Safe Policy Iteration

eo Pirotta MATTEO.PIROTTA@POLIMI.I
cello Restelli MARCELLO.RESTELLI@POLIMI.I
sio Pecorino ALESSIO.PECORINO@MAIL.POLIMI.I
ele Calandriello DANIELE.CALANDRIELLO@MAIL.POLIMI.I
Elect., Inf., and Bioeng., Politecnico di Milano, piazza Leonardo da Vinci 32, I-20133, Milan, ITALY

【强化学习 88】SPI

张楚珩
清华大学 交叉信息院博士在读

SPI 全称为 Safe Policy Iteration，ICML 2013。

原文传送门

Pirotta, Matteo, et al. "Safe policy iteration." International Conference on Machine Learning. 2013.

特色

Policy iteration 分为两步，policy evaluation 和 policy improvement，如果两步都能够准确进行，（即，policy evaluation 能准确估计到 Q^{π_k}，同时 policy improvement 能找到相对于 Q^{π_k} 的 greedy policy），那么能够保证在有限次迭代内找到最优策略（我记得大概最多 $|S||A|$ 步）。但是通常的情况是这两步都不能保证准确地进行，比如在 model-free 的条件下 policy evaluation 需要大量的样本才能估计准确，同时，使用参数化的策略会导致能够表示出来的策略空间并不是全部的策略空间，从而无法准确地进行 policy improvement 步。

文章提到若干 approximate policy iteration（API）算法提出减小 policy evaluation 步骤的误差（approximation error）的方法，然后再采用相应的 greedy operator 来做 policy improvement。我想，比如在之前的价值函数估计基础上更新得到 on-policy 的价值函数应该也是一种减小 approximation error 的方法。

Conservative policy iteration（CPI）提出在 policy improvement step 中做小步的更新，这样相近策略的状态分布相近，从而产生相近的价值函数。这样能保证 policy improvement。

文章在 (Kakade&Langford, 2002) 的基础上改进得到新的 policy improvement bound，从而得到新的 CPI 算法。

过程

1. Policy improvement bound

策略 π' 相比于策略 π 的 policy improvement 写作 $J_{\pi'} - J_{\pi}$，即最大化 $J_{\pi'}$。

$$J_D^\pi = \sum_{s\in\mathcal{S}} D(s)V^\pi(s) = \sum_{s\in\mathcal{S}} d_D^\pi(s) \sum_{a\in\mathcal{A}} \pi(a|s)\mathcal{R}(s,a),$$

由于 $d_\mu^{\pi'}$ 无法得到，因此用 $\mathbb{A}_{\pi,\mu}^{\pi'}$ 来近似估计这个 policy improvement 的量。

$$A_\pi^{\pi'}(s) = \sum_{a\in\mathcal{A}} \pi'(a|s)A^\pi(s,a)$$

$$\mathbb{A}_{\pi,\mu}^{\pi'} = \sum_{s\in\mathcal{S}} d_\mu^\pi(s)A_\pi^{\pi'}(s).$$

这里『近似』要求 $d_\mu^{\pi'} \sim d_\mu^\pi$ ，该关系可以从下面的引理看出（注意蓝色框）：

Lemma 3.3. *(Kakade & Langford, 2002)*
For any stationary policies π and π' and any starting state distribution μ:

$$J_\mu^{\pi'} - J_\mu^\pi = \mathbf{d}_\mu^{\pi'^T}\mathbf{A}_\pi^{\pi'}.$$

下面的引理说明，只要一步转移概率矩阵相差不大，则它们的状态分布相差也不大（model-based version）。

Lemma 3.1. *Let π and π' be two stationary policies for an infinite horizon MDP M with state transition matrix $\mathbf{P}$. The L_1–norm of the difference between their γ–discounted future state distributions under starting state distribution μ can be upper bounded as follows:*

$$\left\|\mathbf{d}_\mu^{\pi'} - \mathbf{d}_\mu^\pi\right\|_1 \leq \frac{\gamma}{1-\gamma}\left\|\mathbf{P}^{\pi'} - \mathbf{P}^\pi\right\|_\infty \left\|\left(\mathbf{I} - \gamma\mathbf{P}^{\pi'}\right)^{-1}\right\|_\infty$$

上述 $P^{\pi'} = \Pi^{\pi'}P$ ，需要知道 model P 。下面提供一个 model-free 版本，不过更 loose（注意到多了一个 $\beta = 1/(1-\gamma)$ ）。

Corollary 3.2. *Let π and π' two stationary policies for an infinite horizon MDP M. The L_1–norm of the difference between their γ–discounted future state distributions under starting state distribution μ can be upper bounded as follows:*

$$\left\|\mathbf{d}_\mu^{\pi'} - \mathbf{d}_\mu^{\pi}\right\|_1 \leq \frac{\gamma}{(1-\gamma)^2}\left\|\mathbf{\Pi}^{\pi'} - \mathbf{\Pi}^{\pi}\right\|_\infty$$

本文的主要贡献就是提供了一个 model-free 版本的更为紧的定理，不过描述的和上面推论是一样的事情。它主要基于以下观察：

Lemma 3.4. *(Haviv & Heyden, 1984, Corollary 2.4) For any vector $\mathbf{d}$ and any vector $\mathbf{c}$ such that $\mathbf{c}^T\mathbf{e} = 0$,*

$$|\mathbf{c}^T\mathbf{d}| \leq \|\mathbf{c}\|_1 \frac{\Delta\mathbf{d}}{2},$$

where $\Delta\mathbf{d} = \max_{i,j} |\mathbf{d}_i - \mathbf{d}_j|$.

