Notes on State Abstractions

Nan Jiang

September 28, 2018

【强化学习理论 63】StatisticalRL 7

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）课程的第四讲的第三部分，第四讲主要内容是state abstraction，这一部分主要是对 state abstraction 做 finite sample analysis，即要学习到一个 abstracted MDP 需要多少样本。

原文传送门

CS598 Note4
nanjiang.cs.illinois.edu

回顾之前讲到的 certainty-equivalence，它从每个 state-action pair 出发采样 n 个样本，然后利用估计到的 reward 和 dynamics 来找出估计到的 MDP 模型下的最优策略。理论分析表明，certainty-equivalence算法中对于每个 state-action pair 只需要【与状态空间大小无关的数目 $\times(\frac{1}{1-\gamma})^{3/2}$ / 状态空间大小的对数数目 $\times(\frac{1}{1-\gamma})$ 】个样本，即可保证所找出最优策略的性能距离真实最优策略性能差距不会太大。当然，由于每个 state-action pair 都需要这么多样本，因此最后样本的数目还是和 state-action pair 的数目成正比。

当我们使用 state abstraction 的时候，state 的数目减小了，因此所需要的样本就会相应的减少，下面就来用数学说明这件事情。

考虑每个 abstract state-action pair 都最少有若干个样本

$$n_\phi(D) := \min_{x\in\phi(\mathcal{S}),a\in\mathcal{A}} |D_{x,a}|, \quad \text{where} \quad D_{x,a} := \sum_{s\in\phi^{-1}(x)} |D_{s,a}|.$$

考虑一个对于 abstract model 的估计

Before that, we need a few more notations: Let $\widehat{M}_\phi = (\phi(\mathcal{S}), \mathcal{A}, \widehat{P}_\phi, \widehat{R}_\phi, \gamma)$ be the estimated model using the abstract representation. Let $M_\phi = (\phi(\mathcal{S}), \mathcal{A}, P_\phi, R_\phi, \gamma)$ be the following MDP:

$$R_\phi(x,a) = \frac{\sum_{\tilde{s}\in\phi^{-1}(x)} |D_{\tilde{s},a}| R(s,a)}{|D_{\phi(s),a}|}, \quad P_\phi(x'|x,a) = \frac{\sum_{\tilde{s}\in\phi^{-1}(x)} |D_{\tilde{s},a}| P(x'|s,a)}{|D_{\phi(s),a}|}.$$

考虑真实模型的最优 Q 函数相比于估计模型最优 Q 函数的差

Bounding $\left\|Q^\star_M - [Q^\star_{\widehat{M}_\phi}]_M\right\|_\infty$: We bound it by introducing an intermediate term:

$$\left\|Q^\star_M - [Q^\star_{\widehat{M}_\phi}]_M\right\|_\infty \le \left\|Q^\star_M - [Q^\star_{M_\phi}]_M\right\|_\infty + \left\|[Q^\star_{M_\phi}]_M - [Q^\star_{\widehat{M}_\phi}]_M\right\|_\infty.$$

其中第一项表示真实模型最优 Q 函数相比于抽象模型最优 Q 函数的差，这一项我们叫做 approximation error，它与 ϕ 的选取有关，这一项我们在前一讲已经详细分析过了；第二项表述抽象模型最优 Q 函数相比于估计的抽象模型最优 Q 函数的差，这一项我们叫做 estimation error，当样本数目足够多的时候，这一项变为零，它与 ϕ 的选取无关。下面我们就要来 bound 这一项。

一个简单的想法是就直接利用前面对于 certainty-equivalence 的分析，把原来的状态空间 $\mathcal{S}$ 直接替换为抽象的状态空间 $\phi(\mathcal{S})$ 。但是这样直接替换会面临一个问题。之前的分析中，我们直接认为 $D_{s,a}$ 中的数据是从独立同分布 $P(s,a)$ 中采用得到的，但是我们不能说现在 $D_{s,a}$ 中的数据是从独立同分布 $P_\phi(x,a)$ 里面得到的，因为对于不同的 $s\in\phi^{-1}(x)$ ，其采样样本的分布是独立但是不一定是同分布的。

考虑可能的这个问题，我们需要对于不同的 $s\in\phi^{-1}(x)$ 分别展开，并且利用之前的concentration inequality。

$$\left\|[Q^\star_{M_\phi}]_M - [Q^\star_{\widehat{M}_\phi}]_M\right\|_\infty = \left\|Q^\star_{M_\phi} - Q^\star_{\widehat{M}_\phi}\right\|_\infty$$
$$\le \frac{1}{1-\gamma}\left\|Q^\star_{M_\phi} - \mathcal{T}_{\widehat{M}_\phi} Q^\star_{M_\phi}\right\|_\infty = \frac{1}{1-\gamma}\left\|\mathcal{T}_{\widehat{M}_\phi} Q^\star_{M_\phi} - \mathcal{T}_{M_\phi} Q^\star_{M_\phi}\right\|_\infty.$$

For each $(x,a)\in\phi(S)\times\mathcal{A}$,

$$\begin{aligned}
&|(\mathcal{T}_{\widehat{M}_\phi} Q^\star_{M_\phi})(x,a) - (\mathcal{T}_{M_\phi} Q^\star_{M_\phi})(x,a)| \\
&= |\widehat{R}_\phi(x,a) + \gamma\langle \widehat{P}_\phi(x,a), V^\star_{M_\phi}\rangle - R_\phi(x,a) - \gamma\langle P_\phi(x,a), V^\star_{M_\phi}\rangle| \\
&= \left|\frac{1}{|D_{x,a}|}\sum_{s\in\phi^{-1}(x)}\sum_{(r,s')\in D_{s,a}} \left(r + \gamma V^\star_{M_\phi}(\phi(s')) - R(s,a) - \gamma\langle P(s,a), [V^\star_{M_\phi}]_M\rangle\right)\right|.
\end{aligned}$$

If we view the nested sum as a flat sum, the expression is the sum of the differences between random variables $r+\gamma V^\star_{M_\phi}(s')$ and their expectation w.r.t. the randomness of (r,s'), so Hoeffding's inequality applies (although for different $s\in\phi^{-1}(x)$ the random variables have non-identical distributions): with probability at least $1-\delta$,

$$\left\|\mathcal{T}_{\widehat{M}_\phi} Q^\star_{M_\phi} - \mathcal{T}_{M_\phi} Q^\star_{M_\phi}\right\|_\infty \le \frac{R_{\max}}{1-\gamma}\sqrt{\frac{1}{2n_\phi(D)}\ln\frac{2|\phi(\mathcal{S})\times\mathcal{A}|}{\delta}}.$$

This completes the analysis.

结论还是类似，每个 abstract state-action pair 需要的样本数为【状态空间大小的对数数目 $\times(\frac{1}{1-\gamma})$ 】个样本。

发布于 2019-05-26

强化学习 (Reinforcement Learning)

文章被以下专栏收录

强化学习前沿
读呀读paper

进入专栏