Lecture 7: Policy gradient convergence

Akshay Krishnamurthy
akshay@cs.umass.edu

March 7, 2022

1 Recap and the performance difference lemma

We continue to work in the discounted MDP defined by $(\mathcal{S}, \mathcal{A}, R, P, \mu, \gamma)$. Recall that we are interested in finding a policy that maximizes the $J(\pi) := \mathbb{E}[\sum_h \gamma^h r_h \mid \pi]$ and last time we introduced the policy gradient family of algorithms. In this approach, we set up a continuous parametrization for policies, denoted π_θ, and then we use first-order methods to optimize $\max_\theta J(\pi_\theta)$. The most basic version, the standard policy gradient method applies the iteration

$$\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta \nabla_\theta J(\pi_{\theta^{(t)}}).$$

The main result in the last lecture provided simple/easy-to-use expressions for this gradient. Today we will primarily work with the advantage version:

$$\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma} \mathbb{E}_{\substack{s \sim d^{\pi_\theta} \\ a \sim \pi_\theta(s)}} \left[A^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a \mid s) \right]. \tag{1}$$

To recap the definitions, d^π is the discounted state occupancy measure given by $d^\pi(s) = (1-\gamma)\sum_h \gamma^h \Pr[s_h = s \mid \pi]$ and A^π is the advantage function given by $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$. Sometimes we also use d^π to denote the state-action occupany measure. Throughout the lecture we use the superscript (t) to denote an object derived from the t^{th} iterate of the (natural) policy gradient update, e.g., $\pi^{(t)} = \pi_{\theta^{(t)}}$, etc.

At the end of the last lecture we also saw the natural policy gradient algorithm. This method uses a Fisher information matrix as a preconditioner to the gradient update, to induce a better geometry. This leads to a particularly clean form of update when we use a tabular softmax parametrization. Recall that here we consider the parametrized policy class:

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})} \text{ for } \theta \in \mathbb{R}^{S \times A}.$$

With this parametrization, the NPG update can be expressed in the policy space directly, where it is given by:

$$\theta^{(t+1)} \leftarrow \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)} + v, \qquad \pi^{(t+1)}(a \mid s) = \pi^{(t)}(a \mid s) \cdot \frac{\exp(\eta A^{(t)}(s,a)/(1-\gamma))}{Z_t(s)},$$

where v is a state-dependent offset, $\pi^{(t)}$ is the previous iterate, $A^{(t)} = A^{\pi^{(t)}}$ is the advantage function for that policy, and Z_t is a normalizing factor that ensures that $\pi^{(t+1)}(\cdot \mid s)$ is a distribution.

Today we will discuss convergence properties of these methods, their sample-based analogues, and the function approximation setting. The key lemma to analyze these methods is the so-called "Performance Difference Lemma," which is also used throughout the theory of reinforcement learning.

Lemma 1 (Performance Difference Lemma)**.** *Let π_1, π_2 be any two policies. Then*

$$J(\pi_1) - J(\pi_2) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d^{\pi_1}} \left[A^{\pi_2}(s,a) \right].$$

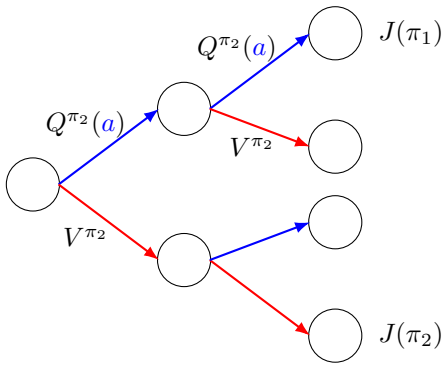

Figure 1: An illustration of the performance difference lemma. Policy π_1 takes the top path (and blue actions) while π_2 takes the bottom path (and red actions). The difference in rewards can be decomposed in terms of 1-step differences of π_2's Q-function along the path taken by π_1.

Observe that this lemma decomposes the multi-step difference in reward into many "one-step" differences, and in some sense address credit assignment. A picture of the intuition for this lemma is in Figure 1.

