Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes

Alekh Agarwal* Sham M. Kakade† Jason D. Lee‡ Gaurav Mahajan§

【强化学习 90】PG Theory 2

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，这是第二部分。

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

前面说到本文分为两大部分：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（agnostic results）。这里讲的是第二部分。

---

1. 概述

这一部分主要研究参数化的策略族

$$\Pi = \{\pi_\theta | \theta \in \Theta \subseteq \mathbb{R}^d\} \tag{16}$$

这一部分主要研究两种情况，第一种情况是该函数族对于参数无约束，即 $\Theta = \mathbb{R}^d$ ；第二种情况是该函数族对于参数有约束，这样每次做梯度更新之后都需要进行投影（projection）的操作。所研究的函数族不一定包含 optimal policy，通常假设函数族的表示能力是比较局限的，即 $d \ll |S||A|$ 。

2. NPG for unconstrained policy classes

2.1 问题描述

考虑一个无约束的优化问题

$$\max_{\theta \in \mathbb{R}^d} V^{\pi_\theta}(\rho)$$

每一轮更新参数

$$\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}. \tag{18}$$

2.2 Compatible function approximation

Sutton 书里面讲 policy gradient 的时候就提到了 compatible 这个概念。compatibility 考虑如下的问题，policy gradient 公式中出现价值函数，在 model-free 的情形下，我们需要估计这个价值函数（function approximation），那么采用什么样的 function approximation 才能够使得其 approximation error 不影响策略梯度的估计？换句话说，就是找到估计 $\hat{A}^{\pi_\theta}$ ，使得

$$\begin{aligned}\nabla V^{\pi_\theta}(\mu) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\mathbb{E}_{a\sim\pi^\theta(\cdot|s)}\big[\nabla \log \pi_\theta(a|s) A^{\pi_\theta}(s,a)\big] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\mathbb{E}_{a\sim\pi^\theta(\cdot|s)}\big[\nabla \log \pi_\theta(a|s) \widehat{A}^{\pi_\theta}(s,a)\big].\end{aligned}$$

引用 (Sutton, 1999) 的结论，对应的 compatible function approximation 为：

$$\widehat{A}^{\pi_\theta}(s,a) = w^\star(\theta)\cdot \nabla \log \pi_\theta(a|s),$$

其中

$$w^\star(\theta) \in \mathrm{argmin}_w\, L_\nu(w;\theta).$$

为 compatible function approximation error L_ν 的 minimizer：

$$L_\nu(w;\theta) = \mathbb{E}_{s,a\sim\nu}\left[\big(A^\theta(s,a) - w\cdot\nabla\log\pi_\theta(a|s)\big)^2\right], \quad L_\nu^\star(\theta) := \min_w L_\nu(w;\theta). \tag{19}$$

$$\nu(s,a) = d_\mu^{\pi_\theta}(s)\pi_\theta(a|s)$$

注意到， $\hat{A}^{\pi_\theta}$ 的形式和策略的 parameterization 有关。

2.3 Natural policy gradient

文章直接引用了 (Kakade, 2011) 的结论，natural policy gradient 每一步的更新就是 compatible function approximation error L_ν 的 minimizer 。

$$F_\rho(\theta)^\dagger \nabla V^\theta(\rho) \in \mathrm{argmin}_w\, L_\nu(w;\theta), \tag{20}$$

where $\nu(s,a) = d_\rho^{\pi_\theta}(s)\pi_\theta(a|s)$.

2.4 Minimal compatible function approximation error measures policy expressivity

对于 2.2 中的结论，我们可以换个角度理解： A^{π_θ} 的 contour 可能很复杂，但是在策略的 expressivity 不够的情况下（即 L_ν 较大），这么精细的价值函数估计中的信息也没法被 distill 到策略中。下面举一个例子来说明，考虑一个 linear softmax policy：

$$\pi_\theta(a|s) = \frac{\exp(\theta \cdot \phi_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta \cdot \phi_{s,a'})}$$

通过计算不难得出，

$$\nabla_\theta \log \pi_\theta(a|s) \;=\; \widetilde{\phi}_{s,a}, \text{ where } \widetilde{\phi}_{s,a} \;=\; \phi_{s,a} \;-\; \mathbb{E}_{a' \sim \pi_\theta(\cdot|s)}[\phi_{s,a'}]$$

