TEX: Regret Bounds for Policy Iteration Using Expert Predi

-Yadkori [1] Peter L. Bartlett [2] Kush Bhatia [2] Nevena Lazić [3] Csaba Szepesvári [4] Gel

【强化学习 82】POLITEX

张楚珩
清华大学 交叉信息院博士在读

POLITEX 全称是 POLicy ITeration with EXpert advice。

原文传送门

Abbasi-Yadkori, Yasin, et al. "Politex: Regret Bounds for Policy Iteration using Expert Prediction." (2019).

特色

使用 expert problem 研究中的 exponential weighted average（EWA）算法，得到了一个强化学习策略，并且得到了该算法的 regret bound。注意到对 regret bound 的分析等价于 finite sample analysis。该算法可以使用任意的 Q 函数估计。文章还证明了对于 Q 函数的线性函数拟合，使用 least-squares policy evaluation（LSPE）方法来估计参数，有相应的估计误差的上界。

过程

1. Expert problem and EWA

这里的 expert problem 设定如下。这是一个多轮的游戏，每一轮分为如下几步： A 个专家分别给出对于未来的预测建议（advice） $f_{t,i} \in \mathcal{D}$ ，属于 decision space；玩家依照一定的策略，挑选出一个专家 $I_t \in [A]$ 并且采取它给出的建议；环境揭示该时刻的情形 $y_t \in \mathcal{O}$ ，属于 outcome space；接下来，对于专家给出的每一个建议，都能够通过损失函数计算出相应的损失，该损失可以使用一个向量来表示 $l_t \in [0,1]^A$ 。

Expert problem 一般考虑环境可能是 adversarial，甚至 expert 也是不靠谱的，如果目标是减小玩家的损失，那么只要 expert 总是给出不靠谱的“建议”，那么玩家不可能有很小的损失。然而，退一步地，expert problem 的目标是最小化玩家损失和任意一个专家（或者可以理解为最好的那个专家）损失差距。注意到，我们一开始并不知道那个专家是最好的，甚至专家也可能是 adversarial ，根据你做的决策而改变。具体来说，优化目标为最小化 regret：

$$\mathrm{Regret}_T = \sum_{t=1}^{T} (l_{t,I_t} - l_{t,i})$$

表示截止第 T 轮时，玩家累积受到的损失相比于第 i 个专家受到的累积损失的差距。

EWA 每次按照一定概率选择相应的专家，即使用一个随机策略，选择第 i 个专家的概率为 $\pi_t(i) \propto \exp(-\eta \sum_{s=1}^{t-1} l_{s,i})$ 。即，如果该专家历史上表现都比较好，就会以比较高的概率选择该专家。这个策略也叫做 Boltzmann policy。如果使用该策略，可以证明，当时间较长的时候，玩家遭受的损失基本上和最好专家遭受的损失差不多。具体地，见如下定理：

Theorem 4.2 (Corollary 4.2, Cesa-Bianchi & Lugosi (2006)). *Set* $\eta = \sqrt{8\log(A)/T}$. *Then, regardless of how the environment plays, for any* $0 < \delta < 1$ *and* $i \in [A]$, *with probability* $1 - \delta$, *the regret of EWA with the above choice of* η *satisfies* $\mathfrak{R}_{T,i} \le \sqrt{T\log(A)/2} + \sqrt{T\log(1/\delta)/2}$.

注意到，参数 η 的选择要求我们实现知道游戏需要玩多少轮，这个条件可以被放松，代价是大概会给 regret bound 带来常数倍的影响。只要 regret 增长速度小于 T （比如这里是 $\sqrt{T}$ ），就说明当轮数变多的时候，每轮和最好专家之间产生的 regret 差距逐渐趋向于零。

2. POLITEX 算法框架

受到 EWA 算法的启发，文章也希望采取类似的 Boltzmann policy，并且试图分析一下能不能得到相应的 regret bound。

Algorithm 1 POLITEX: POLicy ITeration using EXperts

Input: phase length $\tau > 0$, initial state x_0

Set $\widehat{Q}_0(x,a) = 0 \ \forall x, a$
for $i := 1, 2, \ldots,$ **do**
  Set $\pi_i(a|x) \propto \exp\left(-\eta \sum_{j=0}^{i-1} \widehat{Q}_j(x,a)\right)$
  Execute π_i for τ time steps and collect data

$$\mathcal{Z}_i = \{(x_t, a_t, c_t, x_{t+1})\}_{t=\tau(i-1)+1}^{\tau i}$$

  Compute $\widehat{Q}_i$ from $\mathcal{Z}_1, \ldots, \mathcal{Z}_i, \pi_1, \ldots, \pi_i$
end for