Proof. The proof is based on an unrolling argument. Observe that

$$\begin{aligned}
J(\pi_1) - J(\pi_2) &= \mathbb{E}_{s\sim\mu}\left[V^{\pi_1}(s) - V^{\pi_2}(s)\right] \\
&= \mathbb{E}_{s\sim\mu}[\mathbb{E}_{a\sim\pi_1(\cdot|s)}Q^{\pi_1}(s,a) - V_2^{\pi}(s)] \\
&= \mathbb{E}_{s\sim\mu}\left[\mathbb{E}_{a\sim\pi_1(\cdot|s)}\left[Q^{\pi_1}(s,a) - Q^{\pi_2}(s,a)\right]\right] + \mathbb{E}_{s\sim\mu}\left[\mathbb{E}_{a\sim\pi_1(\cdot|s)}Q^{\pi_2(s,a)} - V_2^{\pi}(s)\right] \\
&= \mathbb{E}_{s\sim\mu}\left[\mathbb{E}_{a\sim\pi_1(\cdot|s)}\left[Q^{\pi_1}(s,a) - Q^{\pi_2}(s,a)\right]\right] + \mathbb{E}_{(s,a)\sim\mu\circ\pi_1}\left[A^{\pi_2}(s,a)\right]
\end{aligned}$$

Now the first term can be expressed as the difference in value from the second state visited by π_1. Let us call this distribution $d_1^{\pi_1}$. Then:

$$\mathbb{E}_{s\sim\mu}\left[\mathbb{E}_{a\sim\pi_1(\cdot|s)}\left[Q^{\pi_1}(s,a) - Q^{\pi_2}(s,a)\right]\right] = \gamma\mathbb{E}_{s\sim d_1^{\pi_1}}\left[V^{\pi_1}(s) - V^{\pi_2}(s)\right]$$

Repeating the argument above indefinitely yields

$$J(\pi_1) - J(\pi_2) = \sum_{h=0}^{\infty}\gamma^h\mathbb{E}_{s,a\sim d_h^{\pi_1}\circ\pi_1}[A^{\pi_2}(s,a)] = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d^{\pi_1}}[A^{\pi_2}(s,a)]. \qquad \square$$

2 Convergence of vanilla PG

The vanilla policy gradient method does converge, but establishing this is non-trivial and requires overcoming several barriers. Most notably, the optimization problem is not convex. More subtle issues include: (1) coverage of the state space and (2) the fact that with the softmax policy class, deterministic policies have parameters tending to infinity. For the former, suppose we have an MDP with just a few rewarding states and we have some policy that does not visit these states at all. Then the gradient for this policy will be zero, meaning gradient ascent will not move away from this policy, even though it could be highly suboptimal. For the latter, any sequence of policies that become deterministic have parameters tending to infinity, and this implies that the gradient tends to zero. (This can be seen by looking at the A-version of the policy gradient, using the fact that $\nabla_\theta \log \pi_\theta(a \mid s) \approx e_{s,a}$ if the policy is deterministic and the fact that the average-over-actions advantage is 0.)

Despite this it is possible to establish an asymptotic convergence result for vanilla policy gradient with the tabular-softmax parametrization. While we will not prove this here, the theorem is stated below.

Theorem 2. *If μ is strictly positive, that is $\mu(s) > 0$ for all states s and $\eta \le (1-\gamma)^3/8$, then for all states s we have $V^{(t)}(s) \to V^\star(s)$ as $t \to \infty$.*

The tabular softmax representation overcomes the first barrier described above, since unless the parameters are at infinity, the policy has some probability of playing every action and hence some probability of visiting every state. However, this probability may be exponentially small. In line with this, note that no convergence rate is provided here and the conjecture is that convergence may be exponentially slow. This is somewhat undesireable and it can be remedied with various techniques like regularization, although we will not study these methods here.

