

Notes on State Abstractions

Nan Jiang

September 28, 2018

【强化学习理论 62】Statistical RL 6



张楚珩

清华大学 交叉信息院博士在读

7 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）课程的第四讲的第二部分，主要讲的内容是state abstraction。

原文传送门

CS598 Note4

nanjiang.cs.illinois.edu



回顾

对于如下的定义

Definition 3 (Approximate abstractions). Given MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and state abstraction ϕ that operates on \mathcal{S} , define the following types of abstractions:

1. ϕ is an ϵ_{π^*} -approximate π^* -irrelevant abstraction, if there exists an abstract policy $\pi : \phi(\mathcal{S}) \rightarrow \mathcal{A}$, such that $\|V_M^* - V_M^{[\pi]} \|_\infty \leq \epsilon_{\pi^*}$.
2. ϕ is an ϵ_{Q^*} -approximate Q^* -irrelevant abstraction if there exists an abstract Q -value function $f : \phi(\mathcal{S}) \times \mathcal{A} \rightarrow \mathbb{R}$, such that $\|[f]_M - Q_M^* \|_\infty \leq \epsilon_{Q^*}$.
3. ϕ is an (ϵ_R, ϵ_P) -approximate model-irrelevant abstraction if for any $s^{(1)}$ and $s^{(2)}$ where $\phi(s^{(1)}) = \phi(s^{(2)})$, $\forall a \in \mathcal{A}$,

$$|R(s^{(1)}, a) - R(s^{(2)}, a)| \leq \epsilon_R, \quad \left\| \Phi P(s^{(1)}, a) - \Phi P(s^{(2)}, a) \right\|_1 \leq \epsilon_P. \quad (3)$$

Note that Definition 1 is recovered when all approximation errors are set to 0.

知乎 @张楚珩

Definition 2 (*lifting*). For any function f that operates on $\phi(\mathcal{S})$, let $[f]_M$ denote its *lifted* version, which is a function over \mathcal{S} , defined as $[f]_M(s) := f(\phi(s))$. Similarly we can also lift a state-action value function. Lifting a real-valued function f over states can also be expressed in vector form: $[f]_M = \Phi^\top f$.

有以下性质

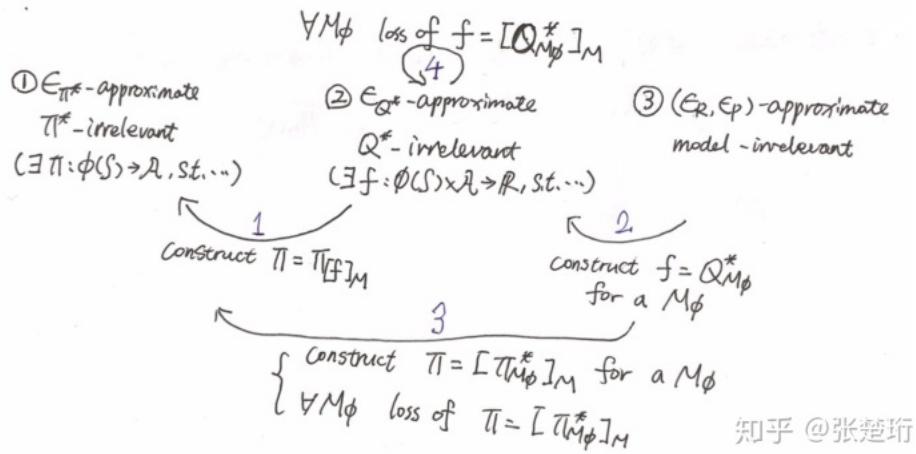
Theorem 2. (1) If ϕ is an (ϵ_R, ϵ_P) -approximate model-irrelevant abstraction, then ϕ is also an approximate Q^* -irrelevant abstraction with approximation error $\epsilon_{Q^*} = \frac{\epsilon_R}{1-\gamma} + \frac{\gamma\epsilon_P R_{\max}}{2(1-\gamma)^2}$.
(2) If ϕ is an ϵ_{Q^*} -approximate Q^* -irrelevant abstraction, then ϕ is also an approximate π^* -irrelevant abstraction with approximation error $\epsilon_{\pi^*} = 2\epsilon_{Q^*}/(1-\gamma)$.

