Notes on Fitted Q-iteration

Nan Jiang

October 16, 2018

【强化学习理论 65】StatisticalRL 8

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）课程的第五讲，这一讲的主要内容是Fitted Q-iteration。

原文传送门

CS598 Note5
nanjiang.cs.illinois.edu

回顾

在很多基础的强化学习书里面我们已经证明了Bellman operator是一个 γ -contraction，它说明在**tabular case**和**infinite sample**的情况下，一定能指数快地（误差随着迭代数指数级下降）学习到最优Q函数。

实际情况中，state space会非常大。在前一讲里面，我们做了state abstraction，把抽象之后的状态空间看做tabular case，在此情况下，分析了finite sample下所找到策略的误差上界。因此，前一讲可以看做是进一步的推广：1）变成了abstracted state space上的tabular case；2）分析了finite sample的情形。

state abstraction可以看做一种histogram regression，即拟合成一个piece-wise constant的函数。这里则考虑一个更为一般的情况，我们不仅仅局限于做histogram regression，而是考虑使用function approximation做更为一般的supervised learning。

过程

1. 定义和假设

为了方便研究上述更为一般的情形，考虑如下定义的Fitted Q-iteration。假设价值函数是一个函数族 $\mathcal{F}\subset\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ 中的函数，每次都在样本上把它朝着Bellman算子作用后的方向更新一步，并且把它投影回到这个函数族中，即

Fitted Q-Iteration (FQI):
[Ernst et al'05]; see also [Gordon'95]

$$f_t = \arg\min_{f\in\mathcal{F}} \sum_{(s,a,r,s')\in D} \left(f(s,a) - \left(r + \gamma \max_{a'\in\mathcal{A}} f_{t-1}(s',a')\right)\right)^2$$

为了便于分析，我们做如下假设：

首先，对于函数族需要满足一下要求。第一个要求说明我们的结论中最多可以含有 $\log|\mathcal{F}|$ ；第二个结论说明规定的函数族 $\mathcal{F}$ 不排除真实的optimal value function；第三个结论说明Bellman算子作用之后，函数仍然在函数族中。注意这一点和每次FQI迭代之后都要往函数族 $\mathcal{F}$ 上投影并不矛盾，因此每次FQI迭代作用的不是这个准确的Bellman算子，而是在样本数估计的Bellman算子。

1. $\mathcal{F}$ is finite but can be exponentially large.
2. Realizability: $Q^\star \in \mathcal{F}$.
3. $\mathcal{F}$ is closed under Bellman update: $\forall f\in\mathcal{F}, \mathcal{T}f\in\mathcal{F}$. (For finite $\mathcal{F}$, this implies realizability.)

下面，再做出如下假设。

4. The dataset $D=\{(s,a,r,s')\}$ is generated as follows: $(s,a)\sim\mu\times U$ (U is uniform over actions), $r\sim R(s,a)$, $s'\sim P(s,a)$. Define the empirical update $\widehat{\mathcal{T}}_{\mathcal{F}}f'$ as

$$\mathcal{L}_D(f;f') := \frac{1}{|D|}\sum_{(s,a,r,s')\in D}\left(f(s,a)-r-\gamma V_{f'}(s')\right)^2.$$
$$\widehat{\mathcal{T}}_{\mathcal{F}}f' := \arg\min_{f\in\mathcal{F}} \mathcal{L}_D(f;f'),$$

where $V_{f'}(s') := \max_{a'} f'(s',a')$. Note that by completeness, $\mathcal{T}f'\in\mathcal{F}$ is the Bayes optimal regressor for the regression problem defined in $\mathcal{L}_D(f;f')$. It will also be useful to define

$$\mathcal{L}_{\mu\times U}(f;f') := \mathbb{E}_D[\mathcal{L}_D(f;f')].$$

5. For any function $g:\mathcal{S}\to\mathbb{R}$, any distribution $\nu\in\Delta(\mathcal{S})$, and $p\geq 1$, define $\|g\|_{p,\nu} := (\mathbb{E}_{s\sim\nu}[|g(s)|^p])^{1/p}$, and let $\|g\|_\nu$ be a shorthand for $\|g\|_{2,\nu}$. Such norms are similarly defined for functions over $\mathcal{S}\times\mathcal{A}$.
6. Let η_h^π be the distribution of s_h under π, that is, $\eta_h^\pi(s) := \Pr[s_h = s \mid s_1\sim\rho_0,\pi]$.
7. μ is exploratory: for a distribution $\nu\in\Delta(\mathcal{S})$ generated by any (non-stationary) policy at any time step (that is, any distribution ν of the form η_h^π where π may be non-stationary),

$$\forall s\in\mathcal{S},\ \frac{\nu(s)}{\mu(s)}\leq C.$$

As a consequence, $\|\cdot\|_\nu \leq \sqrt{C}\|\cdot\|_\mu$. Similarly, when we couple μ with a uniform distribution over $\mathcal{A}$, we have similar results for state-action distributions: $\|\cdot\|_{\nu\times\pi} \leq \sqrt{|\mathcal{A}|C}\,\|\cdot\|_{\mu\times U}$. See slides for example scenarios where C is naturally bounded.