注意到，这里研究的是离散动作空间的问题，可以认为每个 action 都是一个专家。在某状态下，选择某个 action（专家）的概率取决于该 action 的历史表现。

该算法分为 $i \in [B]$ 个 iteration，每个 iteration 都需要采样 τ 个样本，假设总共采样 T 个样本。采样到第 $t \in [T]$ 个样本的时候，对应的策略记为 $\pi_{(t)}$ ；第 $i \in [B]$ 个 iteration 的策略记为 π_i 。按照上述算

法，采样 T 步，每一步的 cost 记做 $c(x_t, a_t)$ 。假设有一个最优策略 π^* ，采样 T 步，每一步的 cost 记做 $c(x_t^*, a_t^*)$ 。

和 expert problem 不同的是：在 expert problem 里面，即使没有选择其他的专家，也告知假如选择其他专家而产生的 loss；而在强化学习里面，并不会告知其他 action 产生的 cumulative cost。因此这里会使用估计的价值函数来代替实际的 cumulative cost，即算法中的 $\hat{Q}$ 。每一轮， $\hat{Q}_t$ 的目标是估计相对于策略 π_t 的价值函数。

该算法的目标就是最小化如下 regret：

$$\Re_T = \sum_{t=1}^{T}[c(x_t, a_t) - c(x_t^*, a_t^*)]$$

3. MDP 设定

该工作使用的是 infinite horizon average cost 的设定，而不是我个人比较习惯的 infinite horizon discounted reward 的设定，不过两者基本上是等价的。

$$\lambda_\pi := \lim_{T\to\infty} \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} c(x_t^\pi, a_t^\pi)\right].$$

相当于是 performance measure。

$$V_\pi(x) := \lim_{T\to\infty} \mathbf{E}\left[\sum_{t=1}^{T}(c(x_t^\pi, a_t^\pi) - \lambda_\pi) \mid x_1^\pi = x\right]$$

$$Q_\pi(x,a) = c(x,a) - \lambda_\pi + \mathbf{E}[V_\pi(x')|x,a]\,, \qquad (2)$$

V 函数和 Q 函数的定义。

4. Regret bound：重要的部分

第一步，对于 regret 进行如下拆分。这也算是分析的常见思路，要 bound 一个随机变量，把它拆成“均值”和“随机变量 - 均值”的形式，NNQL 分析也是类似的思路。

$$\begin{aligned}
\Re_T &= \overline{\Re}_T + V_T + W_T\,, \quad \text{where}\\
\overline{\Re}_T &= \sum_{t=1}^{T}(\lambda_{\pi_{(t)}} - \lambda_{\pi^*})\,,\\
V_T &= \sum_{t=1}^{T}(c(x_t,a_t) - \lambda_{\pi_{(t)}})\,,\\
W_T &= \sum_{t=1}^{T}(\lambda_{\pi^*} - c(x_t^*, a_t^*))\,.
\end{aligned}$$

“随机变量 - 均值”总是能在各种假设条件被 bound 的，因此这里面最关键的还是需要 bound $\overline{\mathfrak{R}}_T$ 。关于这一项的上界，先放出结论：

Theorem 4.1. *Let* $E = \lfloor T/\tau \rfloor$ *and fix* $0 < \delta < 1$. *Let* $\varepsilon(\delta, \tau) > 0$ *and* $Q_{\max} > 0$ *and* $b \in \mathbb{R}$ *be such that for any* $i \in [E]$, *with probability* $1 - \delta$,

$$\|Q_{\pi_i} - \widehat{Q}_i\|_{\nu^*},\ \|Q_{\pi_i} - \widehat{Q}_i\|_{\mu^* \otimes \pi_i} \leq \varepsilon(\delta, \tau) \qquad (6)$$

and $\widehat{Q}_i(x, a) \in [b, b + Q_{\max}]$ *for any* $(x, a) \in \mathcal{X} \times \mathcal{A}$. *Letting* $\eta = \sqrt{8 \log(A)/E}/Q_{\max}$, *with probability* $1 - \delta$, *the regret of* POLITEX *relative to the reference policy* π^* *satisfies*

$$\overline{\mathfrak{R}}_T \leq 2T\, \varepsilon(\delta/(2E), \tau) + E^{1/2} \tau Q_{\max} S_\delta(A, \mu^*)\,,$$

where

$$S_\delta(A, \mu^*) = \sqrt{\frac{\log(A)}{2}} + \left\langle \mu^*, \sqrt{\frac{\log(2/\delta) + \log(1/\mu^*)}{2}} \right\rangle$$

