MDP Preliminaries

Nan Jiang

April 10, 2019

【强化学习理论 58】StatisticalRL 2

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）课程的第一讲（第二部分），之所以拆成两个部分是因为图片贴多了之后知乎编辑变得特别卡。我们继续。

原文传送门

CS598 Note1
nanjiang.cs.illinois.edu

二、MDP上的规划

规划（Planning）的目标是optimal control即找到最优策略或者最优价值函数，一般它隐含有已知环境的 P,R 或者对它们有建模估计的意思。

2.1. Policy Iteration

policy iteration主要分为两步，即policy evaluation和policy improvement。

The policy iteration algorithm starts from an arbitrary policy π_0, and repeat the following iterative procedure: for $k = 1, 2, \ldots$

$$\pi_k = \pi_{Q^{\pi_{k-1}}}.$$

Essentially, in each iteration we compute the Q-value function of π_{k-1} (e.g., using the analytical form given in Equation 4), and then compute the greedy policy for the next iteration. The first step is often called *policy evaluation*, and the second step is often called *policy improvement*.

这里主要讲两个结论：

- policy iteration中每次迭代更新策略，每次迭代之后该策略下的价值函数对于任意的状态都不会变得更差；如果策略不再变化，那么就得到了最优策略。
- policy iteration中策略的价值函数距离最优价值函数的距离是呈指数级减小的。

第一个结论：policy improvement

Theorem 1 (Policy improvement theorem). *In policy iteration, $V^{\pi_k}(s) \geq V^{\pi_{k-1}}(s)$ holds for all $k \geq 1$ and $s \in \mathcal{S}, a \in \mathcal{A}$, and the improvement is strictly positive in at least one state until $\pi^\star$ is found.*

Therefore, the termination criterion for policy iteration is $Q^{\pi_k} = Q^{\pi_{k-1}}$. Since we are only searching over stationary and deterministic policies, and a new policy that is different from all previous ones is found every iteration, the algorithm is guaranteed to terminate in $|\mathcal{A}|^{|\mathcal{S}|}$ iterations.

$|\mathcal{A}|^{|\mathcal{S}|}$ 这个bound是怎么来的？可以证明如果前一次迭代的策略和后一次迭代的策略一样，那么迭代收敛，最优策略被找到；每次迭代价值函数都不会变的更差，因此已经找到过的策略肯定不会再回去了；总共可能有的策略就是这么多个，因此最差情况每次就找到比上一次好那么一点点的策略，需要迭代这么多次。

证明思路如下：构造关于策略的advantage function，由于policy improvement step都是关于前一轮价值函数的贪心策略，因此单步的advantage function都是正数；把总体的策略价值函数改变写成advantage function的线性组合，即可证明每轮迭代价值函数只增不减。

To prove the policy improvement theorem, we introduce an important concept called *advantage*.

Definition 1 (Advantage). The advantage of action a at state s over policy π is defined as $A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s)$. The advantage of policy π' over policy π is defined as $A^\pi(s,\pi') := A^\pi(s,\pi'(s))$.

Since policy iteration always takes the greedy policy of the current policy's Q-value function, by definition the advantage of the new policy over the old one is non-negative. The next result shows that the value difference between two policies can be expressed using the advantage function. The policy improvement theorem immediately follows, since $V^{\pi_k}(s) - V^{\pi_{k-1}}(s)$ can be decomposed into the sum of nonnegative terms.

Proposition 2 (Advantage decomposition of policy values). *For any π, π', and any state $s \in \mathcal{S}$,*

$$V^{\pi'}(s) - V^\pi(s) = \frac{1}{1-\gamma}\mathbb{E}_{s' \sim \eta_s^{\pi'}}[A^\pi(s',\pi')].$$

where $\eta_s^{\pi'}$ is the normalized discounted occupancy induced by policy π' from starting state s.