8. Algorithm (simplified for analysis): let $f_0 \equiv \mathbf{0}$ (assuming $\mathbf{0} \in \mathcal{F}$), and for $k \geq 1$, $f_k := \widehat{\mathcal{T}}_{\mathcal{F}} f_{k-1}$.
9. Uniform deviation bound (can be obtained by concentration inequalities and union bound):

$$\forall f, f' \in \mathcal{F}, |\mathcal{L}_D(f; f') - \mathcal{L}_{\mu\times U}(f; f')| \leq \epsilon.$$

(Note: at the end we will show how to obtain fast rates.)

假设4-9归纳如下。

- （假设4）定义了empirical的Bellman算子，其中数据是从一个特定的状态分布 μ （后面会要求这个状态分布是exploratory的）和均匀的动作分布 U 中产生的。即，这里规定了一个具有探索性的behavior policy。
- （假设4和8）给出了FQI的更新规则，即每一步在给定函数族上找到一个函数最小化empirical Bellman residual，更新算子 $\hat{\mathcal{T}}_{\mathcal{F}}$ 可以看做是函数族的投影算子加上empirical Bellman算子。
- （定义5）定义了状态价值函数（V函数）和行动价值函数（Q函数）的metric（norm）。
- （定义6）定义了state visitation probability。
- （假设7）要求采集数据使用的状态分布状态分布 μ 要具有探索性，即对于任意的状态分布 ν 和任意的状态 s ， μ 都能以不低于 $1/C$ 倍 ν 能访问到的概率访问到该状态。紧接着后面给出了两个推论：1） $||f||_\nu = \sqrt{\mathbb{E}_{s\sim\nu}[f^2(s)]} = \sqrt{\sum_s \nu(s) f^2(s)} \leq \sqrt{C}\sqrt{\sum_s \mu(s) f^2(s)} = \sqrt{C}||f||_\mu$ ；2） $||f||_{\nu\times\pi} = \sqrt{\sum_s\sum_a \nu(s)\pi(a) f^2(s,a)} \leq \sqrt{|\mathcal{A}|C}\sqrt{\sum_s\sum_a \mu(s)\frac{1}{|\mathcal{A}|} f^2(s)} = \sqrt{|\mathcal{A}|C}||f||_{\mu\times U}$
- （推论9）当样本足够多的时候，对于任意两个价值函数，他们在数据集上的loss相比于其期望loss都小于某个误差，该误差随着样本的增大而减小。

2. FQI有限步迭代后的策略性能损失上界

有限步迭代之后得到一个价值函数，考虑相对此价值函数的一个greedy策略，考虑该策略的性能相比于最优策略的性能差距，即

Goal Let $\hat{\pi} := \pi_{f_k}$. Derive an upper bound on $v^\star - v^{\hat{\pi}}$.

第一步：策略性能损失表示为 $||Q^* - f_k||$

其中 $v^* = \mathbb{E}_{s_1\sim\rho_0}[V^{\pi^*}(s_1)]$ ， $v^{\hat{\pi}} = \mathbb{E}_{s_1\sim\rho_0}[V^{\hat{\pi}}(s_1)]$ ，它们相减有

$$
\begin{aligned}
v^\star - v^{\hat\pi} &= \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{s \sim \eta_h^{\hat\pi}} [V^\star(s) - Q^\star(s, \hat\pi)] \\
&\le \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{s \sim \eta_h^{\hat\pi}} [Q^\star(s, \pi^\star) - f_k(s, \pi^\star) + f_k(s, \hat\pi) - Q^\star(s, \hat\pi)] \\
&\le \sum_{h=1}^{\infty} \gamma^{h-1} \left(\|Q^\star - f_k\|_{1, \eta_h^{\hat\pi} \times \pi^\star} + \|Q^\star - f_k\|_{1, \eta_h^{\hat\pi} \times \hat\pi} \right) \\
&\le \sum_{h=1}^{\infty} \gamma^{h-1} \left(\|Q^\star - f_k\|_{\eta_h^{\hat\pi} \times \pi^\star} + \|Q^\star - f_k\|_{\eta_h^{\hat\pi} \times \hat\pi} \right). \qquad (1)
\end{aligned}
$$

The last line contains two terms, both in the form of $\|Q^\star - f_k\|_{\nu \times \pi}$. So it remains to bound $\|Q^\star - f_k\|_{\nu\times\pi}$ for any $\nu \times \pi \in \Delta(\mathcal{S} \times \mathcal{A})$ that combines any $\nu \in \Delta(\mathcal{S})$ that satisfies bullet 4 with any $\pi : \mathcal{S} \to \mathcal{A}$.

