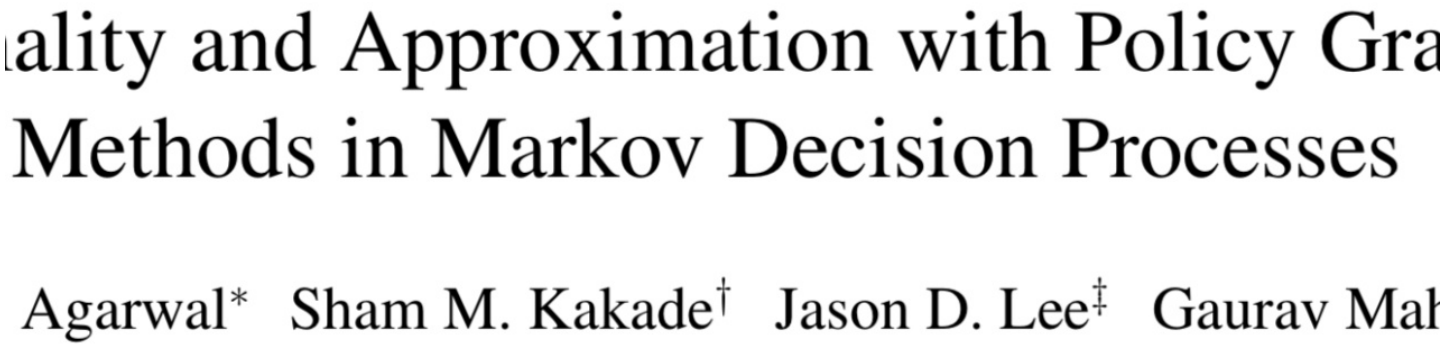

【强化学习 89】PG Theory 1

张楚珩
清华大学 交叉信息院博士在读

是非常新的一篇理论工作，从理论上分析 policy gradient 算法相关的各种性质。这篇文章比较长（有 71 页），我们分两次讲，这是第一部分。

原文传送门

Agarwal, Alekh, et al. "Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes." arXiv preprint arXiv:1908.00261 (2019).

特色

看到 Kakade 的强化学习文章，别的不用说，只需要读读读就好了！个人感觉，基于策略的算法（包括策略梯度类算法）含有 global optimality 和 finite sample analysis 的理论比较少，这里就出现了一篇比较系统讲这件事情的文章，非常值得一读。

基于策略的算法算法每一轮都使用当前的策略进行采样，这样并不能保证所有的状态都被访问到，更不能保证均匀地访问到所有的状态，在这种情况下，一般很难有 global optimality。（考虑有些可能很『好』的区域根本没有被策略访问到）对比基于值函数的方法（比如 Q-learning、certainty-equivalence），它们只需要零散的 transition，这样就能要求在足够多的样本上把任意状态访问足够多次，从而有相应的 contraction。

本文分为两大部分：第一个部分为对 tabular policy parameterization 的分析，即 policy class 一定包含 optimal policy，在这一部分文章中给出了 global optimality 的分析；第二部分为对 restricted policy class 的分析，即 policy class 不一定包含 optimal policy，在这一部分中文字给出了 agnostic result，即相比于该 policy class 中最优 policy 的 loss。这里先讲第一部分。

这一部分主要贡献为：1）引入 distribution mismatch coefficient，来描述状态空间上访问不均匀的情形；2）引入 gradient domination，说明只要策略梯度较小，则策略接近全局最优；3）给出了不同 parameterization 和不同设定下的分析。

过程

1. 综述

Policy gradient 算法每次使用当前的策略进行采样，因此每次对于策略参数的更新都更加侧重于当前策略访问频率较高的状态。而要想得到一个最优策略，需要对所有状态（更准确地说是从初始状态分布出发能够 reachable 的所有状态）都有足够的访问频率。该问题就是强化学习里面非常重要的『探索』（exploration）问题。对于该问题，本文引入 distribution mismatch coefficient 来对其进行描述。