Theorem 4. Let ϕ be an (ϵ_R, ϵ_P) -approximate model-irrelevant abstraction of M , and M_ϕ be an abstract model defined as in Lemma 3 with arbitrary distributions $\{p_x\}$, then

$$\left\| V_M^* - V_M^{[\pi_{M_\phi}^*]_M} \right\|_\infty \leq \frac{2\epsilon_R}{1-\gamma} + \frac{\gamma\epsilon_P R_{\max}}{(1-\gamma)^2}.$$

Theorem 5. Let ϕ be an ϵ_{Q^*} -approximate Q^* -irrelevant abstraction for M . Then, for M_ϕ constructed as in Lemma 3 with arbitrary distributions $\{p_x\}$, we have $\|[Q_{M_\phi}^*]_M - Q_M^*\|_\infty \leq 2\epsilon_{Q^*}/(1-\gamma)$.

这几个定理的关系如下



1. 要用第二种近似bound第一种近似，即找到一个 π 满足第一种近似中的条件即可；
2. 要用第三种近似bound第二种近似，即找到一个 f 满足第二种近似中的条件即可；为了找到这个 f ，我们将新定义一个与 ϕ 有关的MDP M_ϕ ，并说明这个新MDP下的最优价值函数满足第二种近似中的条件；
3. 可以连用1和2来用第三种近似bound第一种近似；为了得到更紧的bound，可以直接找到一个 π 满足第一种近似中的条件，同样，我们找到的这个 π 就是 M_ϕ ；这个证明又告诉了我们另外一件事情：第三种近似下，直接在 M_ϕ 上找到的最优策略的损失有多大，其中我们使用 M_ϕ 上最优策略和真实最优策略之间价值函数的差来衡量“损失”。
4. 这里证明的是，在第二种近似下，直接在 M_ϕ 上找到的最优策略的损失有多大。

在证明这四件事情之前，我们先看看如何定义一个与 ϕ 有关的MDP M_ϕ ，并且如果 ϕ 满足第三种近似，那么这个MDP M_ϕ 有什么样的性质。

Construct a MDP with respect to the approximate model-irrelevant

Lemma 3. Let ϕ be an (ϵ_R, ϵ_P) -approximate model-irrelevant abstraction of M . Given any distributions $\{p_x : x \in \phi(S)\}$ where each p_x is supported on $\phi^{-1}(x)$, define $M_\phi = (\phi(S), \mathcal{A}, P_\phi, R_\phi, \gamma)$, where $R_\phi(x, a) = \mathbb{E}_{s \sim p_x}[R(s, a)]$, and $P_\phi(x'|x, a) = \mathbb{E}_{s \sim p_x}[P(x'|s, a)]$. Then for any $s \in S, a \in \mathcal{A}$,

$$|R_\phi(\phi(s), a) - R(s, a)| \leq \epsilon_R, \quad \|P_\phi(x, a) - \Phi P(s, a)\|_1 \leq \epsilon_P.$$

即对于满足第三种近似的 ϕ 来说，这个关于它的新MDP的reward和dynamics都距离真实MDP相差不远。这样构造的原因是当原MDP中很多状态都抽象到 M_ϕ 中的一个状态之后，原MDP中的状态就无法区分了，我们需要对于每一个抽象出来的状态，都规定一个条件概率分布，即给定 x 它对应 s 的概率是多少。这样才能把原MDP中的reward/dynamics联系起来并且得到一个完全定义在 x -space 上的reward/dynamics。

Proof. We only prove for the transition part; the reward part follows from a similar (and easier) argument. Consider any fixed x and a . Let $q_s := \Phi P(s, a)$. By the definition of approximate bisimulation we have $\|q_{s^{(1)}} - q_{s^{(2)}}\|_1 \leq \epsilon_P$ for any $\phi(s^{(1)}) = \phi(s^{(2)})$. The LHS of the claim on transition function is (let $x := \phi(s)$)

$$\begin{aligned} & \|P_\phi(x, a) - \Phi P(s, a)\|_1 \\ &= \left\| \sum_{\tilde{s} \in \phi^{-1}(x)} p_x(\tilde{s}) q_{\tilde{s}} - q_s \right\|_1 = \left\| \sum_{\tilde{s} \in \phi^{-1}(x)} p_x(\tilde{s}) (q_{\tilde{s}} - q_s) \right\|_1 \\ &\leq \sum_{\tilde{s} \in \phi^{-1}(x)} \|p_x(\tilde{s}) (q_{\tilde{s}} - q_s)\|_1 \leq \sum_{\tilde{s} \in \phi^{-1}(x)} p_x(\tilde{s}) \epsilon_P = \epsilon_P. \end{aligned} \quad \text{知乎 @张楚珩}$$