Instead, NPG overcomes this "low visitation" issue by reweighting according to the Fisher information matrix. Recall that we are preconditioning by the pseudoinverse of $\mathbb{E}_{(s,a)\sim d^\pi}\left[(\nabla_\theta \log \pi_\theta(a \mid s))(\nabla_\theta \log \pi_\theta(a \mid s))^\top\right]$. If d^π places very low mass on some state, preconditioning with this matrix will dramatically increase the magnitude of the update, even though the gradient itself may be very small. In this way, NPG intuitively addresses some of the issues with vanilla PG. On a technical level, it is somewhat more straightforward to understand the behavior of the natural policy gradient, since it is so closely related to Exponential Weights. We will study this algorithm in more detail in this lecture.

3 Convergence of NPG

For NPG, the key theorem is a "regret bound" that is quite related to the analysis of Exponential weights. This may not be surprising by examining the form of the update.

Theorem 3. *Suppose we run NPG for T rounds starting with $\theta^{(1)} = 0$ and with learning rate $\eta > 0$. Then*

$$J(\pi^\star) - J(\pi^{(T+1)}) \leq \frac{\log|\mathcal{A}|}{\eta T} + \frac{1}{(1-\gamma)^2 T}.$$

Proof of Theorem 3. Let us start by comparing $J(\pi^\star)$ with the value of the t^{th} iterate of NPG. Recall that the KL divergence between two distributions $p, q \in \Delta(\mathcal{A})$ is given by $\mathrm{KL}(p||q) := \sum_a p(a)\log(p(a)/q(a))$ and that this quantity is bounded between 0 and $\log|\mathcal{A}|$ (or more generally the support size of the distributions).

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

Here we use the PDL, then the closed form of our updates and finally the definition of the KL divergence. Next let us look at the "average regret":

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T J(\pi^\star) - J(\pi^{(t)}) &= \frac{1}{\eta T}\sum_{t=1}^T \mathbb{E}_{s\sim d^{\pi^\star}} \Big(\mathrm{KL}(\pi^\star(\cdot \mid s)||\pi^{(t)}(\cdot \mid s)) - \mathrm{KL}(\pi^\star(\cdot \mid s)||\pi^{(t+1)}(\cdot \mid s)) + \log(Z_t(s))\Big) \\
&\leq \frac{\mathbb{E}_{s\sim d^{\pi^\star}}\left[\mathrm{KL}(\pi^\star(\cdot \mid s)||\pi^{(1)}(\cdot \mid s))\right]}{\eta T} + \frac{1}{\eta T}\sum_{t=1}^T \mathbb{E}_{s\sim d^{\pi^\star}} \log(Z_t(s)) \\
&\leq \frac{\log|\mathcal{A}|}{\eta T} + \frac{1}{\eta T}\sum_{t=1}^T \mathbb{E}_{s\sim d^{\pi^\star}} \log(Z_t(s))
\end{aligned}$$

When looking at the average regret, we conveniently obtain a telescoping sum, which leads to a much simpler bound. We also are using that $\pi^{(0)}$ is the uniform distribution. To finish the proof, we need to do two things: (1) relate the "average regret" to the performance of the last iterate and (2) upper bound the normalizing constant terms. The key inequality to establishing both of these steps is that for any initial distribution ρ

$$0 \leq \frac{1-\gamma}{\eta}\mathbb{E}_{s\sim\rho} \log Z_t(s) \leq J(\pi^{(t+1)};\rho) - J(\pi^{(t)};\rho) \tag{2}$$

(Here $J(\pi;\rho)$ means we change the starting distribution of the MDP to ρ but otherwise compute the value as before.) Observe that this establishes a form of policy improvement for the NPG iteration, very similar to what we say in the policy iteration analysis. Indeed, this inequality establishes that $V^{\pi^{(t+1)}} \succeq V^{\pi^{(t)}}$ in a pointwise sense.