Proof. Consider a sequence of (potentially non-stationary) policies $\{\pi_i\}_{i\geq 0}$, where $\pi_0 = \pi$, $\pi_\infty = \pi'$. For any intermediate i, π_i is the non-stationary policy that follows π' for the first i time-steps and switches to π for the remainder of the trajectory. Now we can rewrite the LHS of the statement as:

$$V^{\pi'}(s) - V^\pi(s) = \sum_{i=0}^{\infty}\left(V^{\pi_{i+1}}(s) - V^{\pi_i}(s)\right).$$

For each term on the RHS, observe that π_i and π_{i+1} share the same "roll-in" policy π' for the first i steps, which defines a roll-in distribution $\mathbb{P}[s_{i+1}|s_1 = s, \pi']$. They also share the same "roll-out" policy π starting from the $(i+2)$-th time step, so conditioned on $s_{i+1} = s, a_{i+1} = a$, the total expected reward picked up in the remainder of the trajectory is $\gamma^i Q^\pi(s,a)$ for both π_i and π_{i+1}. Putting together, we have

$$
\begin{aligned}
V^{\pi'}(s) - V^{\pi}(s) &= \sum_{i=0}^{\infty} \gamma^i \sum_{s' \in \mathcal{S}} \mathbb{P}[s_{i+1} = s' | s_1 = s, \pi'] \left(Q^{\pi}(s', \pi'(s')) - Q^{\pi}(s', \pi(s')) \right) \\
&= \sum_{i=0}^{\infty} \gamma^i \sum_{s' \in \mathcal{S}} \mathbb{P}[s_{i+1} = s' | s_1 = s, \pi'] A^{\pi}(s', \pi').
\end{aligned}
$$

The result follows by noticing that $\sum_{i=0}^{\infty} \gamma^i \mathbb{P}[s_{i+1} = s' | s_1 = s, \pi'] = \frac{1}{1-\gamma} \eta_s^{\pi'}(s')$. □

第二个结论：policy iteration enjoys exponential convergence

这个证明构造性比较强，直接看结论吧。

Theorem 3 (Policy iteration enjoys exponential convergence). $\|Q^\star - Q^{\pi_{k+1}}\|_\infty \leq \gamma \|Q^\star - Q^{\pi_k}\|_\infty$.

Proof. We will use two facts: (a) $\mathcal{T}^{\pi_{k+1}} Q^{\pi_k} \geq \mathcal{T}^{\pi} Q^{\pi_k}\ \forall \pi$, (b) $\mathcal{T}^{\pi_{k+1}} Q^{\pi_k} \leq Q^{\pi_{k+1}}$. Here "$\leq$" and "$\geq$" are element-wise, and we will verify (a) and (b) at the end of this proof. Given (a) and (b), we have

$$Q^\star - Q^{\pi_{k+1}} = (Q^\star - \mathcal{T}^{\pi_{k+1}} Q^{\pi_k}) + (\mathcal{T}^{\pi_{k+1}} Q^{\pi_k} - Q^{\pi_{k+1}}) \leq \mathcal{T}^{\pi^\star} Q^\star - \mathcal{T}^{\pi^\star} Q^{\pi_k}.$$

The first step just adds and subtracts the same quantity. The second step applies (a) and (b) to the two parentheses, respectively. Now

$$\begin{aligned} \|Q^\star - Q^{\pi_{k+1}}\|_\infty &\leq \|\mathcal{T}^{\pi^\star} Q^\star - \mathcal{T}^{\pi^\star} Q^{\pi_k}\|_\infty && (Q^\star - Q^{\pi_{k+1}} \text{ is non-negative}) \\ &\leq \gamma \|Q^\star - Q^{\pi_k}\|_\infty. && (\mathcal{T}^\pi \text{ is a } \gamma\text{-contraction for any } \pi) \end{aligned}$$