证明很容易，大体上就是平均值小于等于最大值。

下面证明第一件事情：

Approximate Q-star-irrelevant abstraction bounds approximate pi-star-irrelevant abstraction

第二种近似里面提到存在一个价值函数 f 满足某个性质，第一种近似要求存在一个策略 π 满足某个性质。这样我们就规定策略 π 是相对于价值函数 f 的最优策略，并且证明它满足第一种近似的要求。

这件事情在第一讲中已经证明了，可以拿来直接使用。

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

下面证明第二件事情：

Approximate model-irrelevant abstraction bounds approximate Q-star-irrelevant abstraction

Define M_ϕ to be an abstract model as in Lemma 3 w.r.t. arbitrary distributions $\{p_x\}$. We will use $Q_{M_\phi}^*$ as the f function in the definition of approximate Q^* -irrelevance, and upper bound $\|Q_{M_\phi}^* - Q^*\|_\infty$ as:

$$\|Q_{M_\phi}^* - Q^*\|_\infty \leq \frac{1}{1 - \gamma} \|Q_{M_\phi}^* - \mathcal{T}[Q_{M_\phi}^*]\|_\infty = \frac{1}{1 - \gamma} \|(\mathcal{T}_{M_\phi} Q_{M_\phi}^*) - \mathcal{T}[Q_{M_\phi}^*]\|_\infty.$$

For any (s, a) ,

$$\begin{aligned} & |(\mathcal{T}_{M_\phi} Q_{M_\phi}^*)(s, a) - (\mathcal{T}[Q_{M_\phi}^*])(s, a)| \\ &= |(\mathcal{T}_{M_\phi} Q_{M_\phi}^*)(\phi(s), a) - (\mathcal{T}[Q_{M_\phi}^*])(s, a)| \\ &= |R_\phi(\phi(s), a) + \gamma \langle P_\phi(\phi(s), a), V_{M_\phi}^* \rangle - R(s, a) - \gamma \langle P(s, a), [V_{M_\phi}^*]_M \rangle| \\ &\leq \epsilon_R + \gamma \left| \langle P_\phi(\phi(s), a), V_{M_\phi}^* \rangle - \langle P(s, a), \Phi^\top V_{M_\phi}^* \rangle \right| \\ &= \epsilon_R + \gamma \left| \langle P_\phi(\phi(s), a), V_{M_\phi}^* \rangle - \langle \Phi P(s, a), V_{M_\phi}^* \rangle \right| \quad (*) \\ &\leq \epsilon_R + \gamma \epsilon_P \|V_{M_\phi}^* - \frac{R_{\max}}{2(1-\gamma)} \mathbf{1}\|_\infty \\ &\leq \epsilon_R + \gamma \epsilon_P R_{\max} / (2(1 - \gamma)). \end{aligned}$$

In step (*), we notice that $[V_{M_\phi}^*]_M$ is piece-wise constant, so when we take its dot-product with $P(s, a)$, we essentially first collapse $P(s, a)$ onto $\phi(S)$ (which is done by the Φ operator) and then take its dot-product with $V_{M_\phi}^*$. The rest of the proof is similar to that of the simulation lemma. \square

其中，第一个式子中的第一个不等式的推导和下面的推导方法类似，第二个等号是由于Bellman算子的不动点性质。

$$\|Q_{\widehat{M}}^* - Q_M^*\|_{\infty} \leq \frac{1}{1-\gamma} \|Q_M^* - \mathcal{T}_{\widehat{M}} Q_M^*\|_{\infty}. \quad (8)$$