这一部分主要分析四种情况：1）direct parameterized policy classes；2）direct softmax parameterization；3）softmax parameterization with relative entropy regularization；4）softmax parameterization with natural policy gradient ascent。分别对应下表中的含有『Thm』的项目。

Algorithm	**Iteration complexity**
Projected Gradient Ascent on Simplex (Thm 4.2)	$O\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^6\epsilon^2}\left\lVert\frac{d_\rho^{\pi^*}}{\mu}\right\rVert_\infty^2\right)$
Policy Gradient, softmax parameterization (Thm 5.1)	asymptotic
Policy Gradient + relative entropy regularization, softmax parameterization (Thm 5.3)	$O\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2}{(1-\gamma)^6\epsilon^2}\left\lVert\frac{d_\rho^{\pi^*}}{\mu}\right\rVert_\infty^2\right)$
MDP Experts Algorithm [Even-Dar et al., 2009]	$O\left(\frac{\ln\lvert A\rvert}{(1-\gamma)^4\epsilon^2}\right)$
MD-MPI Geist et al. [2019]	$\frac{2+(1-\gamma)\ln\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon}$
Natural Policy Gradient (NPG) softmax parameterization (Thm 5.6)	$\frac{2}{(1-\gamma)^2\epsilon}$

Table 1: **Iteration Complexities for the Tabular Case:** A summary of the number of steps required by different algorithms to return a policy π satisfying $\mathbb{E}_{s\sim\rho}[V^\star(s) - V^\pi(s)] \leq \epsilon$. First three algorithms optimize the objective $\mathbb{E}_{s\sim\mu}[V^\pi(s)]$, where μ is the initial starting state distribution; as we show, this implies a guarantee with respect to all starting state distributions ρ. The MDP has $|\mathcal{S}|$ states, $|\mathcal{A}|$ actions, and discount factor $0 \leq \gamma < 1$; the worst case ratio $\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty = \max_s\left(\frac{d_\rho^{\pi^*}(s)}{\mu(s)}\right)$ is termed the *distribution mismatch coefficient*, where, roughly speaking, $d_\rho^{\pi^*}(s)$ is the fraction of time spent in state s when executing the optimal policy (see (2)), when started in the state $s_0 \sim \rho$. Both the MDP Experts Algorithm [Even-Dar et al., 2009] and the MD-MPI algorithm Geist et al. [2019] (see Corollary 3) imply guarantees for the same update rule as the NPG for the softmax parameterization. NPG directly optimizes $\mathbb{E}_{s\sim\rho}[V^\pi(s)]$ and incurs no distribution mismatch coefficient dependence. See Section 2 for further discussion.

该表中的 iteration complexity 指需要多少次迭代能够收敛到距离不动点 $\star$ 的位置。由于每一次迭代中所用到的价值函数都认为是准确的，而不需要样本去做估计，因此 iteration complexity 可以看做是 sample complexity 的 lower bound。

2. 设定

和通常的强化学习设定一样，这里只讲一下稍微特殊一点的部分。

State visitation distribution

$$d_{s_0}^{\pi}(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr^{\pi}(s_t = s|s_0) \tag{2}$$

$$d_{\rho}^{\pi}(s) = \mathbb{E}_{s_0\sim\rho}\left[d_{s_0}^{\pi}(s)\right] .$$

能够归一化为 1 的一般叫做 distribution，否则（没有前面的常数项系数）一般叫 frequency。

Policy parameterization

对于离散状态空间和离散动作空间，这篇文章研究三种策略参数化的方法：

- *Direct parameterization:* The policies are parameterized by

$$\pi_\theta(a|s) = \theta_{s,a}, \tag{5}$$

where $\theta \in \Delta(\mathcal{A})^{|\mathcal{S}|}$, i.e. θ is subject to $\theta_{s,a} \geq 0$ and $\sum_{a \in \mathcal{A}} \theta_{s,a} = 1$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

- *Softmax parameterization:* For unconstrained $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$,

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}. \tag{6}$$

The softmax parameterization is also complete.