Finally we verify (a) and (b) by noting that

$$(\mathcal{T}^{\pi_{k+1}} Q^{\pi_k})(s,a) = \mathbb{E}[\sum_{h=1}^{\infty} \gamma^{h-1} r_h | s_1 = s, a_1 = a, a_2 \sim \pi_{k+1}, a_{3:\infty} \sim \pi_k], \tag{7}$$

$$(\mathcal{T}^{\pi} Q^{\pi_k})(s,a) = \mathbb{E}[\sum_{h=1}^{\infty} \gamma^{h-1} r_h | s_1 = s, a_1 = a, a_2 \sim \pi, a_{3:\infty} \sim \pi_k], \tag{8}$$

$$Q^{\pi_{k+1}}(s,a) = \mathbb{E}[\sum_{h=1}^{\infty} \gamma^{h-1} r_h | s_1 = s, a_1 = a, a_2 \sim \pi_{k+1}, a_{3:\infty} \sim \pi_{k+1}], \tag{9}$$

where $a_{3:\infty}$ denote all the actions from the 3rd time step onwards, and $a_h \sim \pi$ is a shorthand for $a_h = \pi(s_h)$. Since π_{k+1} greedily optimizes Q^{π_k}, (7) $\geq$ (8) and (a) follows. (b) follows due to the policy improvement theorem, i.e., (9) $\geq$ (7) because π_{k+1} outperforms π_k in all states.

(a) 的主要原因是 $\mathcal{T}^{\pi_{k+1}}$ 算子是关于 Q^{π_k} 最优的，它每次都选择后续状态中 Q^{π_k} 最大的那个行动，因此任意给一个其他的策略，都肯定没有它大；

(b) 的主要原因是不等式左边只是对于临近的一步取 greedy policy 得到最优，比如 a_2 取了贪心的策略， $V(s_2) \leftarrow \text{better } V(s_3)$ ，这样新的价值函数会更大；而右边是把这种最优的性质传播开来，即 $V(s_2) \leftarrow \text{better } V(s_3) \quad V(s_3) \leftarrow \text{better } V(s_4) \quad \cdots$ 。

2.2. Value Iteration

Value iteration就是直接迭代Q函数，而不是像policy iteration那样在Q函数和策略交替优化。

Value Iteration computes a series of Q-value functions to directly approximate $Q^\star$, without going back and forth between value functions and policies as in Policy Iteration. Let $Q^{\star,0}$ be the initial value function, often initialized to $\mathbf{0}_{|\mathcal{S}\times\mathcal{A}|}$. The algorithm computes $Q^{\star,h}$ for $h = 1, 2, \ldots, H$ in the following manner:

$$Q^{\star,h} = \mathcal{T} Q^{\star,h-1}. \tag{10}$$

Recall that $\mathcal{T}$ is the Bellman optimality operator defined in Equation 6.

$$(\mathcal{T}f)(s,a) := R(s,a) + \gamma \langle P(s,a), V_f \rangle, \tag{6}$$

where $V_f(\cdot) := \max_{a\in\mathcal{A}} f(\cdot, a)$. This allows us to rewrite Equation 5 in the following concise form, which implies that $Q^\star$ is the fixed point of the operator $\mathcal{T}$:

$$Q^\star = \mathcal{T} Q^\star.$$

这里讲了三个结论

- 当我们学习到了一个Q函数 f ，相对于此Q函数的贪心策略 π_f 的性能 V^{π_f} 可以被函数 f 离最优Q函数 Q^* 的距离bound。
- 每轮迭代，学习到的Q函数距离最优Q函数 Q^* 的距离都指数衰减。
- 如果只考虑有限步的最优价值函数，随着考虑的步数的增多，其相对于最优Q函数 Q^* 的距离指数减少。