对于不等式左边，我们一般会想到用 L1 和 Linf 来 bound，但是考虑到 c 的特殊性质，可以有这个更紧的 bound。 $c^T e = 0$ 这个条件让我们联想到 $\sum_a \pi(a|s) A^\pi(s,a) = 0$ 。文章给出如下定理：

Theorem 3.5. *For any stationary policies π and π' and any starting state distribution μ, given any baseline policy π_b, the difference between the performance of π' and the one of π can be lower bounded as follows:*

$$J_\mu^{\pi'} - J_\mu^{\pi} \geq \mathbf{d}_\mu^{\pi_b T}\mathbf{A}_\pi^{\pi'} - \frac{\gamma}{(1-\gamma)^2}\left\|\mathbf{\Pi}^{\pi'} - \mathbf{\Pi}^{\pi_b}\right\|_\infty \frac{\Delta\mathbf{A}_\pi^{\pi'}}{2}$$

这里还用到了一个任意的策略 π_b ，当然也可以就用 $\pi_b = \pi$ ，这样 $d_\mu^{\pi_b}$ 是可以获得的。

下面给出一个更松的推论：

Corollary 3.6. *For any stationary policies π and π' and any starting state distribution μ, the difference between the performance of π' and the one of π can be lower bounded as follows:*

$$J_\mu^{\pi'} - J_\mu^{\pi} \geq \mathbf{d}_\mu^{\pi\,T} \mathbf{A}_\pi^{\pi'} - \frac{\gamma}{(1-\gamma)^2} \left\| \mathbf{\Pi}^{\pi'} - \mathbf{\Pi}^{\pi} \right\|_\infty^2 \frac{\|\mathbf{q}^\pi\|_\infty}{2}.$$

2. Exact safe policy iteration

Exact case 指的是价值函数能够被准确估计的情况（approximation error = 0）。

当 Q^π 能够被准确估计的时候，选择一个相对于 Q^π 的贪心策略肯定能有 policy improvement，但是可能存在更好的策略。当 Q^π 不能够被准确估计的时候，选择一个关于它的贪心策略，可能该策略比之前的策略更差。SPI 的想法就是每次去最大化 policy improvement 的下界，并且当不能保证 policy improvement 的时候停止更新。

文章提出两种算法：unique-parameter safe policy improvement（USPI）和 multiple-parameter safe policy improvement（MSPI）。

USPI

USPI 是从 Theorem 3.5 推导出来的，和 (Kakade&Langford 02) 一样，它们都有个共同的特征，如果存在一个 π 使得 $\mathbb{A}^{\pi}_{\pi,\mu}$ 大于零，那么总存在一个在 π 和 π 之间的策略使得 policy improvement 的下界大于零。因此，只需要找出这样一个合适的步长即可。根据该思路，考虑一个 conservative policy update：

$$\pi' = \alpha\overline{\pi} + (1-\alpha)\pi,$$

有如下 policy improvement 的保证：

Corollary 4.1. *If $\mathbb{A}^{\overline{\pi}}_{\pi,\mu} \geq 0$, then, using $\alpha^* = \frac{(1-\gamma)^2 \mathbb{A}^{\overline{\pi}}_{\pi,\mu}}{\gamma \|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi}\|_\infty \Delta\mathbf{A}^{\overline{\pi}}_{\pi}}$, we set $\alpha = \min(1, \alpha^*)$, so that when $\alpha^* \leq 1$ the following policy improvement is guaranteed:*

$$J_\mu^{\pi'} - J_\mu^{\pi} \geq \frac{(1-\gamma)^2 {\mathbb{A}^{\overline{\pi}}_{\pi,\mu}}^2}{2\gamma \left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi}\right\|_\infty \Delta\mathbf{A}^{\overline{\pi}}_{\pi}},$$

and when $\alpha^* > 1$, we perform a full update towards the target policy $\overline{\pi}$ with a policy improvement equal to the one specified in Theorem 3.5.