- *Restricted parameterizations:* We also study parametric classes $\{\pi_\theta | \theta \in \Theta\}$ that may not contain all stochastic policies. Here, the best we may hope for is an agnostic result where we do as well as the best policy in this class.

Direct parameterization 的好处在于对于离散的状态和行动空间，它有绝对完整的表示能力，坏处在于每次梯度更新完之后，不能保证 $\theta \in \Delta(\mathcal{A})^{|\mathcal{S}|}$ ，因此需要 projection step。Softmax parameterization 的好处在于任意 $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ 的参数都代表合法的策略，坏处在于不能表示确定性的策略。

Lemmas and definitions

强化学习的目标是最大化 $f(\theta) = V^{\pi_\theta}(s)$ ，下面引理说明该目标相对于 θ 不是 concave optimization。

Lemma 3.1. *There is an MDP M (described in Figure 1) such that the optimization problem $V^{\pi_\theta}(s)$ is not concave for both the direct and softmax parameterizations.*

这里 Figure 1 就不放了，其实很简单，如果整个 MDP 是奖励稀疏的，很有可能在大多是从参数空间上， $f(\theta)$ 都是平的，因此不是 concave。

接下来是非常熟悉、反复出现的引理：performance difference lemma！

Lemma 3.2. *(The performance difference lemma [Kakade and Langford, 2002]) For all policies π, π' and states s_0,*

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi}_{s_0}} \mathbb{E}_{a \sim \pi(\cdot|s)} \left[A^{\pi'}(s,a) \right].$$

最后是 distribution mismatch coefficient 的定义

Definition 3.3 (Distribution mismatch coefficient)**.** Given a policy π and measures $\rho, \mu \in \Delta(\mathcal{S})$, we refer to $\left\| \frac{d^{\pi}_{\rho}}{\mu} \right\|_\infty$ as the *distribution mismatch coefficient* of π relative to μ.

其中， μ 表示算法进行 rollout 时使用的初始状态分布， ρ 表示用来进行评价算法性能时使用的初

始状态分布。

3. Gradient Domination

本文的目标是证明 global optimality，而策略梯度算法中我们只知道一些局部的信息，如何使用局部的信息来判断全局的最优性呢？这就需要 gradient domination property 了，简单说来，即：

$$f(\theta^\star) - f(\theta) = O(G(\theta)),$$

where $\theta^\star \in \operatorname{argmax}_{\theta' \in \Theta} f(\theta')$ and where $G(\theta)$ is some suitable scalar notion of first-order stationarity, which can be considered a measure of how large the gradient is

假设有了上属性值，那么假设某个位置上的梯度较小，那么就可以推出这个地方距离全局最优差距不大。下面引理说明了强化学习中策略梯度满足该性质。

Lemma 4.1 (Gradient domination). *For the direct policy parameterization (as in (5)), for all state distributions $\mu, \rho \in \Delta(\mathcal{S})$, we have*

$$\begin{aligned} V^\star(\rho) - V^\pi(\rho) &\leq \left\| \frac{d_\rho^{\pi^\star}}{d_\mu^\pi} \right\|_\infty \max_{\bar{\pi}} (\bar{\pi} - \pi)^\top \nabla_\pi V^\pi(\mu) \\ &\leq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty \max_{\bar{\pi}} (\bar{\pi} - \pi)^\top \nabla_\pi V^\pi(\mu), \end{aligned}$$

where the max is over the set of all policies, i.e. $\bar{\pi} \in \Delta(\mathcal{A})^{|S|}$, and where $\frac{d_1}{d_2}$ denotes componentwise division.