This is because

$$\begin{aligned} \|Q_{\widehat{M}}^* - Q_M^*\|_{\infty} &= \|\mathcal{T}_{\widehat{M}} Q_{\widehat{M}}^* - \mathcal{T}_{\widehat{M}} Q_M^* + \mathcal{T}_{\widehat{M}} Q_M^* - Q_M^*\|_{\infty} \\ &\leq \gamma \|Q_{\widehat{M}}^* - Q_M^*\|_{\infty} + \|\mathcal{T}_{\widehat{M}} Q_M^* - Q_M^*\|_{\infty}. \end{aligned} \quad (\mathcal{T}_{\widehat{M}} \text{ is a } \gamma\text{-contraction})$$

第二串式子的推导难点在于分清楚每个符号表示的含义；倒数第二个不等号要考虑到P不是普通的向量，它的每一项都为正数并且加起来为1，因此它与任何单位向量的内积都为1，因此在倒数第二行里面减去一个常向量，这样能够让bound缩小一半；最后一个不等式用到 $\|u^*v\|_1 \leq \|u\|_1 \|v\|_{\infty}$ 。

下面证明第三件事情：

Approximate model-irrelevant abstraction bounds approximate pi-star-irrelevant abstraction (How lossy is the optimal policy for \mathcal{M}_{ϕ})

为了证明 \mathcal{M}_{ϕ} 下的最优策略的损失，先证明：对于任意 \mathcal{M}_{ϕ} 下的策略 π ，策略 π 在 \mathcal{M}_{ϕ} 下的性能和策略 $\pi|_{\mathcal{M}}$ 在 \mathcal{M} 下的性能的差距上界。注意到这个上界刚好是我们要证明损失的一半。

Proof. We first prove that for any abstract policy $\pi : \phi(S) \rightarrow \mathcal{A}$,

$$\left\| [V_{M_\phi}^\pi]_M - V_M^{[\pi]_M} \right\|_\infty \leq \frac{\epsilon_R}{1-\gamma} + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)^2}. \quad (4)$$

To prove this, first recall the contraction property of policy-specific Bellman update operator for state-value functions, which implies that

$$\left\| [V_{M_\phi}^\pi]_M - V_M^{[\pi]_M} \right\|_\infty \leq \frac{1}{1-\gamma} \left\| [V_{M_\phi}^\pi]_M - \mathcal{T}^{[\pi]_M} [V_{M_\phi}^\pi]_M \right\|_\infty = \frac{1}{1-\gamma} \left\| [\mathcal{T}_{M_\phi}^\pi V_{M_\phi}^\pi]_M - \mathcal{T}^{[\pi]_M} [V_{M_\phi}^\pi]_M \right\|_\infty.$$

For notation simplicity let $R^{\pi'}(s) := R(s, \pi'(s))$ and $P^{\pi'}(s) := P(s, \pi'(s))$. For any $s \in S$,

$$\begin{aligned} & |[\mathcal{T}_{M_\phi}^\pi V_{M_\phi}^\pi]_M(s) - \mathcal{T}^{[\pi]_M} [V_{M_\phi}^\pi]_M(s)| \\ &= |(\mathcal{T}_{M_\phi}^\pi V_{M_\phi}^\pi)(\phi(s)) - \mathcal{T}^{[\pi]_M} [V_{M_\phi}^\pi]_M(s)| \\ &= |R_\phi^\pi(\phi(s)) + \gamma \langle P_\phi^\pi(\phi(s)), V_{M_\phi}^\pi \rangle - R^{[\pi]_M}(s) - \gamma \langle P^{[\pi]_M}(s), V_M^{[\pi]_M} \rangle| \\ &\leq \epsilon_R + \gamma |\langle P_\phi^\pi(\phi(s)), V_{M_\phi}^\pi \rangle - \langle P^{[\pi]_M}(s), [V_{M_\phi}^\pi]_M \rangle| \\ &= \epsilon_R + \gamma \left| \langle P_\phi^\pi(\phi(s)), V_{M_\phi}^\pi \rangle - \langle \Phi P^{[\pi]_M}(s), V_{M_\phi}^\pi \rangle \right| \\ &\leq \epsilon_R + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)}. \end{aligned}$$

Now that we have a uniform upper bound on evaluation error, it might be attempting to argue that we under-estimate π_M^* and over-estimate $\pi_{M_\phi}^*$ at most this much, hence the decision loss is twice the evaluation error. This argument does not apply here because π_M^* cannot be necessarily expressed as a lifted abstract policy when ϕ is not an exact bisimulation!