其中第一个等号其实还需要一些推导： $\sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{s \sim \eta_h^{\pi}} [Q^*(s, \hat\pi)] = \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{s \sim \eta_h^{\pi}} [r(s, \hat\pi) + \gamma \mathbb{E}_{s' \sim P_\pi} [V^*(s')]] = v^{\hat\pi} + \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{s \sim \eta_h^{\pi}} [V^*(s)] - v^*$

第二行的不等式是由于 $\hat\pi$ 是相对于 f_k 的greedy策略；第三行的不等式是由于分开两项求绝对值相加；最后一行的不等式是由于 Jesen 不等式，注意到 $\|\cdot\|_{p,\nu}$ 的定义。

第二步： $\|Q^* - f_k\|$ ***的上界***

下面我们想证明对于任意的状态概率分布和任意的行动概率分布， $\|Q^* - f\|_{\nu\times\pi}$ 都不会太大。

先放出一个引理，它说明V函数的差距和Q函数差距的关系。

Lemma 1. *Define* $\pi_{f,f_k}(s) := \arg\max_{a \in \mathcal{A}} \max\{f(s,a), f_k(s,a)\}$. *Then we have* $\forall \nu \in \Delta(\mathcal{S})$,

$$
\|V_f - V_{f_k}\|_\nu \le \|f - f_k\|_{\nu \times \pi_{f,f_k}}.
$$

Proof.

$$
\begin{aligned}
\|V_f - V_{f_k}\|_\nu^2 &= \sum_{s \in \mathcal{S}} \nu(s) (\max_{a \in \mathcal{A}} f(s,a) - \max_{a' \in \mathcal{A}} f_k(s, a'))^2 \\
&\le \sum_{s \in \mathcal{S}} \nu(s) (f(s, \pi_{f,f_k}) - f_k(s, \pi_{f,f_k}))^2 = \|f - f_k\|_{\nu \times \pi_{f,f_k}}^2
\end{aligned}
$$

其中不等式是由于 π_{f,f_k} 使得 f 和 f_k 中较大的那个最大，另外较小的那个如果还使用 π_{f,f_k} 的话，会导致它更小一些（或不变），因此差距会变大，可以作为一个上界。

Now we can bound $\|Q^\star - f_k\|_{\nu\times\pi}$ using Lemma 1. Define $P(\nu \times \pi)$ as a distribution over $\mathcal{S}$ generated as $s' \sim P(\nu\times\pi) \Leftrightarrow (s,a) \sim \nu \times \pi, s' \sim P(s,a)$, and

$$
\begin{aligned}
\|f_k - Q^\star\|_{\nu\times\pi} &= \|f_k - \mathcal{T} f_{k-1} + \mathcal{T} f_{k-1} - Q^\star\|_{\nu\times\pi} \\
&\le \|f_k - \mathcal{T} f_{k-1}\|_{\nu\times\pi} + \|\mathcal{T} f_{k-1} - \mathcal{T} Q^\star\|_{\nu\times\pi} \\
&\le \sqrt{|\mathcal{A}| C}\, \|f_k - \mathcal{T} f_{k-1}\|_{\mu \times U} + \gamma \|V_{f_{k-1}} - V^\star\|_{P(\nu\times\pi)} \qquad (*) \\
&\le \sqrt{|\mathcal{A}| C}\, \|f_k - \mathcal{T} f_{k-1}\|_{\mu \times U} + \gamma \|f_{k-1} - Q^\star\|_{P(\nu\times\pi) \times \pi_{f_{k-1}, Q^\star}}. \qquad \text{(Lemma 1)}
\end{aligned}
$$