为中心化的 feature，在 policy 形式固定的情况下，feature 表征了模型的 expressivity。相应的 compatible function approximation error 为

$$L_\nu(w;\theta) = \mathbb{E}_{s,a\sim\nu}\left[\left(A^\theta(s,a) - w \cdot \widetilde{\phi}_{s,a}\right)^2\right].$$

即，描述了 feature 对于 advantage function 的表示能力。

2.5 Bound: NPG for unconstrained policy classes w.r.t. a fixed policy

固定一个策略 π ，以下定理说明做 T 轮 NPG 之后，得到策略性能距离策略 π 的性能的差距。注意，在下面定理中，不可以把这个固定的策略看做是最优策略，因为最优策略是未知的，而这里的算法假定能获得最优策略产生的状态分布。再后面的定理会通过一个 exploratory 的初始分布来绕过该限制。

Theorem 6.4. *(NPG approximation) Fix a comparison policy π and a state distribution ρ. Define ν as the induced state-action measure under π, i.e.*

$$\nu(s,a) = d_\rho^\pi(s)\pi(a|s).$$

Suppose that the update rule (18) starts with $\theta^{(0)} = 0$ and uses the sequence of weights $w^{(0)}, \dots, w^{(T)}$; that Assumption 6.2 holds; and that for all $t < T$,

$$\frac{1}{T}\sum_{t=0}^{T-1} L_\nu(w^{(t)};\theta^{(t)}) \le \widetilde{\epsilon}_{approx}, \quad \|w^{(t)}\| \le W.$$

We have that:

$$\min_{t<T}\left\{V^\pi(\rho) - V^{(t)}(\rho)\right\} \le \frac{1}{1-\gamma}\left(\sqrt{\widetilde{\epsilon}_{approx}} + \frac{\log|\mathcal{A}|}{\eta T} + \frac{\eta\beta W^2}{2}\right)$$

Assumption 6.2. (Policy Smoothness) Assume for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ that $\log \pi_\theta(a|s)$ is a β-smooth function (in θ).

证明的过程比较巧妙，值得一看，这里不贴出来了。关键点如下：1）利用 smoothness 的条件，说明相邻两轮的策略差别不大；2）每一轮策略相对于策略 π 的 KL 散度都在减小，并且减小程度的 lower bound 与 compatible function approximation error 和 $(V^\pi(\rho) - V^{(t)}(\rho))$ 有关；3）计算 $\sum_{t=0}^{T-1}(V^\pi(\rho) - V^{(t)}(\rho))$ 的上界，利用前述结论，并且前后交叠消项。

2.6 Bound: NPG for unconstrained policy classes w.r.t. optimal policy (agnostic result)

前面得到的 bound 需要作为比较的策略预先知道，一般我们需要和策略函数族中最优策略相比（agnostic analysis），而最优策略是不知道的。因此，前面的方法不构成可行的算法，这里研究一个可行的算法，即把 compatible function approximation error 中需要用到的

$$\nu^\star(s,a) = d_\rho^{\pi^\star}(s)\pi^\star(a|s).$$

换成

$$\nu_{\nu_0}^{\pi}(s,a) := (1-\gamma)\mathbb{E}_{s_0,a_0\sim\nu_0} \sum_{t=0}^{\infty} \gamma^t \Pr^\pi(s_t = s, a_t = a | s_0, a_0)$$

其中 π 为最新的策略，也把上述分布记为 $\nu^{(t)}$ 。每一轮参数更新公式为：

$$w^{(t)} \in \mathrm{argmin}_w\, L_{\nu^{(t)}}(w; \theta^{(t)}). \tag{21}$$

有如下结论：

Corollary 6.5. *(Agnostic Learning with NPG) Suppose that we follow the update rule in (21) starting with $\theta^{(0)} = 0$. Fix a state distribution ρ and a state-action distribution ν_0; let $\pi^\star = \pi_{\theta^\star}$ the best policy in Π for ρ, i.e. $\theta^\star \in \mathrm{argmax}_{\theta\in\Theta} V^{\pi_\theta}(\rho)$. Define $\nu^\star$ as the induced state-action measure under $\pi^\star$, i.e.*

$$\nu^\star(s,a) = d_\rho^{\pi^\star}(s)\pi^\star(a|s).$$

Suppose $\eta = \sqrt{2\log|\mathcal{A}|/(\beta W^2 T)}$; Assumption 6.2 holds; and that for all $t < T$,

$$L^\star_{\nu^{(t)}}(\theta^{(t)}) \le \epsilon_{approx}, \quad \|w^{(t)}\| \le W.$$

