

同时文章还提到另外一种可能的做法是不正确的，其主要原因是 Hoeffding 使用中只能是 x_i 有随机性，其他变量应该是确定的。



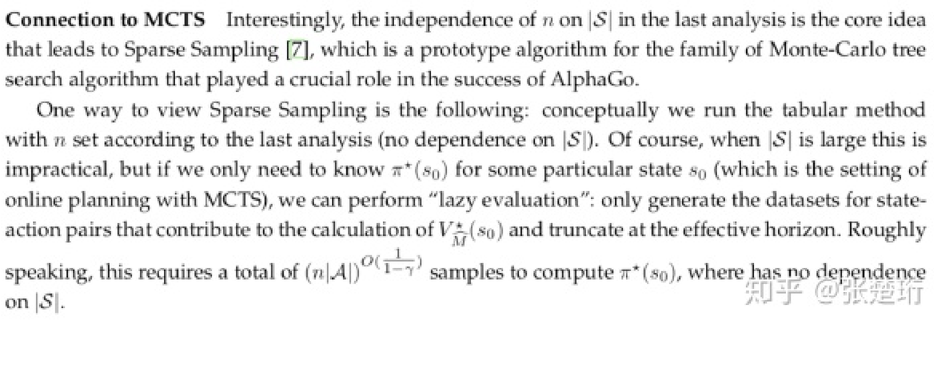
Now the (s,a)-th entry of the above expression is

$$\left(\widehat{R}(s,a)+\gamma\langle\widehat{P}(s,a),V^\star_{\widehat{M}}\rangle\right)-\left(R(s,a)+\gamma\langle P(s,a),V^\star_{\widehat{M}}\rangle\right)$$

It is attempting to use the techniques in the proof of Lemma 3, by claiming that $(r+V^\star_{\widehat{M}}(s'))$ are i.i.d. random variables for $(r,s')\in D_{s,a}$, with expected value $R(s,a)+\gamma\langle P(s,a),V^\star_{\widehat{M}}\rangle$. This is not true in general, because the function $V^\star_{\widehat{M}}(s')$ itself is random and depends on the data in $D_{s,a}$! Hence Hoeffding does not apply. One workaround is to consider a deterministic function class that always contains $V^\star_{\widehat{M}}$ and do a union bound over that class; in fact, if we choose all tabular functions in the range of $[0,R_{\max}/(1-\gamma)]$, the analysis is basically identical to Section 2.2.

Now you should see why we use $Q^\star_M$ and $\mathcal{T}_{\widehat{M}}$ in Eq.(8), as this way we compare $\mathcal{T}_M$ and $\mathcal{T}_{\widehat{M}}$ against $V^\star_M$, which is a *deterministic* function.

以上的结论告诉我们每个 (s,a) 对上所需要的采样数 n 与总的状态数无关，但是这也产生至少 $n|\mathcal{S}\times\mathcal{A}|$ 的总采样数目。但其实我们可以只对于我们关心的初始状态出发的，有限 horizon 内能达到的那一部分 MDP 。因此实际需要的采样数其实可以更少，即 $(n|\mathcal{A}|)^{O(1/1-\gamma)}$ 。



Connection to MCTS Interestingly, the independence of n on $|\mathcal{S}|$ in the last analysis is the core idea that leads to Sparse Sampling [7], which is a prototype algorithm for the family of Monte-Carlo tree search algorithm that played a crucial role in the success of AlphaGo.

One way to view Sparse Sampling is the following: conceptually we run the tabular method with n set according to the last analysis (no dependence on $|\mathcal{S}|$). Of course, when $|\mathcal{S}|$ is large this is impractical, but if we only need to know $\pi^\star(s_0)$ for some particular state s_0 (which is the setting of online planning with MCTS), we can perform "lazy evaluation": only generate the datasets for state-action pairs that contribute to the calculation of $V^\star_{\widehat{M}}(s_0)$ and truncate at the effective horizon. Roughly speaking, this requires a total of $(n|\mathcal{A}|)^{O(\frac{1}{1-\gamma})}$ samples to compute $\pi^\star(s_0)$, where has no dependence on $|\mathcal{S}|$.

发布于 2019-05-15

强化学习 (Reinforcement Learning)

▲ 赞同 3 ▼ 添加评论 分享 喜欢 收藏 ···

文章被以下专栏收录

强化学习前沿

读呀读paper

进入专栏