假设部分显然是必要的，因为 regret 的衡量标准是基于策略在 MDP 上性能的，而策略的输入只是对于性能的估计 $\hat{Q}$ ，因此要有有用的结论，起码需要假设性能的估计比较准确。这里使用 $\varepsilon(\delta,\tau)$ 来表征其估计的准确性， δ 表示相应的概率， τ 表示估计 Q 函数所使用的样本长度，显然，样本越多，estimation error 越小，估计越准确。

下面开始证明：

首先观察到

$$\lambda_{\pi_{(t)}} - \lambda_{\pi^*} = \langle \mu_{\pi^*}, Q_{\pi_{(t)}}(\cdot, \pi_{(t)}) - Q_{\pi_{(t)}}(\cdot, \pi^*) \rangle\,.$$

在 discounted reward setting 下，该公式我们非常的熟悉，就是 TRPO paper 里面的 (1) 式，在 02 年 Kakade&Lanford 的 paper 里面也有提到。考虑到策略使用的是 $\hat{Q}$ ，只有替换成 $\hat{Q}$ 才能讨论 EWA 相关的结论。因此，做如下拆分：

$$\overline{\mathfrak{R}}_{T,1} = \sum_{t=1}^{T} \langle \mu_{\pi^*}, \widehat{Q}_{\pi_{(t)}}(\cdot, \pi_{(t)}) - \widehat{Q}_{\pi_{(t)}}(\cdot, \pi^*) \rangle$$

$$\overline{\mathfrak{R}}_{T,2} = \sum_{t=1}^{T} \langle \mu_{\pi^*}, Q_{\pi_{(t)}}(\cdot, \pi_{(t)}) - \widehat{Q}_{\pi_{(t)}}(\cdot, \pi_{(t)}) \rangle$$

$$+ \sum_{t=1}^{T} \langle \mu_{\pi^*}, \widehat{Q}_{\pi_{(t)}}(\cdot, \pi^*) - Q_{\pi_{(t)}}(\cdot, \pi^*) \rangle$$

第一项可以套用 EWA 的结论，第二项可以套用定理中的假设。

对于第二项，使用定理中的假设，再加上 union bound（注意到，要求对于 E 个不同的策略，要求*每一个*都以大概率小于某个值），即可得到：

$$\overline{\mathfrak{R}}_{T,2} \leq 2T\varepsilon(\delta, \tau)$$

对于第一项，目标是套用 EWA 的结论。EWA 中损失函数需要在 $[0,1]$ 区间内，因此这里对 $\hat{Q}$ 做 scale；同时每个 iteration 的 τ 步都是一样的，因此可以拆为 $\sum_{t=1}^{T} \to \sum_{t=1}^{\tau} \sum_{t=1}^{E}$ 。令， $l_{i,a} = (\hat{Q}_i(x,a) - b)/Q_{\max}$ ，套用 EWA 结论，有：

$$\widetilde{\mathfrak{R}}_T(x) := \sum_{i=1}^{E} \langle \pi_i(\cdot|x), \ell_i \rangle - \langle \pi^*(\cdot|x), \ell_i \rangle$$

$$\leq \sqrt{E \log(A)/2} + \sqrt{E \log(1/\delta)/2}$$

下面是个人感觉比较难的一步（费了我一上午。。文章就单说使用 union bound。。）：对于每个状态 x 都大概率有关于 $R(x)$ 的上界，特别地， $Pr[R(x) \geq \epsilon] \leq \exp(-\epsilon^2)$ ，那么对于 $\mathbb{E}_{x\sim\mu}[R(x)]$ 的上界应为多少？

首先，对于任意的 $\epsilon(x), s.t. \sum_x \epsilon(x)\mu(x) = \epsilon$ ，都有

$$\begin{aligned}
Pr[\mathbb{E}_{x\sim\mu} R(x) \geq \epsilon] &\leq Pr[\exists x, \mu(x) > 0, R(x) \geq \epsilon(x)] \\
&\leq \sum_x Pr[R(x) \geq \epsilon(x)] \quad \text{(union bound)} \\
&\leq \sum_x \exp(-\epsilon^2(x)) = \delta = \sum_x \mu(x)\delta \\
&\leq \mathbb{E}_{x\sim\mu}[\sqrt{\log(1/\mu(x)\delta)}] \quad \text{(Let } \epsilon(x) = \sqrt{\log(1/\mu(x)\delta)}\text{)}
\end{aligned}$$

最后可以得到

$$\langle \mu^*, \widetilde{\mathfrak{R}}_T \rangle \leq \sqrt{E \log(A)/2} + \langle \mu^*, \sqrt{E \log(1/\delta\mu^*)/2} \rangle \,.$$

$$\overline{\mathfrak{R}}_{T,1} \;\leq\; \tau Q_{\max} \langle \mu^*, \widetilde{\mathfrak{R}}_T \rangle$$