Before we prove (2) let us see how we can use it to establish the theorem. First, taking $\rho = \mu$ this implies that $J(\pi^{(t+1)}) \geq J(\pi^{(t)})$, or in other words $J(\pi^{(T+1)})$ is the largest value among all of our iterates. Formally

$$J(\pi^\star) - J(\pi^{(T+1)}) \leq \frac{1}{T}\sum_{t=1}^{T} J(\pi^\star) - J(\pi^{(t)}),$$

which addresses our first issue. For the second issue, taking $\rho = d^{\pi^\star}$ we get

$$\frac{1}{\eta T}\sum_{t=1}^{T} \mathbb{E}_{s\sim d^{\pi^\star}} \log Z_t(s) \leq \frac{1}{(1-\gamma)T}\sum_{t=1}^{T} J(\pi^{(t+1)}; d^{\pi^\star}) - J(\pi^{(t)}; d^{\pi^\star}) = \frac{J(\pi^{(T+1)}; d^{\pi^\star}) - J(\pi^{(1)}; d^{\pi^\star})}{(1-\gamma)T} \leq \frac{1}{(1-\gamma)^2 T}$$

Finally, let us prove (2). The first inequality is a consequence of Jensen's inequality. For any state s

$$\log Z_t(s) = \log\left(\sum_a \pi^{(t)}(a \mid s)\exp(\eta A^{(t)}(s,a)/(1-\gamma))\right) \geq \frac{\eta}{1-\gamma}\sum_a \pi^{(t)}(a \mid s)A^{(t)}(s,a) = 0$$

The second inequality follows from PDL and the form of the updates. For this derivation, let d^π_ρ be the discounted occupancy measure when we start in distribution ρ and execute policy π.

$$\begin{aligned}
J(\pi^{(t+1)};\rho) - J(\pi^{(t)};\rho) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi^{(t+1)}}_\rho}\sum_a \pi^{(t+1)}(a \mid s)A^{(t)}(s,a) \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^{\pi^{(t+1)}}_\rho}\sum_a \pi^{(t+1)}(a \mid s)\log\left(\frac{\pi^{(t+1)}(a \mid s)Z_t(s)}{\pi^{(t)}(a \mid s)}\right) \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^{\pi^{(t+1)}}_\rho}\mathrm{KL}(\pi^{(t+1)}(\cdot \mid s)||\pi^{(t)}(\cdot \mid s)) + \frac{1}{\eta}\mathbb{E}_{s\sim d^{\pi^{(t+1)}}_\rho}\log Z_t(s) \\
&\geq \frac{1}{\eta}\mathbb{E}_{s\sim d^{\pi^{(t+1)}}_\rho}\log Z_t(s) \geq \frac{(1-\gamma)}{\eta}\mathbb{E}_{s\sim\rho}\log Z_t(s). \qquad \square
\end{aligned}$$

4 NPG convergence with sampling errors

There is also a fairly clean way to analyze NPG with sampling errors. Recall that in the parameter space, the update is $\theta^{(t+1)} \leftarrow \theta^{(t)} + \frac{\eta}{1-\gamma}A^{(t)}$ up to a state dependent offset. To capture estimation errors, we'll (a) absorb the $1-\gamma$ term into the learning rate and (b) replace $A^{(t)}$ with some estimator $w^{(t)}$ that we will fit from samples. At a high level, we'll want to train $w^{(t)}$ so that

$$A^{(t)}(s,a) \approx \left\langle w^{(t)}, \nabla_\theta \log \pi^{(t)}(a \mid s)\right\rangle.$$

The intuition for this arises from the tabular softmax parametrization where $\nabla_\theta \log \pi^{(t)}(a \mid s) = e_{s,a} - \sum_{a'} e_{s,a'}\pi^{(t)}(a' \mid s)$. With this expression for the gradient, we are asking that $w^{(t)}(s,a) = A^{(t)}(s,a)$ up to a state-dependent offset, which matches the form of the NPG update. Indeed, if we set $w^{(t)}(s,a) = Q^{(t)}(s,a)$ then the above will be satisfied with equality. However, while this intuition stems from the tabular softmax representation, it also works more generally and indeed can capture the function approximation setting where we will treat the gradient as features and train a linear function to predict the advantage.