类似地， (Kakade&Langford 02) 的 policy improvement 下界可以写为

$$J_{\mu}^{\pi'} - J_{\mu}^{\pi} \geq \frac{(1-\gamma)^2 \mathbb{A}_{\pi,\mu}^{\overline{\pi}}{}^2}{8\frac{\gamma}{1-\gamma}}.$$

由于

$$\left\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi}\right\|_{\infty} \Delta\mathbf{A}_{\pi}^{\overline{\pi}} \leq \frac{4}{1-\gamma}$$

因此，本文的下界更紧，但是计算会更复杂一些。

MSPI

MSPI 还是考虑 conservative policy update，但是对于每个 state 使用不同的步长，即

$$\pi'(a|s) = \alpha(s)\overline{\pi}(a|s) + (1-\alpha(s))\pi(a|s), \forall s, a, \text{ where } \alpha(s) \in [0,1], \forall s.$$

由于对每个状态单独考虑，因此效果可能会更好，但是如果考虑 Theorem3.5 中的 bound 会比较复杂，因此对每个状态只考虑 Corollary3.6 中的简化的 bound。

Corollary 4.2. *Let $\mathcal{S}_{\pi}^{\overline{\pi}}$ be the subset of states where the advantage of policy $\overline{\pi}$ over policy π is positive: $\mathcal{S}_{\pi}^{\overline{\pi}} = \{s \in \mathcal{S} | A_{\pi}^{\overline{\pi}}(s) > 0\}$.*
The bound in Corollary 3.6 is optimized by taking $\alpha(s) = 0, \forall s \notin \mathcal{S}_{\pi}^{\overline{\pi}}$ and $\alpha(s) = \min\left(1, \frac{\overline{\Upsilon}^*}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1}\right)$, $\forall s \in \mathcal{S}_{\pi}^{\overline{\pi}}$, where $\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1 = \sum_{a \in \mathcal{A}} |\overline{\pi}(a|s) - \pi(a|s)|$ and $\overline{\Upsilon}^*$ is the value that maximizes the following function:

$$B(\overline{\Upsilon}) = \sum_{s \in \mathcal{S}_{\pi}^{\overline{\pi}}} \min\left(1, \frac{\overline{\Upsilon}}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1}\right) \mathbf{d}_{\mu}^{\pi} \mathbf{A}_{\pi}^{\overline{\pi}} - \overline{\Upsilon}^2 \frac{\gamma}{(1-\gamma)^2} \frac{\|\mathbf{q}^{\pi}\|_{\infty}}{2}$$

这里比较麻烦的点在于每个状态上的步长 $\alpha(s)$ 不能够被 closed-form 写出来，不过 $B(\hat{\Gamma})$ 的形态是有一定的规律的，因此能够在 $O(|S|(|A|+\log|S|))$ 时间内找到最大值。形象地来说，找 $\hat{\Gamma}$ 的过程就是找一个阈值的过程，我们希望在所有 S_{π}^{0} 的状态中，如果 π 和 π 差距比较小，那么我们就分配 $\alpha(s)=1$ ，如果差距比较大，就分配一个 $\alpha(s)<1$ 。

3. Approximate safe policy iteration

在 exact 情况下，不管用 Theorem3.5 还是 Corollary3.6 都需要准确地估计 ΔA 或者 $\|q^{\pi}\|_{\infty}$ ，在 approximate 的情形下都需要很多样本去做估计。为了绕过这件事情，我们直接把这两项替换为相应的 upper bound，得到如下更松的 bound：

$$J_{\mu}^{\pi'} - J_{\mu}^{\pi} \geq \mathbb{A}_{\pi,\mu}^{\pi'} - \frac{\gamma}{2(1-\gamma)^3} \left\| \mathbf{\Pi}^{\pi'} - \mathbf{\Pi}^{\pi} \right\|_{\infty}^{2}, \qquad (1)$$

同时，使用 $\mathbb{A}_{\pi,\mu}^{\pi'}$ 的一个 ϵ -approximate 的估计 $\hat{\mathbb{A}}_{\pi,\mu}^{\pi'}$ ，得到相应的 aUSPI 和 aMSPI 算法：

(1) Choose an initial policy at random. (2) Select the target policy $\overline{\pi} \in \hat{\Pi} \subseteq \Pi$ through the maximization of a sample-based version of the Q-function. (3) Produce an $\frac{\epsilon}{3(1-\gamma)}$−accurate estimate of the average advantage: $\hat{\mathbb{A}}_{\pi}^{\overline{\pi}}$. (4) If $\hat{\mathbb{A}}_{\pi}^{\overline{\pi}}$ is larger than $\frac{2\epsilon}{3(1-\gamma)}$, then compute (according to the USPI or the MSPI approach) the new policy for the next iteration using the bound in Eq. 1. For instance, in the case of aUSPI, the value of the parameter α, to take into account the approximation error, is $\alpha = \frac{(1-\gamma)^3(\hat{\mathbb{A}}_{\pi}^{\overline{\pi}} - \frac{\epsilon}{3(1-\gamma)})}{\gamma\|\mathbf{\Pi}^{\overline{\pi}} - \mathbf{\Pi}^{\pi}\|_{\infty}^{2}}$. (5) When $\hat{\mathbb{A}}_{\pi}^{\overline{\pi}} \leq \frac{2\epsilon}{3(1-\gamma)}$ the algorithm stops returning the current policy.

发布于 2019-08-17

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