证明的过程从 performance difference lemma 开始，定理中的第一行右边第一项与策略 π 有关，考虑到稳态分布和初始分布的关系，能够进一步放缩得到与 π 无关的不等式（只有梯度项与 π 有关）。注意到该引理和策略的参数化方式无关，因为它只管 $\nabla_\pi V^\pi(\mu)$ ，不管策略的参数化形式。文章把该引理放到 direct policy parameterization 中，而我单独把这个引理列出来。

4. Direct policy parameterization

Direct policy parameterization 的策略梯度为：

$$\frac{\partial V^\pi(\mu)}{\partial \pi(a|s)} = \frac{1}{1-\gamma} d_\mu^\pi(s) Q^\pi(s,a), \tag{7}$$

考虑一个 projected gradient ascent update：

$$\pi^{(t+1)} = P_{\Delta(\mathcal{A})^{|S|}}(\pi^{(t)} + \eta \nabla V^{(t)}(\mu)), \tag{9}$$

where $P_{\Delta(\mathcal{A})^{|S|}}$ is the projection on the probability simplex $\Delta(\mathcal{A})^{|S|}$.

有如下的 global optimality + finite iteration result：

Theorem 4.2. *The projected gradient ascent algorithm (9) on $V^{\pi}(\mu)$ with stepsize $\eta = \frac{(1-\gamma)^3}{2\gamma|\mathcal{A}|}$ satisfies for all distributions $\rho \in \Delta(\mathcal{S})$,*

$$\min_{t<T}\left\{V^{\star}(\rho) - V^{(t)}(\rho)\right\} \leq \epsilon \quad \textit{whenever} \quad T > \frac{64\gamma|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^6\epsilon^2}\left\|\frac{d_{\rho}^{\pi^{\star}}}{\mu}\right\|_{\infty}^2$$

这里会用到 Lemma4.1，梯度大小的衡量为 ，-stationary，即

$$\text{for all } \pi_\theta + \delta \in \Delta(\mathcal{A})^{|S|} \text{ and } \|\delta\|_2 \leq 1, \delta^{\top}\nabla V^{\pi_\theta}(\mu) \leq \epsilon.$$

换句话说，即 stationarity implies optimality。这其中包含的 distribution mismatch coefficient 是必须的，同样考虑一个 sparse reward 的环境，如果策略们都 exponentially hard 地得到非零奖励，那么这些策略附近也是 stationary 的，但是并没有 optimality。利用该思路，可以证明相应的 lower bound，即至少存在某种 MDP 使得 iteration 的数量大致上得有这么多。

证明的过程需要看一下，特别是 projection operator 是如何分析和处理的。

Lemma 4.3 (Vanishing gradients at suboptimal parameters). *Consider the chain MDP of Figure 2, with $\gamma = H/(H+1)$, and with the direct policy parameterization (with $3|\mathcal{S}|$ parameters, as described in the text above). Suppose θ is such that $0 < \theta < 1$ (componentwise) and $\theta_{s,a_1} < 1/4$ (for all states s). For all $k \leq \frac{H/8}{\ln 2H}$, we have $\|\nabla^k V^{\pi_\theta}(s_0)\| \leq (1/4)^{H/2}$, where $\nabla^k V^{\pi_\theta}(s_0)$ is a tensor of the k_{th} order derivatives of $V^{\pi_\theta}(s_0)$ and the norm is the operator norm [3]. Furthermore, $V^{\star}(s_0) - V^{\pi_\theta}(s_0) \geq (H+1)/4 - (1/4)^{H+1}$.*

该引理给出了一个探索难的例子，说明如果无视探索的问题（即，distribution mismatch coefficient 很大），在梯度较小的情况下，仍然离最优策略较远。

5. Softmax parameterization without regularization

在 softmax parameterization 下，策略梯度可以写为：

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_{\mu}^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi_\theta}(s,a) \tag{10}$$