To capture both estimation and approximation errors, we will keep track of how well $w^{(t)}$ approximates $A^{(t)}$. To do this, let $\tilde{\pi}$ be some reference policy that we will compare to in the analysis. This could be the optimal policy, or something else. Then we define

$$\mathrm{err}_t := \mathbb{E}_{(s,a)\sim d^{\tilde{\pi}}}\left[A^{(t)}(s,a) - \left\langle w^{(t)}, \nabla_\theta \log \pi^{(t)}(a \mid s)\right\rangle\right].$$

The main lemma here is the following "regret lemma"

Lemma 4. *Fix comparison policy $\tilde{\pi}$ and assume that $\log \pi_\theta(a \mid s)$ is β-smooth with respect to the ℓ_2 norm:*

$$\forall \theta, \theta', s, a : |\log \pi_{\theta'}(a \mid s) - \log \pi_\theta(a \mid s) - \nabla \log \pi_\theta(a \mid s) \cdot (\theta' - \theta)| \leq \frac{\beta}{2}\|\theta - \theta'\|_2. \tag{3}$$

Assume that $\sup_t \|w^{(t)}\|_2 \leq W$ and that err_t is defined as above. Then the NPG iterates given by $\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta w^{(t)}$ starting with $\theta^{(1)} = 0$ satisfy

$$\min_{t \leq T} \left\{ J(\tilde{\pi}) - J(\pi^{(t)}) \right\} \leq \frac{1}{1-\gamma} \left(\underbrace{\frac{\log |\mathcal{A}|}{\eta T} + \frac{\eta \beta W^2}{2}}_{\textit{Exponential Weights regret}} + \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t \right)$$

Remark 5. *Note that the tabular softmax representation is indeed 1 smooth. To see why this is true, by Taylor's theorem with remainder, it suffices to check that for any $\zeta \in [0, 1]$*

$$\frac{1}{2}(\theta' - \theta) \nabla^2 \log \pi_{\zeta\theta + (1-\zeta)\theta'}(a \mid s)(\theta' - \theta) \leq \frac{1}{2}\|\theta' - \theta\|_2^2.$$

A sufficient condition for this is that for any *parametrization θ the operator norm of the hessian is at most* 1*. Since the gradient is $\nabla_\theta \log \pi_\theta(a \mid s) = e_{s,a} - \sum_{a'} \pi_\theta(a' \mid s) e_{s,a'}$, one can check that the hessian is an $SA \times SA$ matrix that is all zeros except for the $A \times A$ block corresponding to state s, where it is:*

$$\nabla_\theta^2 \log \pi_\theta(a \mid s) = \pi_\theta(\cdot \mid s)\pi_\theta(\cdot \mid s)^\top - \mathrm{diag}(\pi_\theta(\cdot \mid s)) \in \mathbb{R}^{A \times A}.$$

Considering any vector $v \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ with $\|v\|_2 \leq 1$ it can now be easily seen that

$$\left| v^\top \left(\nabla_\theta^2 \log \pi_\theta(a \mid s) \right) v \right| = \left| (\mathbb{E}_{a \sim \pi_\theta(\cdot \mid s)} v_{s,a})^2 - \mathbb{E}_{a \sim \pi_\theta(\cdot \mid s)} v_{s,a}^2 \right| \leq 1.$$

Proof. Let us try to obtain a telescoping sum, as we did in the exponential weights analysis. We start by relating the reference policy to the t^{th} iterate:

