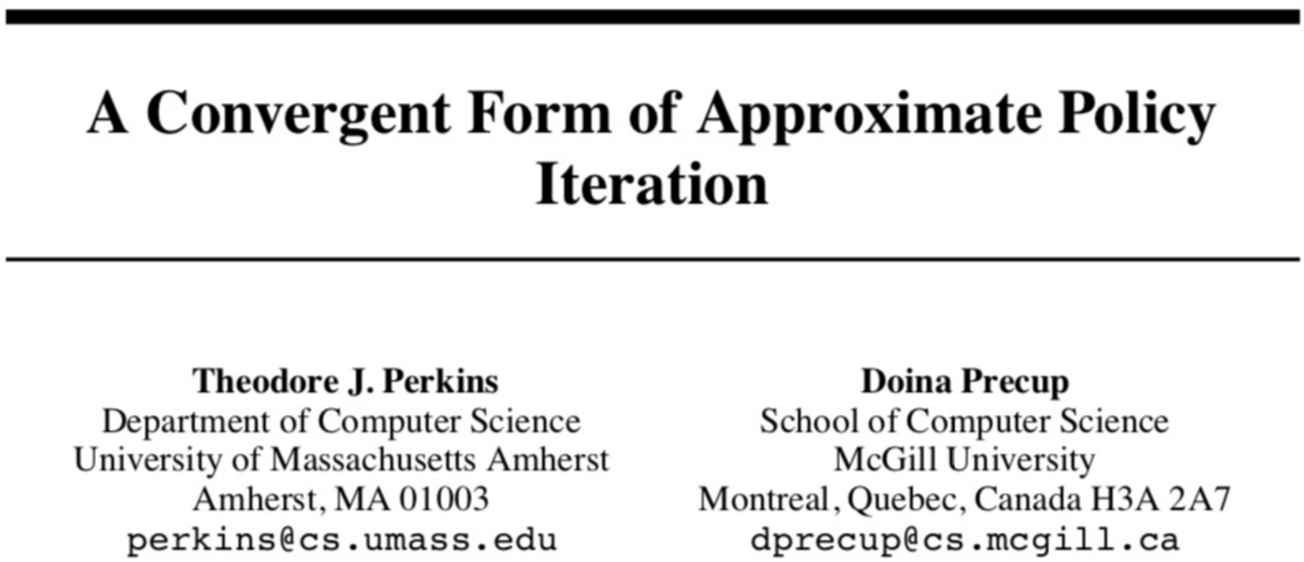

A Convergent Form of Approximate Policy Iteration

Theodore J. Perkins
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
perkins@cs.umass.edu

Doina Precup
School of Computer Science
McGill University
Montreal, Quebec, Canada H3A 2A7
dprecup@cs.mcgill.ca

【强化学习 86】ConvergentAPI

张楚珩
清华大学 交叉信息院博士在读

一篇 2002 年的理论工作（NIPS），证明了一种能够保证收敛的 approximate policy iteration。

原文传送门

Perkins, Theodore J., and Doina Precup. "A convergent form of approximate policy iteration." Advances in neural information processing systems. 2003.

特色

Tabular case 下很多 value-based 算法都能保证收敛到 optimal，在如果对于价值函数使用 function approximation，则不容易保证收敛，比如在实验中可能产生振荡。其主要原因是价值函数产生微小变动时，策略可能产生不连续的变化。比如本来行动 a1 的 Q 值比行动 a2 的 Q 值略高一点，策略应该确定性地选择 a1；当 Q 值变化一点点，使得 a2 的 Q 值刚好比 a1 的 Q 值略高一点，策略这时就会确定性地选择 a2。这里假设 value function 到 policy 的过程连续并且足够有探索性，这样，相应的 API 算法能保证收敛到唯一的不动点。

本文的优势在于对于一个 general 的 greedy operator 进行分析，缺点在于 value function 是针对 linear approximated SARSA 分析的。

过程

1. 设定与假设

MDP

假设离散的状态和行动空间， $m=|S|, n=|A|$ 。

MDP 需要满足如下假设：

- MDP behaves as an irreducible and aperiodic Markov Chain。Irreducible 表面整个状态空间是联通的，不然初始状态给到哪一片无论如何控制都会在这一片了。只有 aperiodic 的设定下，stationary state distribution 才有好的定义。

Policy Improvement

这里假设了一个 general 的 (greedy) operator $\pi=\Gamma(Q)$ ，表示从估计的 Q 函数中导出的 (greedy) policy。对于该算子，有两个假设：