考虑一个正常的 gradient ascent 更新：

$$\theta^{(t+1)} = \theta^{(t)} + \eta\nabla V^{(t)}(\mu). \tag{11}$$

文章给出了 optimality 的结果，没有 finite iteration 的结果，实际上该方法可能收敛地指数级地慢。

Theorem 5.1 (Global convergence for softmax parameterization). *Assume we follow the gradient descent update rule as specified in Equation 11 and that the distribution μ is strictly positive i.e. $\mu(s) > 0$ for all states s. Suppose $\eta \leq \frac{(1-\gamma)^2}{5}$, then we have that for all states s, $V^{(t)}(s) \to V^{\star}(s)$ as $t \to \infty$.*

该结果要求算法 rollout 的初始状态分布就要布满整个状态空间，但是原则上来讲，transition dynamics 应该也能把采样的状态带到各个需要的状态上，这个条件不是很必须。不过文章的证明还是用到了这个条件。

注意到

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi_\theta}(s,a) = \pi_\theta(a|s)\frac{\partial V^{\pi_\theta}(\mu)}{\partial \pi_\theta(a|s)}.$$

即，和 direct policy parameterization 不一样的地方在于， $\nabla_\theta V$ 较小不代表 $\nabla_\pi V$ 很小。因此， $\nabla_\theta V$ 较小的时候，更新就比较缓慢了，但实际上可能离最优策略还比较远。这就是该算法可能到最后收敛缓慢的原因。后面会讲到，这个问题可以通过增加 entropic regularization （代价是，根据正则项强度的不同，会产生大小不同的 bias）和使用 natural policy gradient 来解决。

6. Softmax parameterization with relative entropy regularization

首先，区分一下常用的两种熵正则：entropy / relative entropy。

Relative entropy 指相对于一个均匀分布的相对熵，即，目标函数变为：

$$L_\lambda(\theta) := V^{\pi_\theta}(\mu) - \frac{\lambda}{|\mathcal{S}|}\sum_s \mathrm{KL}(\mathbf{Uni}, \pi_\theta(\cdot|s)) \tag{12}$$

$$= V^{\pi_\theta}(\mu) + \frac{\lambda}{|\mathcal{S}|\,|\mathcal{A}|}\sum_{s,a}\log \pi_\theta(a|s) + \lambda \log|\mathcal{A}|\,, \tag{13}$$

由于后面要对它求导，因此最后一个常数项可以被扔掉。

Entropy 指的是策略自身的熵：

$$\frac{1}{|\mathcal{S}|}\sum_s H(\pi_\theta(\cdot|s)) = \frac{1}{|\mathcal{S}|}\sum_s\sum_a -\pi_\theta(a|s)\log \pi_\theta(a|s).$$

区别在于 relative entropy 中蓝色框的部分为 $1/|\mathcal{A}|$ 。可以看出，relative entropy 对于确定性策略（在其他行动上产生非常小的 probability）的惩罚更大。文章这里分析 relative entropy 正则项的情形。

考虑如下更新策略：

$$\theta^{(t+1)} = \theta^{(t)} + \eta\nabla L_\lambda(\theta^{(t)}). \tag{14}$$

有如下 global optimality + finite iteration result：

Theorem 5.3. *(Relative entropy regularization) Let* $\beta_\lambda := \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$. *Starting from any initial* $\theta^{(0)}$, *consider the updates (14) with* $\lambda = \frac{\epsilon(1-\gamma)}{2\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty}$ *and* $\eta = 1/\beta_\lambda$. *Then for all starting state distributions* ρ, *we have*

$$\min_{t<T}\left\{V^\star(\rho) - V^{(t)}(\rho)\right\} \le \epsilon \quad \textit{whenever} \quad T \ge \frac{320|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\,\epsilon^2}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty^2$$

注意到一点，需要先知道所需要的精度 ϵ，再确定正则项的强度 λ，并且如果要求的精度高，则需要少加一些正则，这样才能够保证 bias 足够小，相应的代价是需要更多的 iteration。