$$\begin{aligned}
J(\tilde{\pi}) - J(\pi^{(t)}) &= \frac{1}{1-\gamma} \mathbb{E}_{s,a \sim d^{\tilde{\pi}}} \left[A^{(t)}(s, a) \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s,a \sim d^{\tilde{\pi}}} \left[\left\langle w^{(t)}, \nabla_\theta \log \pi^{(t)}(a \mid s) \right\rangle \right] + \frac{\mathrm{err}_t}{1-\gamma} \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s,a \sim d^{\tilde{\pi}}} \left[\frac{1}{\eta} \left\langle \theta^{(t+1)} - \theta^{(t)}, \nabla_\theta \log \pi^{(t)}(a \mid s) \right\rangle \right] + \frac{\mathrm{err}_t}{1-\gamma} \\
&\leq \frac{1}{1-\gamma} \mathbb{E}_{s,a \sim d^{\tilde{\pi}}} \left[\frac{1}{\eta} \log \left(\frac{\pi^{(t+1)}(a \mid s)}{\pi^{(t)}(a \mid s)} \right) + \frac{\eta\beta}{2} \|w^{(t)}\|_2^2 \right] + \frac{\mathrm{err}_t}{1-\gamma} \\
&\leq \frac{1}{1-\gamma} \left(\frac{1}{\eta} \mathbb{E}_{s,a \sim d^{\tilde{\pi}}} \left[\log \left(\frac{\pi^{(t+1)}(a \mid s)}{\pi^{(t)}(a \mid s)} \right) \right] + \frac{\eta\beta W^2}{2} + \mathrm{err}_t \right)
\end{aligned}$$

The first equality is PDL, the second uses the definition of err_t and the third uses the form of the NPG update, namely that $w^{(t)} = (\theta^{(t+1)} - \theta^{(t)})/\eta$. The first inequality uses the smoothness property and the second inequality upper bounds $\|w^{(t)}\|_2^2$. We finish the proof by observing that the best of the T iterates has value that is better than the average, and the average results in a telescoping sum:

$$\begin{aligned}
\min_{t \leq T} \left\{ J(\tilde{\pi}) - J(\pi^{(t)}) \right\} &\leq \frac{1}{T} \sum_{t=1}^{T} J(\tilde{\pi}) - J(\pi^{(t)}) \\
&\leq \frac{1}{1-\gamma} \left(\frac{1}{\eta \cdot T} \sum_{t=1}^{T} \mathbb{E}_{s,a \sim d^{\tilde{\pi}}} \left[\log \left(\frac{\pi^{(t+1)}(a \mid s)}{\pi^{(t)}(a \mid s)} \right) \right] + \frac{\eta\beta W^2}{2} + \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t \right) \\
&= \frac{1}{1-\gamma} \left(\frac{1}{\eta \cdot T} \mathbb{E}_{s,a \sim d^{\tilde{\pi}}} \left[\log \left(\frac{\pi^{(T+1)}(a \mid s)}{\pi^{(1)}(a \mid s)} \right) \right] + \frac{\eta\beta W^2}{2} + \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t \right) \\
&\leq \frac{1}{1-\gamma} \left(\frac{\log |\mathcal{A}|}{\eta \cdot T} + \frac{\eta\beta W^2}{2} + \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t \right).
\end{aligned}$$

□

Regression and bounds on err_t. The last step to completely instantiate the algorithm is to fit $w^{(t)}$ from samples and control the err_t terms. Recall that we want $w^{(t)}(s,a) \approx Q^{(t)}(s,a)$ and we already saw how to obtain unbiased estimates of the latter using the geometric stopping trick. So, on the t^{th} iteration of NPG, let us collect a dataset of the form $\{(s_i, a_i, \widehat{Q}^{(t)})_i\}_{i=1}^N$ where $s_i \sim \mu$, $a_i \sim \mathrm{Unif}(\mathcal{A})$ and $\widehat{Q}_i^{(t)}$ is obtained by taking actions according to $\pi^{(t)}$, stopping the trajectory at each step with probability $1-\gamma$ and returning the undiscounted sum of rewards. From this dataset, we can solve the regression problem

$$w^{(t)} \leftarrow \operatorname*{argmin}_w \frac{1}{N} \sum_{i=1}^{N} (w(s_i,a_i) - \widehat{Q}_i^{(t)})^2, \tag{4}$$

and from a standard regression analysis, we will get

$$\mathbb{E}_{s,a \sim \mu \circ \mathrm{unif}(\mathcal{A})} \left[(w^{(t)}(s,a) - Q^{(t)}(s,a))^2 \right] \le O\left(\frac{SA \log(1/\delta)}{N(1-\gamma)} \right),$$

with probability at least $1-\delta$. Here $SA/(1-\gamma)$ arises because SA is the dimensionality of the regression problem and $1/(1-\gamma)$ is the range of the Q-function, reflecting the scale of the problem. (There is one detail here in the regression analysis since the reward estimates are technically unbounded. There are two ways to handle this, one is by truncating the roll out at some point (like a large constant time $1/(1-\gamma)$ time steps) and showing that this incurs very little bias, the other is via a conditioning argument showing that with high probability all of the trajectories you collect are relatively short.)