- 要求该算子足够连续，即 c-Lipschitz： $\|\Gamma(Q_1)-\Gamma(Q_2)\|\le c\|Q_1-Q_2\|$ ，其中的 norm 为 L2 norm（Euclidean norm）。可以把策略看做是一个 mn 维的向量（或者是一个 $m\times mn$ 的对角矩阵），把 Q 函数看做是一个 mn 维的向量，由此计算相应的 norm。前面提到了，approximate Q learning 不收敛的原因主要就是 Q-learning 中每一步 greedy policy 选择 argmax，导致该算子不连续。
- 要求该算子足够具有探索，即要求该算子产生的任何策略都要满足 ϵ -soft： $\pi(a|s)\ge\epsilon,\forall s,a$ 。这样的策略族定义为 Π_ϵ 。可见，它是一个 compact set。如果不满足该假设，收敛到的策略就可能和初始策略有关。因为探索不够，找到的只是它能探索到的这部分策略中的解。

Policy evaluation

这一部分使用 linear approximated SARSA，价值函数的表示为 $\hat{Q}=\Phi w$ ，其中 Φ 为 $mn\times k$ 的矩阵，表示 state-aciton pair 的 representation； w 为 k 维的向量，为需要学习的参数。对于 Φ 有以下假设：

- Φ 的每一列是线性无关的。如果有两列线性相关，那么 w 中对应位置可以产生两个绝对数值很大权重，但是产生相同的价值函数。这样就没法 bound w 的 norm 了。

2. 算法

Inputs: initial policy π_0, and policy improvement operator Γ.

for i=0,1,2,... **do**
  Policy evaluation: *Sarsa updates under policy π_i, with linear function approximation.*
  Initialize $\mathbf{w}_i \in \mathbb{R}^k$ arbitrarily.
  With environment in state s_0:
    Choose a_0 according to $\pi_i(s_0,\cdot)$.
    Observe r_0, s_1.
  Repeat for $t=1,2,3,\ldots$ until $\mathbf{w}_i$ converges:
    Choose a_t according to $\pi_i(s_t,\cdot)$.
    $\mathbf{w}_i \leftarrow \mathbf{w}_i + \alpha_t\Phi(s_{t-1},a_{t-1})(r_{t-1}+\gamma\Phi'(s_t,a_t)\mathbf{w}_i - \Phi'(s_{t-1},a_{t-1})\mathbf{w}_i)$
    Observe r_t, s_{t+1}.
  Policy improvement:
  $\pi_{i+1} \leftarrow \Gamma(\Phi\mathbf{w}_i)$.
end for

Figure 1: The version of approximate policy iteration that we study.

3. 定理

Theorem 1 *For any infinite-horizon Markov decision process satisfying Assumption 1, and for any $\epsilon > 0$, there exists $c > 0$ such that if Γ is ϵ-soft and Lipschitz continuous with constant c, then the sequence of policies generated by the approximate policy iteration algorithm in Figure 1 converges to a unique limiting policy $\pi \in \Pi_\epsilon$, regardless of the choice of π_0.*

在上述假设下，对于任何的初始策略，上述算法都能收敛到唯一的不动点。收敛的证明思路是一轮迭代下来有 contraction，即

$$||\Gamma(\hat{Q}^{\pi_1}) - \Gamma(\hat{Q}^{\pi_2})|| \le c||\hat{Q}^{\pi_1} - \hat{Q}^{\pi_2}|| = c||\Phi(w^{\pi_1} - w^{\pi_2})|| \le c_?||\pi_1 - \pi_2||$$

根据前面对于 Γ 算子的假设，第一个不等式直接成立，比较困难的是如何说明相近的策略，其对应的价值函数估计也相近，即最后一个不等式。这一步过程显然和所使用的 policy evaluation 有关，不过在此之前，我们先推导一些和 policy evaluation 无关的 MDP 连续性方面的结论。

4. MDP 连续性

策略相近，则其一步转移概率相近（一步转移概率相对于策略的连续性）

一步转移概率矩阵 P^π 就是 MDP 的转移概率矩阵 P 乘上策略矩阵 Π ，而这里策略矩阵相近，而转移概率矩阵给定，自然可以推出策略到一步转移概率矩阵的连续性。

Lemma 1 *There exists c_P such that for all π_1, π_2, $||P^{\pi_1} - P^{\pi_2}|| \le c_P||\pi_1 - \pi_2||$.*

Proof: Let π_1 and π_2 be fixed, and let $i = (s, a)$ and $j = (s', a')$. Then $|P^{\pi_1}_{i,j} - P^{\pi_2}_{i,j}| = |p^a_{s,s'}(\pi_1(s', a') - \pi_2(s', a'))| \le |\pi_1(s', a') - \pi_2(s', a')| \le \max_{s',a'} |\pi_1(s', a') - \pi_2(s', a')| = \|\pi_1 - \pi_2\|_\infty \le \|\pi_1 - \pi_2\|$. It is readily shown that for any two l-by-l matrices