7. Softmax parameterization with natural policy gradient

考虑如下更新 natural policy gradient 的更新公式

$$F_\rho(\theta) = \mathbb{E}_{s\sim d_\rho^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\nabla\log\pi_\theta(a|s)\left(\nabla\log\pi_\theta(a|s)\right)^\top\right]$$
$$\theta^{(t+1)} = \theta^{(t)} + \eta F_\rho(\theta^{(t)})^\dagger \nabla V^{(t)}(\rho), \tag{15}$$

where $M^\dagger$ denotes the Moore-Penrose pseudoinverse of the matrix M.

softmax parameterization + NPG updates 能够得到 closed-form 的 update rule，文章直接引用了一下结论：

Lemma 5.5. *For the softmax parameterization (6), the NPG updates (15) take the form:*

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma}A^{(t)} \quad \textit{and} \quad \pi^{(t+1)}(a|s) = \pi^{(t)}(a|s)\frac{\exp(\eta A^{(t)}(s,a)/(1-\gamma))}{Z_t(s)},$$

where $Z_t(s) = \sum_{a\in\mathcal{A}}\pi^{(t)}(a|s)\exp(\eta A^{(t)}(s,a)/(1-\gamma))$.

这个是怎么来的日后还需要看一下，特别是 NPG update rule 应该怎么分析。

以下引理证明了 policy improvement：

Lemma 5.7 (Improvement lower bound for NPG). *For the iterates* $\pi^{(t)}$ *generated by the NPG updates (15), we have for all starting state distributions* μ

$$V^{(t+1)}(\mu) - V^{(t)}(\mu) \ge \frac{(1-\gamma)}{\eta}\mathbb{E}_{s\sim\mu}\ln Z_t(s) \ge 0.$$

其证明过程感觉还是很巧妙的。

Proof: First, let us show that $\ln Z_t(s) \geq 0$. To see this, observe:

$$\ln Z_t(s) = \ln \sum_a \pi^{(t)}(a|s) \exp(\eta A^{(t)}(s,a)/(1-\gamma))$$
$$\geq \sum_a \pi^{(t)}(a|s) \ln \exp(\eta A^{(t)}(s,a)/(1-\gamma)) = \frac{\eta}{1-\gamma} \sum_a \pi^{(t)}(a|s) A^{(t)}(s,a) = 0.$$

where the inequality follows by Jensen's inequality on the concave function $\ln x$ and the final equality uses $\sum_a \pi^{(t)}(a|s)A^{(t)}(s,a) = 0$. Using $d^{(t+1)}$ as shorthand for $d_\mu^{(t+1)}$, the performance difference lemma implies:

$$V^{(t+1)}(\mu) - V^{(t)}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{(t+1)}} \sum_a \pi^{(t+1)}(a|s) A^{(t)}(s,a)$$
$$= \frac{1}{\eta} \mathbb{E}_{s\sim d^{(t+1)}} \sum_a \pi^{(t+1)}(a|s) \ln \frac{\pi^{(t+1)}(a|s) Z_t(s)}{\pi^{(t)}(a|s)}$$
$$= \frac{1}{\eta} \mathbb{E}_{s\sim d^{(t+1)}} \mathrm{KL}(\pi_s^{(t+1)} || \pi_s^{(t)}) + \frac{1}{\eta} \mathbb{E}_{s\sim d^{(t+1)}} \ln Z_t(s)$$
$$\geq \frac{1}{\eta} \mathbb{E}_{s\sim d^{(t+1)}} \ln Z_t(s) \geq \frac{1-\gamma}{\eta} \mathbb{E}_{s\sim\mu} \ln Z_t(s).$$

where the last step uses that $d^{(t+1)} = d_\mu^{(t+1)} \geq (1-\gamma)\mu$, componentwise by (2), and that $\ln Z_t(s) \geq 0$. ■

以下定理说明了 softmax parameterization + NPG updates 的性质（global optimality + finite iteration）