第一个结论：策略性能与价值函数误差之间的关系

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

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

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

 □

第一个等式只是添项和减项。第一个等式最后两项变成了第三行里面的内容，只需要利用Q函数和V函数之间的关系 $Q(s,a) = R(s,a) + p(s'|s,\pi_f(s))V(s')$ 即可。第二个不等式再次添项和减项，注意到 π_f 是关于 f 的贪心策略，插入 $f(s,\pi^*(s)) \le f(s,\pi_f(s))$ 即可。第二行的内容得到第三个不等式的第一项，把 $f-Q^*$ 看成一个函数，该函数上两个点的差值被该函数的inf norm的两倍bound。第三行的内容得到第三个不等式的第二项，即期望（线性组合）被该函数的最大值bound。

第二个结论：随着迭代误差指数减小（不动点视角）

这是最常见的一种证明方法。首先，结论如下面(11)式所示。

Value Iteration can be viewed as solving for the fixed point of $\mathcal{T}$, i.e., $Q^\star = \mathcal{T}Q^\star$. The convergence of such iterative methods is typically analyzed by examining the *contraction* of the operator. In fact, the Bellman optimality operator is a γ-contraction under ℓ_∞ norm [1]: for any $f, f' \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$

$$\|\mathcal{T}f - \mathcal{T}f'\|_\infty \le \gamma \|f - f'\|_\infty . \tag{11}$$

To verify, we expand the definition of $\mathcal{T}$ for each entry of $(\mathcal{T}f - \mathcal{T}f')$:

$$\begin{aligned}\left|[\mathcal{T}f - \mathcal{T}f']_{s,a}\right| &= \left|R(s,a) + \gamma\langle P(s,a), V_f\rangle - R(s,a) - \gamma\langle P(s,a), V_{f'}\rangle\right| \\ &\le \gamma\left|\langle P(s,a), V_f - V_{f'}\rangle\right| \le \gamma \|V_f - V_{f'}\|_\infty \le \gamma \|f - f'\|_\infty .\end{aligned}$$

第一个不等号应该是等号吧。第二个不等号利用了 $P(s,a)$ 向量每个元素都非负并且加起来等于1。第三个不等号note里面给了详细解释，其实只需要注意到 V_f 就是取 f 的最大值，因此被 f 的最大值bound，具体解释如下。

The last step uses the fact that $\forall s \in \mathcal{S}$, $|V_f(s) - V_{f'}(s)| = \max_{a\in\mathcal{A}} |f(s,a) - f'(s,a)|$. The easiest way to see this is to assume $V_f(s) > V_{f'}(s)$ (the other direction is symmetric), and let a_0 be the greedy action for f at s. Then

$$|V_f(s) - V_{f'}(s)| = f(s,a_0) - \max_{a\in\mathcal{A}} f'(s,a) \le f(s,a_0) - f'(s,a_0) \le \max_{a\in\mathcal{A}} |f(s,a) - f'(s,a)|.$$

因此可以得到结论，每轮迭代误差都指数减小。

Using the contraction property of $\mathcal{T}$, we can show that as h increases, $Q^\star$ and $Q^{\star,h}$ becomes exponentially closer under ℓ_∞ norm:

$$\|Q^{\star,h} - Q^\star\|_\infty = \|\mathcal{T}Q^{\star,h-1} - \mathcal{T}Q^\star\|_\infty \le \gamma\|Q^{\star,h-1} - Q^\star\|_\infty .$$

加上初始值bound，可以得到经过若干轮迭代之后的误差上界。

Since $Q^\star$ has bounded range (recall Equation 2), for $Q^{\star,0} = \mathbf{0}_{|\mathcal{S}\times\mathcal{A}|}$ (or any function in the same range) we have $\|Q^{\star,0} - Q^\star\|_\infty \le R_{\max}/(1-\gamma)$. After H iterations, the distance shrinks to

$$\|Q^{\star,H} - Q^\star\|_\infty \le \gamma^H R_{\max}/(1-\gamma). \tag{12}$$

由此可以求到经过 $O(\frac{1}{1-\gamma})$ 轮迭代能够得到一个足够精度的最优Q函数估计。

To guarantee that we compute a value function ϵ-close to $Q^\star$, it is sufficient to set