A and B whose elements different in absolute value by at most ϵ, $\|A - B\| \le \sqrt{l}\epsilon$. Hence, $\|P^{\pi_1} - P^{\pi_2}\| \le \sqrt{mn}\|\pi_1 - \pi_2\|$. □

策略相近，则稳态状态分布相近（稳态状态分布相对于策略的连续性）

Lemma 2 *For any $\epsilon > 0$, there exists c_μ such that for all $\pi_1, \pi_2 \in \Pi_\epsilon$, $\|\mu^{\pi_1} - \mu^{\pi_2}\| \le c_\mu\|\pi_1 - \pi_2\|$.*

Proof: For any $\pi \in \Pi_\epsilon$, let λ^π be the largest eigenvalue of P^π with modulus strictly less than 1. λ^π is well-defined since the transition matrix of any irreducible, aperiodic Markov chain has precisely one eigenvalue equal to one [11]. Since the eigenvalues of a matrix are continuous in the elements of the matrix [9], and since Π_ϵ is compact, there exists $\lambda^{\max} = \max_{\pi \in \Pi_\epsilon} \lambda^\pi = \lambda^{\pi_{\max}} < 1$ for some $\pi_{\max} \in \Pi_\epsilon$. Seneta [12], showed that for any two irreducible aperiodic Markov chains with transition matrices P^1 and P^2 and stationary distributions μ^1 and μ^2, on a state set with l elements, $\|\mu^1 - \mu^2\|_1 \le \frac{l}{|1-\lambda^1|}\|P^1 - P^2\|_\infty$, where λ^1 is the largest eigenvalue of P^1 with modulus strictly less than one. Let $\pi_1, \pi_2 \in \Pi_\epsilon$. $\|\mu^{\pi_1} - \mu^{\pi_2}\| \le \|\mu^{\pi_1} - \mu^{\pi_2}\|_1 \le \frac{mn}{|1-\lambda^{\pi_1}|}\|P^{\pi_1} - P^{\pi_2}\|_\infty \le \frac{mn}{|1-\lambda^{\max}|}\|P^{\pi_1} - P^{\pi_2}\| \le \frac{mn}{|1-\lambda^{\max}|}c_P\|\pi_1 - \pi_2\|$. □

这里的证明最重要的是套用文献[12]中的结论。如何理解 P^π 特征值的物理含义？ P^π 是一步转移概率，而稳态状态分布 μ 是无穷步混合之后的，这中间涉及到 P^π 的高阶矩，因此，其主要特征被最大特征值主导，特征值越大，稳态状态分布 μ 的变化越剧烈。参考：

张楚珩：【强化学习 71】Successor Representation

张楚珩：【强化学习 67】Proto-value Function

5. Linear approximated SARSA 的连续性

SARSA 的不动点由以下方程决定

$$\Phi' D^\pi (I - \gamma P^\pi)\Phi \mathbf{w} = \Phi' D^\pi \mathbf{r} \tag{1}$$

即， $w^\pi = (A^\pi)^{-1} b^\pi$ ，其中

$$A^\pi = \Phi' D^\pi (I - \gamma P^\pi)\Phi \text{ and } b^\pi = \Phi' D^\pi \mathbf{r}$$

下面就分别证明这两部分的连续性：

Lemma 3 *There exist c_b and c_A such that for all π_1, π_2, $\|b^{\pi_1} - b^{\pi_2}\| \le c_b \|\pi_1 - \pi_2\|$ and $\|A^{\pi_1} - A^{\pi_2}\| \le c_A \|\pi_1 - \pi_2\|$.*

Proof: For the first claim, $\|b^{\pi_1} - b^{\pi_2}\| = \|\Phi'(D^{\pi_1} - D^{\pi_2})\mathbf{r}\| \le \|\Phi'\|\|D^{\pi_1} - D^{\pi_2}\|\|\mathbf{r}\| \le c_\mu \|\Phi'\|\|\mathbf{r}\|\|\pi_1 - \pi_2\|$. For the second claim,