5. Regret bound：次要的部分

次要的部分就是要 bound V_T, W_T 两项：

$$V_T = \sum_{t=1}^{T} (c(x_t, a_t) - \lambda_{\pi_{(t)}}) \,,$$

$$W_T = \sum_{t=1}^{T} (\lambda_{\pi^*} - c(x_t^*, a_t^*)) \,.$$

从 λ 的定义可以看到：如果每一轮中样本数目趋向无穷的时候， V_T 应该趋向零；如果总样本数 T 趋向无穷的时候， W_T 趋向零。在样本有限的时候，状态分布更多地类似于初始状态分布（暂态）可能有别于稳态状态，这会导致 V_T, W_T 两项绝对数值较大。因此，需要做暂态状态消失地足够快的假设：

Assumption A2 ***(Uniformly fast mixing)*** There exists a constant $\kappa > 0$ such that for any distribution ν',

$$\sup_{\pi} \|(\nu_\pi - \nu')^\top H_\pi\|_1 \leq \exp(-1/\kappa) \|\nu_\pi - \nu'\|_1 \,.$$

在此假设下能够得到结论：

Lemma 4.4. *Let Assumption A2 hold. With probability at least* $1 - \delta$*, we have that*

$$|W_T| \leq \kappa + 4\kappa \sqrt{2T \log(2/\delta)} \,,$$

$$|V_T| \leq E\kappa + 4E\kappa \sqrt{2\tau \log(2T/\delta)}.$$

暂态消失地足够快时， κ 比较小， V_T, W_T 两项的数值也比较小；同时该误差也随着轨迹长度的增长而 sublinear 地增长，即步数越长，平均每步产生的误差越小；注意到 V_T 描述的策略，每隔 τ 步都会变化，因此还需要套一个 union bound，得到的结果含有 E 。

6. 如何估计 Q 函数

注意到，策略需要使用以前所有的 Q 函数估计，这一点不很友好。但是如果价值函数的估计使用

的是线性函数拟合，就可以累积地维护历史上 Q 估计的和了。文章分析了一个特殊的线性函数估计方法 LSPE，得到了前面公式中 $\epsilon(\delta,\tau)$ 的具体表达形式。联合起来，得到了 LSPE + POLITEX 的 regret bound：

Theorem 5.2. *Under Assumptions A1, A3, A2, and A4, for some constant C, for any fixed T large enough, the regret of* POLITEX *with LSPE with respect to a reference policy π^* is bounded with probability at least $1-\delta$ as*

$$\mathfrak{R}_T \leq 2\varepsilon_0 T + CT^{3/4} \cdot \left(\sqrt{2\log A} + \kappa \sqrt{\log(4T/\delta)} + \langle \mu^*, \sqrt{\log(1/\delta\mu^*)/2} \rangle + \frac{\kappa}{\sigma^3} \sqrt{d \log(4d/\delta)} \right)$$

其中一直存在的误差 ε_0 表示的是线性函数拟合的 approximation error；除此之外，regret 的增长速度大致为 $\tilde{O}(T^{3/4})$ 。这一部分的分析在附录中，很长，个人不太感兴趣，就没仔细看。

在实际中，可以使用神经网络来作为 Q 的估计，这样需要把历史上的神经网络全部存下来。实际操作中可以只存最近的若干个。

回头看一下 POLITEX，如果策略不使用历史上所有 Q 函数估计的和，而是只使用上一轮的 Q 函数估计，该方法几乎是一个 soft 版本（使用 Boltzmann policy）的 Q-learning。相比于 Q-learning，它使用了前些轮估计的 Q 函数，实验效果看，更稳定。

在实验上，使用神经网络估计 Q 函数，相比于 DQN 效果更好。实验环境为 Ms Pacman。

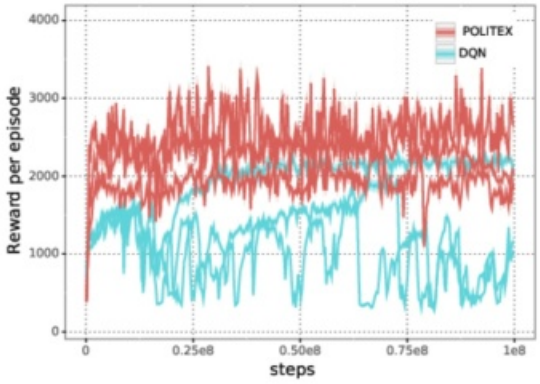

Figure 2. Ms Pacman game scores obtained by the agents at the end of each game. The plots are based on three runs of each algorithm with different random seeds.

发布于 2019-07-21

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