From here we have to check two things. The first is that we should use this guarantee to bound err_t and the second is that we need to make sure that $\|w^{(t)}\|_2 \le W$. The second is easier in the tabular parametrization, since we know that $Q^{(t)}(s,a) \in [0, \frac{1}{1-\gamma}]$ we can constrain w in the regression problem to have $\|w\|_2 \le \sqrt{SA}/(1-\gamma)$ without losing the optimal predictor.

For the former, we have to handle the issue of *distribution shift*. For this we use a reweighting argument:

$$\begin{aligned}
\mathrm{err}_t &:= \mathbb{E}_{(s,a)\sim d^{\tilde{\pi}}} \left[A^{(t)}(s,a) - \left\langle w^{(t)}, \nabla_\theta \log \pi^{(t)}(a \mid s) \right\rangle \right] \\
&= \mathbb{E}_{(s,a)\sim d^{\tilde{\pi}}} \left[Q^{(t)}(s,a) - \mathbb{E}_{a\sim\pi^{(t)}(\cdot|s)} Q^{(t)}(s,a) - w^{(t)}(s,a) + \mathbb{E}_{a\sim\pi^{(t)}(\cdot|s)} w^{(t)}(s,a) \right] \\
&\le \mathbb{E}_{(s,a)\sim d^{\tilde{\pi}}} \left[|Q^{(t)}(s,a) - w^{(t)}(s,a)| \right] + \mathbb{E}_{\substack{s\sim d^{\tilde{\pi}} \\ a\sim\pi^{(t)}(\cdot|s)}} \left[\left| Q^{(t)}(s,a) - w^{(t)}(s,a) \right| \right].
\end{aligned}$$

Both of these terms are addressed using the same argument:

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^{\tilde{\pi}}} \left[|Q^{(t)}(s,a) - w^{(t)}(s,a)| \right] &\le \sqrt{\mathbb{E}_{(s,a)\sim d^{\tilde{\pi}}} \left[(Q^{(t)}(s,a) - w^{(t)}(s,a))^2 \right]} \\
&= \sqrt{\sum_{s,a} \frac{d^{\tilde{\pi}}(s,a)}{\mu(s)/A} \cdot \frac{\mu(s)}{A} \cdot \left(Q^{(t)}(s,a) - w^{(t)}(s,a) \right)^2} \\
&\le \sqrt{|\mathcal{A}| \cdot \left\| \frac{d^{\tilde{\pi}}}{\mu} \right\|_\infty \mathbb{E}_{(s,a)\sim\mu\circ\mathrm{Unif}(\mathcal{A})} \left[\left(Q^{(t)}(s,a) - w^{(t)}(s,a) \right)^2 \right]} \\
&\le O\left(\sqrt{|\mathcal{A}| \cdot \left\| \frac{d^{\tilde{\pi}}}{\mu} \right\|_\infty \frac{|\mathcal{S}||\mathcal{A}| \log(1/\delta)}{N(1-\gamma)}} \right).
\end{aligned}$$

Here the key quantity is the density ratio term: $\left\| \frac{d^{\tilde{\pi}}}{\mu} \right\|_\infty$, which captures how poorly our starting state distribution covers the distribution of states visited by the reference policy $\tilde{\pi}$. Indeed this is the main term demonstrating how policy gradient methods do not address the exploration challenge. They rely on the initial distribution (or some known distribution) to cover the state space sufficiently.

Algorithm overview and analysis steps. This essentially completes the analysis for NPG with sampling with tabular-softmax representation. Just to recap, the algorithm operates as follows. Starting with $\theta^{(1)} = 0$, at each iteration $t \in [T]$ we:

1. Collect N trajectories starting from $\mu \circ \mathrm{Unif}(\mathcal{A})$ and rolling out with $\pi^{(t)}$, using geometric stopping.
2. Fit $w^{(t)}$ by solving the regression problem in (4).
3. Perform the update $\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta w^{(t)}$.