注意到，它变成了 $X_k \le Const + \gamma X_{k-1}$ 的形式，因此可以推导出 X_k 的上界。其中，(*)的推导如下

$$
\begin{aligned}
\|\mathcal{T} f_{k-1} - \mathcal{T} Q^\star\|_{\nu\times\pi}^2 &= \mathbb{E}_{(s,a)\sim\nu\times\pi}\left[((\mathcal{T} f_{k-1})(s,a) - (\mathcal{T} Q^\star)(s,a))^2\right] \\
&= \mathbb{E}_{(s,a)\sim\nu\times\pi}\left[\left(\gamma \mathbb{E}_{s'\sim P(s,a)}[V_{f_{k-1}}(s') - V^\star(s')]\right)^2\right] \\
&\le \gamma^2\, \mathbb{E}_{(s,a)\sim\nu\times\pi, s'\sim P(s,a)}\left[\left(V_{f_{k-1}}(s') - V^\star(s')\right)^2\right] && \text{(Jensen)} \\
&= \gamma^2\, \mathbb{E}_{s'\sim P(\nu\times\pi)}\left[\left(V_{f_{k-1}}(s') - V^\star(s')\right)^2\right] = \gamma^2 \|V_{f_{k-1}} - V^\star\|_{P(\nu\times\pi)}^2.
\end{aligned}
$$

注意到，在infinite sample和 $\mathcal{F} = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ 的情况下， $X_k \le Const + \gamma X_{k-1}$ 中的 $Const$ 一项为零，即每次 $f_k \leftarrow \mathcal{T} f_{k-1}$ 。前面关于 $\mathcal{F}$ closed under Bellman eq和realizability的假设使得 f_k 可以取到 $\mathcal{T} f_{k-1}$ 附近，因此 $Const$ 一项在样本足够多的时候趋向零，下面来证明这件事情。

$$
\begin{aligned}
\|f_k - \mathcal{T} f_{k-1}\|_{\mu\times U}^2 &= \mathcal{L}_{\mu\times U}(f_k; f_{k-1}) - \mathcal{L}_{\mu\times U}(\mathcal{T} f_{k-1}; f_{k-1}) && (\mathcal{L} \text{ squared loss} + \mathcal{T} f_{k-1} \text{ Bayes optimal}) \\
&\le \mathcal{L}_D(f_k; f_{k-1}) - \mathcal{L}_D(\mathcal{T} f_{k-1}; f_{k-1}) + 2\epsilon && (\mathcal{T} f_{k-1} \in \mathcal{F}) \\
&\le 2\epsilon. && (f_k \text{ minimizes } \mathcal{L}_D(\cdot; f_{k-1}))
\end{aligned}
$$

个人认为这里是最难理解的地方。第一个等式也可以弄成下面这种写法（把 ν 换成 $\mu\times U$ ）。

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim\nu}\left[\left(f(s,a) - \left(r + \gamma \max_{a'\in\mathcal{A}} f(s',a')\right)\right)^2\right] \\
&= \mathbb{E}_{(s,a)\sim\nu}\left[(f(s,a) - (\mathcal{T} f)(s,a))^2\right] + \mathbb{E}_{(s,a)\sim\nu}\left[\left((\mathcal{T} f)(s,a) - \left(r + \gamma \max_{a'\in\mathcal{A}} f(s',a')\right)\right)^2\right]
\end{aligned}
$$

注意到当环境为确定性的时候，后面一项为零；并且不考虑平方的部分，后面一项中的第一个变量就等于第二个变量的期望；因此等式左边可以在括号里面添一项减一项，然后展开，交叉项取期望之后为零，最后只剩下等式右边的两项。

解 $X_k \le Const + \gamma X_{k-1}$ 可以得到

$$
\|f_k - Q^\star\|_{\nu\times\pi} \le \frac{1-\gamma^k}{1-\gamma}\sqrt{2|\mathcal{A}|C\epsilon} + \gamma^k \frac{R_{\max}}{1-\gamma}.
$$

第三步：合并

Apply this to Equation (1) and we get

$$v^{\star} - v^{\pi_{f_k}} \le \frac{2}{1-\gamma}\left(\frac{1-\gamma^k}{1-\gamma}\sqrt{2|\mathcal{A}|C\epsilon} + \gamma^k \frac{R_{\max}}{1-\gamma}\right).$$

注意到，这里的 ϵ 代表这相应样本数能够达到的误差上界，由于 $\epsilon \sim O(n^{-1/2})$ ，因此价值函数的收敛速度 $\sim O(n^{-1/4})$ 。

文章还给出了另外一个误差上界，收敛速度会更快 $\sim O(n^{-1/4})$ ，不过误差上界会和函数族的大小 $N = |\mathcal{F}|$ 有关。相关的推导有几处地方实在没太明白，这里就不讲了。

编辑于 2019-08-06