We have that:

$$\min_{t<T}\left\{V^{\pi^\star}(\rho) - V^{(t)}(\rho)\right\} \le \left(\frac{W\sqrt{2\beta\log|\mathcal{A}|}}{(1-\gamma)}\right)\cdot\frac{1}{\sqrt{T}} + \sqrt{\frac{1}{(1-\gamma)^3}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty \epsilon_{approx}}\,.$$

Proof: Since $\nu^{(t)}(s,a) \ge (1-\gamma)\nu_0(s,a)$,

$$L_{\nu^\star}(w^{(t)};\theta^{(t)}) \le \left\|\frac{\nu^\star}{\nu^{(t)}}\right\|_\infty \cdot L_{\nu^{(t)}}(w^{(t)};\theta^{(t)}) \le \left\|\frac{\nu^\star}{\nu^{(t)}}\right\|_\infty \cdot \epsilon_{\text{approx}} \le \frac{1}{(1-\gamma)}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty \cdot \epsilon_{\text{approx}}.$$

Using this in Theorem 6.4 with the choice of η completes the proof. ■

可以看到，如果我们不能事先得知要比较的策略下的稳态状态分布，就会产生一个代价，即一个 mismatch $\frac{1}{1-\gamma}\|\frac{\nu^\star}{\nu_0}\|_\infty$ 。这也其实反映了探索难的问题。

2.7 Finite sample analysis for NPG with unconstrained policy classes

注意到，前面所有的分析都假设能够计算得到准确的策略梯度或者是 L_ν 的最小值点。但是它们需要通过足够的样本来估算，因此，这里考虑一个实际的基于样本的算法，并且分析得到其相应的 sample complexity 和 computation complexity。

考虑策略参数的更新公式：

$$\theta^{(t+1)} = \theta^{(t)} + \eta \widehat{w}^{(t)}. \tag{22}$$

而 L_ν 的最小值点 $\hat{w}^{(t)}$ 通过在 L_ν 上基于样本做 SGD 得到：

$$w \leftarrow w - \alpha\widehat{\nabla_w L_{\nu^{(t)}}}(w;\theta^{(t)}),$$

where $\widehat{\nabla_w L_{\nu^{(t)}}}(w;\theta^{(t)})$ is an unbiased estimate of the gradient and α is a constant learning rate.

具体算法如下所示：

Algorithm 1 Sample-based Natural Policy Gradient with Function Approximation

Require: Learning rate η, SGD learning rate α, and simulation access to the MDP M under starting state-action distribution ν_0.

1: Initialize $\theta^{(0)} = 0$.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3:   Initialize $w_0 = 0$
4:   **for** $i = 0, 1, \ldots, N-1$ **do**
5:     Sample $s, a \sim \nu^{(t)}$. Sample $a' \sim \nu^{(t)}(a|s)$.
6:     Continue the episode by executing π starting from s, a, using a termination probability of $1-\gamma$. Let $\widehat{Q}(s,a)$ be the cumulative (undiscounted) reward from this episode.
7:     Estimate

$$g_i = \left(w_i \cdot \nabla \log \pi^{(t)}(a|s) - \widehat{Q}(s,a)\right) \nabla \log \pi^{(t)}(a|s) + \widehat{Q}(s,a) \nabla \log \pi^{(t)}(a'|s).$$

8:     Update w:

$$w_{i+1} = w_i - \alpha g_i.$$

9:   **end for**
10:   Set $\widehat{w}^{(t)} = \frac{1}{N}\sum_{i=1}^{N} w_i$.
11:   Update $\theta^{(t+1)} = \theta^{(t)} + \eta \widehat{w}^{(t)}$.
12: **end for**

- 策略参数 θ 更新 T 轮，每一轮中，使用 N 个样本来在 L_ν 上做 SGD 得到相应的策略参数变化量 $\hat{w}^{(t)}$ ，该变化量由 averaged SGD 产生。