接下来我们要证明的损失可以拆开，并且使用前面的结论

Instead we can use the following argument: for any $s \in S$,

$$\begin{aligned} V_M^*(s) - V_M^{[\pi_{M_\phi}^*]_M}(s) &= V_M^*(s) - V_{M_\phi}^*(\phi(s)) + V_{M_\phi}^*(\phi(s)) - V_M^{[\pi_{M_\phi}^*]_M}(s) \\ &\leq \left\| Q_M^* - [Q_{M_\phi}^*]_M \right\|_\infty + \left\| [V_{M_\phi}^{\pi_{M_\phi}^*}]_M - V_M^{[\pi_{M_\phi}^*]_M} \right\|_\infty. \end{aligned}$$

Here both terms can be bounded by $\frac{\epsilon_R}{1-\gamma} + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)^2}$ but for different reasons: the bound applies to the first term due to Claim (1) of Theorem 2, and applies to the second term through Eq. (4) as $\pi_{M_\phi}^*$ is an abstract policy. \square

下面证明第四件事情

Under approximate Q-star-irrelevant, how lossy is the optimal policy for μ ,

这件事情看起来不是很自然，因为第二种近似下，有些状态可能本质上不一样，但是它们价值函数差不多，我们现在把它们当做一样的状态来看待。但是这里后面证明了在这种高度抽象的MDP下找最优策略，性能相比于最优策略相差也不多。

为了理解这件事情，我们考虑exact Q-star-irrelevant的情况。考虑我们通过Bellman operator来找最优策略，如果两个状态 $s^{(1)}$ 和 $s^{(2)}$ ，如果它们Q值相同，那么它们在Bellman operator作用下产生的效果类似。进一步来说，可以说明原本MDP中的最优价值函数，也是 M_ϕ 下Bellman operator的不动点。

Exact Q^* -irrelevance To develop intuition, let's see what happens when ϕ is an exact Q^* -irrelevant abstraction: we can prove that $[Q_{M_\phi}^*]_M = Q_M^*$, despite that the dynamics and rewards in M_ϕ "do not make sense". In particular, we know that for any $s^{(1)}$ and $s^{(2)}$ aggregated by ϕ , for any $a \in \mathcal{A}$,

$$R(s^{(1)}, a) + \gamma \langle P(s^{(1)}, a), V_M^* \rangle = Q^*(s^{(1)}, a) = Q^*(s^{(2)}, a) = R(s^{(2)}, a) + \gamma \langle P(s^{(2)}, a), V_M^* \rangle.$$

This equation tells us that, although ϕ aggregates states that can have very different rewards and dynamics, they at least share one thing: the Bellman operator updates Q_M^* in exactly the same way at $s^{(1)}$ and $s^{(2)}$ (for any action).

Let $[Q_M^*]_\phi(x, a) = Q_M^*(s, a)$ for any $s \in \phi^{-1}(x)$; note that the notation $[\cdot]_\phi$ can only be applied to functions that are piece-wise constant under ϕ . We now show that $[Q_M^*]_\phi$ is the fixed point of \mathcal{T}_{M_ϕ} , which proves the claim. This is because, for any $x \in \phi(\mathcal{S})$, $a \in \mathcal{A}$, let s be any state in $\phi^{-1}(x)$:

$$\begin{aligned} (\mathcal{T}_{M_\phi}[Q_M^*]_\phi)(x, a) &= R_\phi(x, a) + \gamma \langle P_\phi(x, a), [V_M^*]_\phi \rangle \\ &= \sum_{s \in \phi^{-1}(x)} p_x(s) (R(s, a) + \gamma \langle \Phi P(s, a), [V_M^*]_\phi \rangle) \\ &= \sum_{s \in \phi^{-1}(x)} p_x(s) (R(s, a) + \gamma \langle P(s, a), V_M^* \rangle) \\ &= \sum_{s \in \phi^{-1}(x)} p_x(s) [Q_M^*]_\phi(x, a) = [Q_M^*]_\phi(x, a). \end{aligned}$$

知乎 @张楚珩

下面的证明也类似，大致上（用语言太难说清楚了，放弃说明了。。）：

M 下最优价值函数相比于 M_ϕ 下最优价值函数

M 下最优价值函数相比于它被某个与 M_ϕ 有关Bellman operator作用过的函数

核心就是：exact情形证明容易是由于 M 下的最优价值函数可以投影到x-space上，然后在x-space上分析都很容易；approximate情形下，不可以这样做，分析都需要在s-space上进行。因此会定义另外的在s-space上但又与 ϕ 有关的一个MDP M_ϕ 来绕一下。其实想法挺直接，证明挺繁琐。

The approximate case The more general case is much trickier, as Q_M^* is not piece-wise constant when ϕ is not exactly Q^* -irrelevant, so we cannot apply T_{M_ϕ} to it.