$$H \geq \frac{\log \frac{R_{\max}}{\epsilon(1-\gamma)}}{1-\gamma}. \tag{13}$$

The base of log is e in this course unless specified otherwise. To verify,

$$\gamma^H \frac{R_{\max}}{1-\gamma} = (1-(1-\gamma))^{\frac{1}{1-\gamma}\cdot H(1-\gamma)} \frac{R_{\max}}{1-\gamma} \leq \left(\frac{1}{e}\right)^{\log \frac{R_{\max}}{\epsilon(1-\gamma)}} \frac{R_{\max}}{1-\gamma} = \epsilon.$$

Here we used the fact that $(1-1/x)^x \leq 1/e$ for $x > 1$.

Equation 13 is often referred to as the **effective horizon**. The bound is often simplified as $H = O(\frac{1}{1-\gamma})$, and used as a rule of thumb to translate between the finite-horizon undiscounted and the infinite-horizon discounted settings.[2] From now on we will often use the term "horizon" generically, which should be interpreted as $O(\frac{1}{1-\gamma})$ in the discounted setting.

第三个结论：随着考虑轨迹越长，误差指数减小（Finite-horizon 视角）

这里考虑的是这样一个问题。如果只针对有限步价值函数 $V^{\pi,H}$ ，找到它的最优价值函数 $V^{*,H}$ （对应的策略可以使non-stationary的），它和全局最优的最优价值函数值有什么关系。结论是关系如下（具体的定义见后面贴出来的截图）

we have $\forall s \in \mathcal{S}$,

$$V^\star(s) - \frac{\gamma^H R_{\max}}{1-\gamma} \leq V^{\star,H}(s) \leq V^\star(s),$$

证明很好理解。

Equation 12 can be derived using an alternative argument, which views Value Iteration as optimizing value for a finite horizon. $V^{\star,H}(s)$ is essentially the optimal value for the expected value of the finite-horizon return: $\sum_{t=1}^{H} \gamma^{t-1} r_t$. For any stationary policy π, define its H-step truncated value

$$V^{\pi,H}(s) = \mathbb{E}\Big[\sum_{t=1}^{H} \gamma^{t-1} r_t \mid \pi, s_1 = s\Big]. \tag{14}$$

Due to the optimality of $V^{\star,H}$, we can conclude that for any $s \in \mathcal{S}$ and $\pi : \mathcal{S} \to \mathcal{A}$, $V^{\pi,H}(s) \leq V^{\star,H}(s)$. In particular,

$$V^{\pi^\star,H}(s) \leq V^{\star,H}(s).$$

Note that the LHS and RHS are not to be confused: $\pi^\star$ is the stationary policy that is optimal for infinite horizon, and to achieve the finite-horizon optimal value on the RHS we may need a non-stationary policy (recall the discussion in Section 1.5).

The LHS can be lower bounded as $V^{\pi^\star,H}(s) \geq V^\star(s) - \gamma^H R_{\max}/(1-\gamma)$, because $V^{\pi^\star,H}$ does not include the nonnegative rewards from time step $H+1$ on. (In fact the same bound applies to all policies.) The RHS can be upper bounded as $V^{\star,H}(s) \leq V^\star(s)$: $V^\star$ should dominate any stationary and non-stationary policies, including the one that first achieves $V^{\star,H}$ within H steps and picks up some non-negative rewards afterwards with any behavior. Combining the lower and the upper bounds, we have $\forall s \in \mathcal{S}$,

$$V^\star(s) - \frac{\gamma^H R_{\max}}{1-\gamma} \leq V^{\star,H}(s) \leq V^\star(s),$$

which immediately leads to Equation 12.

主要利用到1）存在稳态的最优策略，它比任何（稳态或者非稳态）的策略都不差；2）每一步奖励都非负，因此在 H 步被截断了，肯定数值会更少，少了多少是有一个上界的。