- 第 5 行的做法是从 ν_0 出发，按照当前策略进行 rollout，在每个新遇到的状态上，以 $1-\gamma$ 的概率接受这个状态作为 s ，如果该状态被接受，除了把本来采样得到的行动 a 作用于环境之外，再独立地采集另外一个行动样本 a' 。
- 第 6 行中 $Q(s,a)$ 的获取方法如下，从 s,a 出发，每次遇到一个新的状态时，以 $1-\gamma$ 的概率把该状态作为 terminal state，然后计算 undiscounted cumulative reward 并把它作为 $Q(s,a)$ 。容易看到，这样得到的 $Q(s,a)$ 是 infinite horizon discounted cumulative reward 的无偏估计，即 $\mathbb{E}[Q(s,a)] = r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots$ 。
- 第 7 行中的 g_i 为 $\nabla_w L_{\nu^{(t)}}(w,\theta)$ 的无偏估计，最后减去的一项的期望为 V 函数，即

$$\begin{aligned}
\nabla_w L_{\nu^{(t)}}(w;\theta) &= 2\mathbb{E}_{s,a\sim\nu^{(t)}}\left[\left(w \cdot \nabla \log \pi_\theta(a|s) - A^\theta(s,a)\right)\nabla \log \pi_\theta(a|s)\right] \\
&= 2\mathbb{E}_{s,a\sim\nu^{(t)}}\left[\left(w \cdot \nabla \log \pi_\theta(a|s) - Q^\theta(s,a)\right)\nabla \log \pi_\theta(a|s)\right] \\
&\quad + 2\mathbb{E}_{s,a\sim\nu^{(t)}}\left[V^\theta(s)\nabla \log \pi_\theta(a|s)\right].
\end{aligned} \tag{23}$$

- 这里只采样一个另外的 action 用于估计 V 函数，同时也采用 online averaged SGD update。还有很多可以减少 variance 的措施，这里为了便于分析，没有使用。

下面推论给出了上述算法的 sample complexity 和 computational complexity。

Corollary 6.9. *(Sample and Computational Complexity of NPG) Suppose that we follow the sample based NPG update rule specified in Algorithm 1, starting with $\theta^{(0)} = 0$ and using N episodes per update of $\theta^{(t)}$. Fix a state distribution ρ and a state-action distribution ν_0; let $\pi^\star = \pi_{\theta^\star}$ where $\theta^\star \in \mathrm{argmax}_{\theta\in\Theta} V^{\pi_\theta}(\rho)$; and define $\nu^\star(s,a) = d_\rho^{\pi^\star}(s)\pi^\star(a|s)$.*

Define $w^{(t)} := \mathrm{argmin}_w L_{\nu^{(t)}}(w;\theta^{(t)})$. Suppose $\eta = \sqrt{2\log|\mathcal{A}|/(\beta\widehat{W}^2 T)}$ and $\alpha = 1/B$; assumptions 6.2, 6.6, and 6.7 hold. We have that:

$$\mathbb{E}\left[\min_{t<T}\left\{V^{\pi^\star}(\rho) - V^{(t)}(\rho)\right\}\right] \le$$
$$\left(\frac{\widehat{W}\sqrt{2\beta\log|\mathcal{A}|}}{(1-\gamma)}\right)\cdot\frac{1}{\sqrt{T}} + \sqrt{\frac{1}{(1-\gamma)^3}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty\left(\sqrt{\epsilon_{approx}} + \frac{4\sqrt{d}\big(BW + 1/(1-\gamma)\big)}{\sqrt{N}}\right)}.$$

Furthermore, each episode has expected length $2/(1-\gamma)$ so the expected number of total samples is $2NT/(1-\gamma)$; the total number of gradient computations of $\nabla\log\pi(a|s)$ is $2NT$; the total number of scalar multiplies, divides, and additions is $O(dNT + NT/(1-\gamma))$.

Assumption 6.2. (Policy Smoothness) Assume for all $s\in\mathcal{S}$, $a\in\mathcal{A}$ that $\log\pi_\theta(a|s)$ is a β-smooth function (in θ).

Assumption 6.6. (Lipschitz Policy) Assume $\|\nabla\log\pi^{(t)}(a|s)\| \le B$.