Theorem 5.6 (Global convergence for Natural Policy Gradient Ascent). *Suppose we run the NPG updates (15) using $\rho \in \Delta(\mathcal{S})$ and with $\theta^{(0)} = 0$. Fix $\eta > 0$. For all $T > 0$, we have:*

$$V^{(T)}(\rho) \geq V^*(\rho) - \frac{\ln |\mathcal{A}|}{\eta T} - \frac{1}{(1-\gamma)^2 T}.$$

In particular, setting $\eta \geq (1-\gamma)^2 \ln |\mathcal{A}|$, we see that NPG finds an ϵ-optimal policy in a number of iterations that is at most:

$$T \leq \frac{2}{(1-\gamma)^2 \epsilon},$$

证明过程用到前面的 policy improvement lemma，考虑然后创造能够前后抵消的项。

$$
\begin{aligned}
V^{\pi^\star}(\rho) - V^{(t)}(\rho) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\star}\sum_a \pi^\star(a|s)A^{(t)}(s,a) \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\sum_a \pi^\star(a|s)\ln\frac{\pi^{(t+1)}(a|s)Z_t(s)}{\pi^{(t)}(a|s)} \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\left(\mathrm{KL}(\pi_s^\star||\pi_s^{(t)}) - \mathrm{KL}(\pi_s^\star||\pi_s^{(t+1)}) + \sum_a \pi^\star(a|s)\ln Z_t(s)\right) \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\left(\mathrm{KL}(\pi_s^\star||\pi_s^{(t)}) - \mathrm{KL}(\pi_s^\star||\pi_s^{(t+1)}) + \ln Z_t(s)\right),
\end{aligned}
$$

where we have used the closed form of our updates from Lemma 5.5 in the second step.

By applying Lemma 5.7 with $d^\star$ as the starting state distribution, we have:

$$
\frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\ln Z_t(s) \le \frac{1}{1-\gamma}\Big(V^{(t+1)}(d^\star) - V^{(t)}(d^\star)\Big)
$$

which gives us a bound on $\mathbb{E}_{s\sim d^\star}\ln Z_t(s)$.

Using the above equation and that $V^{(t+1)}(\rho) \ge V^{(t)}(\rho)$ (as $V^{(t+1)}(s) \ge V^{(t)}(s)$ for all states s by Lemma 5.7), we have:

$$
\begin{aligned}
V^{\pi^\star}(\rho) - V^{(T-1)}(\rho) &\le \frac{1}{T}\sum_{t=0}^{T-1}(V^{\pi^\star}(\rho) - V^{(t)}(\rho)) \\
&\le \frac{1}{\eta T}\sum_{t=0}^{T-1}\mathbb{E}_{s\sim d^\star}(\mathrm{KL}(\pi_s^\star||\pi_s^{(t)}) - \mathrm{KL}(\pi_s^\star||\pi_s^{(t+1)})) + \frac{1}{\eta T}\sum_{t=0}^{T-1}\mathbb{E}_{s\sim d^\star}\ln Z_t(s) \\
&\le \frac{\mathbb{E}_{s\sim d^\star}\mathrm{KL}(\pi_s^\star||\pi^{(0)})}{\eta T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\Big(V^{(t+1)}(d^\star) - V^{(t)}(d^\star)\Big) \\
&= \frac{\mathbb{E}_{s\sim d^\star}\mathrm{KL}(\pi_s^\star||\pi^{(0)})}{\eta T} + \frac{V^{(T)}(d^\star) - V^{(0)}(d^\star)}{(1-\gamma)T} \\
&\le \frac{\ln|\mathcal{A}|}{\eta T} + \frac{1}{(1-\gamma)^2 T}.
\end{aligned}
$$

The proof is completed using that $V^{(T)}(\rho) \ge V^{(T-1)}(\rho)$.

其中最后一个不等式应该略去了 R_{max} 。

编辑于 2019-08-22

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