The analysis relies heavily on Lemma 4, which in term relies heavily on the performance difference lemma, but to use it we must verify a few things:

- Our policy parametrization is smooth in the sense of (3).
- Our regression problem yields useful bounds on err_t, which will result in some notion of distribution shift.

Beyond this, the analysis is quite general and many pieces can be swapped. We can change the distribution from which we collect data, how we solve the regression problem, etc.

Softmax linear parametrization. Perhaps most interestingly, the analysis can also work with non-tabular parametrizations, like the softmax-linear one. Here the regret lemma holds as is, and we just need to understand how to control err_t. For the softmax-linear parametrization, note that

$$\nabla \log \pi_\theta(a \mid s) = \phi(s,a) - \mathbb{E}_{a' \sim \pi_\theta(\cdot|s)}[\phi(s,a')],$$

So we can solve a regression problem where we train $w^{(t)}$ so that $\left\langle w^{(t)}, \phi(s,a) \right\rangle \approx Q^{(t)}(s,a)$, via linear regression. Then if our features are expressive enough so that $Q^{(t)}$ is linearly realizable, we can obtain a square loss bound of the form

$$\mathbb{E}_{(s,a)\sim\mu\circ\mathrm{Unif}(\mathcal{A})}\left[\left(\left\langle w^{(t)}, \phi(s,a)\right\rangle - Q^{(t)}(s,a)\right)^2\right] \lesssim \frac{d}{N(1-\gamma)}.$$

If we assume that $Q^{(t)}(s,a) = \left\langle w_\star^{(t)}, \phi(s,a)\right\rangle$ then we can use a "linear" version of a distribution shift argument in terms of covariance matrices. Indeed let $\Sigma_{\tilde{\pi}} = \mathbb{E}_{(s,a)\sim d^{\tilde{\pi}}}[\phi(s,a)\phi(s,a)^\top]$ then we can transfer the error as

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^{\tilde{\pi}}}\left[\left\langle w^{(t)} - w_\star^{(t)}, \phi(s,a)\right\rangle^2\right] &= \mathbb{E}_{s,a\sim d^{\tilde{\pi}}}\left[\left\langle \Sigma_\mu^{+1/2}(w^{(t)} - w_\star^{(t)}), \Sigma_\mu^{-1/2}\phi(s,a)\right\rangle^2\right] \\
&\le \|w^{(t)} - w_\star^{(t)}\|_{\Sigma_\mu}^2 \cdot \mathbb{E}_{s,a\sim d^{\tilde{\pi}}}\|\phi(s,a)\|_{\Sigma_\mu^{-1}}^2 \\
&= \mathbb{E}_{s,a\sim\mu\circ\mathrm{Unif}(\mathcal{A})}\left[\left\langle w^{(t)} - w_\star^{(t)}, \phi(s,a)\right\rangle^2\right] \cdot \mathbb{E}_{s,a\sim d^{\tilde{\pi}}}\|\phi(s,a)\|_{\Sigma_\mu^{-1}}^2
\end{aligned}$$

The "relative condition number" $\mathbb{E}_{d^{\tilde{\pi}}}\|\phi\|_{\Sigma_\mu^{-1}}$ is the linear version of the distribution shift coefficient, which you may also see written in many other ways, e.g., $\mathrm{tr}(\Sigma_{\tilde{\pi}}\Sigma_\mu^{-1})$. The key point is that this may not depend on the size of the state space, and this is also true for the regression error. Since both the regression error and the distribution shift term may not explicitly depend on $|\mathcal{S}|$ we do see some ability to generalize across states. However, as before we still require that the initial distribution (or the data collection distribution more generally) covers all directions. Based on this, we can get an end-to-end sample complexity guarantee with no dependence on the cardinality of the state space using the softmax linear representation under the assumption that Q^π is linear for all π.