$$\begin{aligned}
\|A^{\pi_1} - A^{\pi_2}\| &= \|\Phi'[D^{\pi_1}(I - \gamma P^{\pi_1}) - D^{\pi_2}(I - \gamma P^{\pi_2})]\Phi\| \\
&\le \|\Phi'\|\|D^{\pi_1}(I - \gamma P^{\pi_1}) - D^{\pi_2}(I - \gamma P^{\pi_2})\|\|\Phi\| \\
&= \|\Phi'\|\|D^{\pi_1} - D^{\pi_2} - \gamma D^{\pi_1} P^{\pi_1} + \gamma D^{\pi_2} P^{\pi_2}\|\|\Phi\| \\
&= \|\Phi'\|\|D^{\pi_1} - D^{\pi_2} - \gamma D^{\pi_1}(P^{\pi_1} - P^{\pi_2} + P^{\pi_2}) + \gamma D^{\pi_2} P^{\pi_2}\|\|\Phi\| \\
&= \|\Phi'\|\|D^{\pi_1} - D^{\pi_2} - \gamma D^{\pi_1}(P^{\pi_1} - P^{\pi_2}) - \gamma(D^{\pi_1} - D^{\pi_2})P^{\pi_2}\|\|\Phi\| \\
&\le \|\Phi'\|(\|D^{\pi_1} - D^{\pi_2}\| + \gamma\|D^{\pi_1}\|\|P^{\pi_1} - P^{\pi_2}\| + \gamma\|D^{\pi_1} - D^{\pi_2}\|\|P^{\pi_2}\|)\|\Phi\| \\
&\le ((1+\gamma)c_\mu + \gamma c_P)\|\Phi'\|\|\Phi\|\|\pi_1 - \pi_2\|,
\end{aligned}$$

where the last line follows from Lemmas 1 and 2 and the facts $\|D^\pi\| \le 1$ and $\|P^\pi\| = 1$ for any $\pi \in \Pi_\epsilon$. $\square$

比较有技术含量的是第三行到第五行的拆分。

Lemma 6 *For any $\epsilon > 0$, there exists c_{w2} such that for all $\pi_1, \pi_2 \in \Pi_\epsilon$, $\|\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2}\| \leq c_{w2}\|\pi_1 - \pi_2\|$.*

Proof: Let $\pi_1, \pi_2 \in \Pi_\epsilon$ be arbitrary. From Equation 1, $A^{\pi_1}\mathbf{w}^{\pi_1} = b^{\pi_1}$ and $A^{\pi_2}\mathbf{w}^{\pi_2} = b^{\pi_2}$. Thus:

$$
\begin{aligned}
A^{\pi_1}\mathbf{w}^{\pi_1} - A^{\pi_2}\mathbf{w}^{\pi_2} &= b^{\pi_1} - b^{\pi_2} \\
\Rightarrow\ A^{\pi_1}(\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2} + \mathbf{w}^{\pi_2}) - A^{\pi_2}\mathbf{w}^{\pi_2} &= b^{\pi_1} - b^{\pi_2} \\
\Rightarrow\ A^{\pi_1}(\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2}) + (A^{\pi_1} - A^{\pi_2})\mathbf{w}^{\pi_2} &= b^{\pi_1} - b^{\pi_2} \\
\Rightarrow\ A^{\pi_1}(\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2}) &= (b^{\pi_1} - b^{\pi_2}) - (A^{\pi_1} - A^{\pi_2})\mathbf{w}^{\pi_2} \\
\Rightarrow\ \|A^{\pi_1}(\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2})\| &\leq \|b^{\pi_1} - b^{\pi_2}\| + \|A^{\pi_1} - A^{\pi_2}\|\|\mathbf{w}^{\pi_2}\| \\
\Rightarrow\ c_g\|\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2}\| &\leq c_b\|\pi_1 - \pi_2\| + c_w c_A\|\pi_1 - \pi_2\| \qquad (2) \\
\Rightarrow\ \|\mathbf{w}^{\pi_1} - \mathbf{w}^{\pi_2}\| &\leq c_g^{-1}(c_b + c_w c_A)\|\pi_1 - \pi_2\|
\end{aligned}
$$

The left hand side of Equation 2 follows from Lemmas 5 and 7; the right hand side follows from Lemmas 3 and 4. □

6. 讨论

最后证明的过程把上述结论拼起来即可，需要注意的是，contraction 的条件是系数小于1，要满足这个条件，要求 Γ 算子足够连续，即 c 足够小。这意味着什么呢？注意到如果如果算子越 greedy，那么 c 越大，因此 c 足够小就是说不要特别依赖于估计的 Q 值，并做过度地利用。（可以这样考虑， c 比较小就是在要求差距不太大的 Q 值估计都映射为相似的策略；如果比较 greedy，那么某个动作上的 Q 值稍微大一点点，就会更急切地选择这个动作，而无法产生『相似』的策略）。

这篇文章的待改进的地方：1）局限于 discrete case；2）linear function approximation。

编辑于 2019-08-06