Assumption 6.7. (Bounded Error and Weights) Suppose that for all $t<T$ that:

$$\mathbb{E}\left[L^\star_{\nu^{(t)}}(\theta^{(t)})\right] \le \epsilon_{\text{approx}}, \quad \mathbb{E}\left[\|\widehat{w}^{(t)}\|^2\right] \le \widehat{W}^2, \quad \mathbb{E}\left[\|w^{(t)}\|^2\right] \le W^2$$

把它和 Corollary 6.5 比较可以看到，相比于 exact case（假设策略梯度能够精确求到），由于使用样本来估计，这里多了不等式右边的最后一项，当每一轮采样样本足够多时，averaged SGD 的结果就等于精确的 NPG，这时 Corollary 6.9 = Corollary 6.5。这一项正比于 $\frac{1}{\sqrt{N}}$ ，同时和参数的维度相关。证明的过程主要套用了 (Bach and Moulines, 2013) 关于 averaged SGD 的结果（与此相关地，才会出现 Assumption 6.6 和 6.7 这两个看起来略奇怪的假设）。

3. Projected PG for constrained policy classes

3.1 Bellman policy error

回忆到，natural policy gradient 更新中每次更新的方向 $w^\star(\theta)$ 可以通过最小化 compatible function approximation error L_ν 得到，这里 projected policy gradient 每次更新的量通过最小化 Bellman policy error L_{BPE} 得到：

$$w^\star(\theta) = \mathrm{argmin}_{w \in \mathbb{R}^d \,:\, w+\theta \in \Theta} L_{\mathrm{BPE}}(\theta; w). \tag{26}$$

其定义如下

$$\begin{aligned} L_{\mathrm{BPE}}(\theta; w) &= \mathbb{E}_{s\sim d^{\pi_\theta}}\left[\sum_{a\in\mathcal{A}} \left|\pi_\theta^+(a|s) - \pi_\theta(a|s) - w^T \nabla_\theta \pi_\theta(a|s)\right|\right] \\ &= \mathbb{E}_{s\sim d^{\pi_\theta}}\left[\|\pi_\theta^+(\cdot|s) - \pi_\theta(\cdot|s) - w^T \nabla_\theta \pi_\theta(\cdot|s)\|_1\right], \end{aligned} \tag{25}$$

$$\pi_\theta^+(s) = \mathrm{argmax}_{a\in\mathcal{A}} A^{\pi_\theta}(s,a),$$

其中 π_θ^+ 为相对于当前策略对应价值函数下的 1-step greedy policy，Bellman policy error 衡量了从当前策略出发走一个 gradient step（一阶近似）离 π_θ^+ 还有多远。这里的算法就是每步都更新策略参数，使得这一步上的 Bellman policy error 都被最小化。

注意到当 policy class 为 complete 的时候，Bellman policy error 可以被最小化到零，比如对于 direct policy parameterization：

$$L_{\mathrm{BPE}}(\theta; w) = \mathbb{E}_{s\sim d^{\pi_\theta}}\left[\sum_{a\in\mathcal{A}} \left|\pi_\theta^+(a|s) - \theta_{s,a} - w_{s,a}\right|\right].$$

总可以选择 update step 使得 $\theta_{s,a} + w_{s,a} = \pi_\theta^+(a|s)$ ，对应的算法就是 tabular case 情形下的 policy iteration。

3.2 Stationarity

前面提到 stationarity implies optimality，这里再给出 ϵ-stationary 的定义

a policy π_θ parameterized by θ is ϵ-stationary if for all $\theta + \delta \in \Theta$ and $\|\delta\| \le 1$, we have

$$\delta^\top \nabla_\theta V^{\pi_\theta}(\mu) = \mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\left[\sum_{a\in\mathcal{A}} \frac{1}{1-\gamma} \delta^\top \nabla \pi_\theta(s,a) A^{\pi_\theta}(s,a)\right] \le \epsilon. \tag{24}$$

3.2 Agnostic optimality

当 stationarity 和 Bellman policy error 给定时，能够确定相应 optimality。注意到，这里（包括下一个 subsection）分析的是 exact case。

Theorem 6.10. *Given any starting state distribution ρ, suppose we find an ϵ_{opt}-stationary point θ (24) satisfying $L_{BPE}(\theta) \le \epsilon_{approx}$. Let $\pi^\star = \pi_{\theta^\star}$ where $\theta^\star \in \mathrm{argmax}_{\theta\in\Theta} V^{\pi_\theta}(\rho)$. We have the guarantee*

$$V^{\pi^\star}(\rho) - V^{\pi_\theta}(\rho) \le \frac{1}{(1-\gamma)^3}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty \Big(\epsilon_{approx} + (1-\gamma)^2\big(1 + \|w^\star(\theta)\|_2\big)\epsilon_{opt}\Big)$$