To get around this issue, define a new MDP $M'_\phi = (\mathcal{S}, \mathcal{A}, P'_\phi, R'_\phi, \gamma)$, with

$$R'_\phi(s, a) = \mathbb{E}_{\tilde{s} \sim p_{\phi(s)}}[R(\tilde{s}, a)], \quad P'_\phi(s'|s, a) = \mathbb{E}_{\tilde{s} \sim p_{\phi(s)}}[P(s'|\tilde{s}, a)].$$

Recall that $\{p_x\}$ are a set of arbitrary distributions and we use them as weights for defining M_ϕ . The model here, $M_{\phi'}$, also combines parameters from aggregated states, but is defined over the *primitive* state space. This seemingly crazy model has two important properties: (1) Its optimal Q -value function coincides with that of M_ϕ (after lifting), and (2) It's defined over \mathcal{S} so we can apply its Bellman operator to Q_M^* .

We first prove that $[Q_{M_\phi}^*]_M = Q_{M'_\phi}^*$, by showing that $T_{M'_\phi}[Q_{M_\phi}^*]_M = [Q_{M_\phi}^*]_M$:

$$\begin{aligned} (T_{M'_\phi}[Q_{M_\phi}^*]_M)(s, a) &= R'_\phi(s, a) + \gamma \langle P'_\phi(s, a), [V_{M_\phi}^*]_M \rangle \\ &= \sum_{\tilde{s}: \phi(\tilde{s})=\phi(s)} p_x(\tilde{s}) (R(\tilde{s}, a) + \gamma \langle P(\tilde{s}, a), [V_{M_\phi}^*]_M \rangle) \\ &= \sum_{\tilde{s}: \phi(\tilde{s})=\phi(s)} p_x(\tilde{s}) R(\tilde{s}, a) + \sum_{\tilde{s}: \phi(\tilde{s})=\phi(s)} p_x(\tilde{s}) \gamma \langle P(\tilde{s}, a), V_{M_\phi}^* \rangle \\ &= R_\phi(\phi(s), a) + \gamma \langle P_\phi(\phi(s), a), V_{M_\phi}^* \rangle \\ &= Q_{M_\phi}^*(\phi(s), a) = [Q_{M_\phi}^*]_M(s, a). \end{aligned}$$

With this result, we have

$$\|[Q_{M_\phi}^*]_M - Q_M^*\|_\infty = \|Q_{M'_\phi}^* - Q_M^*\|_\infty \leq \frac{1}{1-\gamma} \|T_{M'_\phi}Q_M^* - Q_M^*\|_\infty.$$

And

$$\begin{aligned} & |(T_{M'_\phi}Q_M^*)(s, a) - Q_M^*(s, a)| \\ &= |R'_\phi(s, a) + \gamma \langle P'_\phi(s, a), V_M^* \rangle - Q_M^*(s, a)| \\ &= \left| \left(\sum_{\tilde{s}: \phi(\tilde{s})=\phi(s)} p_x(\tilde{s}) (R(\tilde{s}, a) + \gamma \langle P(\tilde{s}, a), V_M^* \rangle) \right) - Q_M^*(s, a) \right| \\ &= \left| \sum_{\tilde{s}: \phi(\tilde{s})=\phi(s)} p_x(\tilde{s}) (Q_M^*(\tilde{s}, a) - Q_M^*(s, a)) \right| \\ &\leq \left| \sum_{\tilde{s}: \phi(\tilde{s})=\phi(s)} p_x(\tilde{s}) (2\epsilon_{Q^*}) \right| = 2\epsilon_{Q^*}. \end{aligned}$$

知乎 @张楚珩

编辑于 2019-05-26

强化学习 (Reinforcement Learning)

▲ 赞同 7



💬 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