3.3 Iteration complexity

加上一些 regularity 的假设之后，能够得到前述算法的 iteration complexity。

Corollary 6.13. *Suppose that Assumption 6.11 holds and for our definitions (25)-(26) of $L_{BPE}(\theta)$ and $w^\star(\theta)$, assume for all $t < T$*

$$L_{BPE}(\theta^{(t)}) \le \epsilon_{approx} \quad \textit{and} \quad \|w^\star(\theta^{(t)})\|_2 \le W^\star.$$

Let

$$\beta = \frac{\beta_2|\mathcal{A}|}{(1-\gamma)^2} + \frac{2\gamma\beta_1^2|\mathcal{A}|^2}{(1-\gamma)^3}.$$

Then, projected gradient ascent (27) with stepsize $\eta = \frac{1}{\beta}$ satisfies for all starting state distributions ρ and for any other policy $\pi^\star \in \Pi$,

$$\min_{t<T}\left\{V^\star(\rho) - V^{(t)}(\rho)\right\} \le \frac{1}{(1-\gamma)^3}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty \epsilon_{approx} + (W^\star + 1)\epsilon, \ \textit{for}\ T \ge \frac{8\beta}{(1-\gamma)^3\epsilon^2}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty^2$$

Assumption 6.11 (Lipschitz continuous and smooth policies)**.** Assume that for all $\theta, \theta' \in \Theta$ and for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have

$$|\pi_\theta(a|s) - \pi_{\theta'}(a|s)| \le \beta_1 \|\theta - \theta'\|_2 \qquad (\beta_1\text{-Lipschitz})$$
$$\|\nabla_\theta \pi_\theta(a|s) - \nabla_\theta \pi_{\theta'}(a|s)\|_2 \le \beta_2 \|\theta - \theta'\|_2 \qquad (\beta_2\text{-smooth})$$

证明的过程主要需要用到 (Beck, 2017, Theorem 10.15) 关于 projected gradient descent 的结果。

4. 总结

先总结一下本文推导出来的结论和之前工作的对比：

Algorithm	Measure of approximation error	Iteration complexity	Accuracy
Approx. Value/Policy Iteration [Bertsekas and Tsitsiklis, 1996]	ϵ_∞: the ℓ_∞ worst-case error of values	$\frac{1}{1-\gamma} \ln \frac{1}{\epsilon_{\text{opt}}}$	$\epsilon_{\text{opt}} + \frac{2\epsilon_\infty}{(1-\gamma)^2}$
Approx. Policy Iteration, with concentrability [Munos, 2005, Antos et al., 2008]	ϵ_1: an ℓ_1 average-case approx. notion	$\frac{1}{1-\gamma} \ln \frac{1}{\epsilon_{\text{opt}}}$	$\epsilon_{\text{opt}} + \frac{2C_{\rho,\mu}\epsilon_1}{(1-\gamma)^2}$
Conservative Policy Iteration [Kakade and Langford, 2002]	ϵ_1: an ℓ_1 average-case approx. notion	$O\left(\frac{1}{\epsilon_1^2}\right)$	$\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty \frac{\epsilon_1}{(1-\gamma)^2}$
Natural Policy Gradient (Cor 6.5)	ϵ_2: an ℓ_2 average-case approx. notion	$O\left(\frac{1}{(1-\gamma)^2\epsilon_{\text{opt}}^2}\right)$	$\epsilon_{\text{opt}} + \sqrt{\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty \frac{\epsilon_2}{(1-\gamma)^3}}$
Projected Gradient Ascent (Cor 6.13)	ϵ_1: an ℓ_1 average-case approx. notion	$O\left(\frac{1}{(1-\gamma)^6\epsilon_{\text{opt}}^2}\right)$	$\epsilon_{\text{opt}} + \left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty \frac{\epsilon_1}{(1-\gamma)^3}$

这篇文章传递的主要思想有两点：1）尽管 MDP 上的优化是 non-convex 的，但是由于有 stationarity -> optimality 的性质，因此，还是能得到相应的 global optimality 的结论；2）MDP 优化中比较困难的点在于 insufficient exploration，关于这一点可以参考各种结论中出现的 distribution mismatch coefficient 项。

发布于 2019-08-22