COLT 2021 RL Theory Tutorial: Exercises

Akshay Krishnamurthy and Wen Sun

August 4, 2021

Exercises for Natural Policy Gradient

In this exercise, we consider the discounted Markov Decision Process $(\mathcal{S}, \mathcal{A}, r, P, \gamma)$ where the initial distribution and exploratory distribution coincide. We refer to both as $\rho \in \Delta(\mathcal{S})$. Recall that for a policy π we use $d_\rho^\pi \in \Delta(\mathcal{S})$ to denote the discounted state visitation distribution for π starting from ρ:

$$d_\rho^\pi(s) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s \mid s_0 \sim \rho, \pi). \tag{1}$$

We also sometimes overload this notation to denote a distribution over states and actions, where the action is always sampled from π.

We focus on the Natural Policy Gradient (NPG) algorithm with tabular softmax parametrization, that is

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}, \tag{2}$$

where $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ are the parameters. Recall that the NPG update is given by

$$\theta^{(t+1)} = \theta^{(t)} + \eta F_\rho(\theta^{(t)})^\dagger \nabla_\theta V^{(t)}(\rho), \tag{3}$$

$$F_\rho(\theta) = \mathbb{E}_{s \sim d_\rho^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot \mid s)} \left[(\nabla_\theta \pi_\theta(a \mid x))(\nabla_\theta \pi_\theta(a \mid x))^\top \right], \tag{4}$$

and $V^{(t)}(\rho)$ is the value of policy $\pi_{\theta^{(t)}}$ from initial distribution ρ. Throughout we use $\pi^{(t)} = \pi^{(\theta^{(t)})}$, $A^{(t)} = A^{(\pi_{\theta^{(t)}})}$ to simplify the notation.

1 Closed form NPG update

Q1: Prove the following proposition verifying a closed form for the NPG update.

Proposition 1. *For NPG with the softmax parametrization in* (2) *we have that*

$$\pi^{(t+1)}(a \mid s) \propto \pi^{(t)}(a \mid s) \cdot \frac{\exp(\eta A^{(t)}(s,a)/(1-\gamma))}{Z_t(s)}, \tag{5}$$

where $Z_t(s)$ is a normalizing factor that ensures that $\pi^{(t+1)}(\cdot \mid s)$ is a distribution.

It may be helpful to view $A^{(t)}(\cdot,\cdot)$ as a vector in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and instead show that

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}(\cdot,\cdot) + \eta v \tag{6}$$

where $v_{s,a} = v_{s,a'} \forall s, a, a'$ is a state-dependent but action-independent offset. Observe that the result follows immediately from (6). Also note that $A^{(t)}(s,a) = Q^{(t)}(s,a) - V^{(t)}(s)$, where $V^{(t)}$ is state-dependent only, so we can also write the algorithm using the Q functions.

2 Performance difference lemma

The performance difference lemma is one of the cornerstone technical results in RL theory. It provides a mechanism for comparing two policies via one-step differences and has an elegant form in terms of the advantage function.

Q2: Prove the following lemma.

Lemma 2. *Let π_1, π_2 be arbitrary policies. Then*

$$V^{\pi_1}(\rho) - V^{\pi_2}(\rho) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d_\rho^{\pi_1}}\left[A^{\pi_2}(s,a)\right]. \tag{7}$$

3 NPG regret analysis

Owing to (6) and by absorbing the $(1-\gamma)$ term into the learning rate. It is natural to consider using an approximation to the advantage function given by a vector $w \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Informally, we want

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

Then, we can simply perform the updates $\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta w^{(t)}$. This corresponds to NPG, because, with the tabular softmax representation, the gradient term is $e_{s,a} - \sum_{a'} e_{s,a'}\pi^{(t)}(a' \mid s)$. This means that we want $w^{(t)}$ to be equal to $A^{(t)}$ up to a state-dependent offset. In fact, we can see that if we set $w^{(t)}(s,a) = Q^{(t)}(s,a)$ then the above is satisfied with equality.

To capture both approximation and estimation errors, we define

$$\mathrm{err}_t := \mathbb{E}_{s\sim d_\rho^{\tilde{\pi}}}\mathbb{E}_{a\sim\tilde{\pi}(\cdot|s)}\left[A^{(t)}(s,a) - \langle w^{(t)}, \nabla_\theta \log \pi^{(t)}(a \mid s)\rangle\right]. \tag{8}$$

Here $\tilde{\pi}$ is some reference policy that we will compete with in our analysis, e.g., it could be the optimal policy $\pi^\star$.

Q3: Prove the following regret lemma using Lemma 2.

Lemma 3 (NPG Regret Lemma)**.** *Fix comparison policy $\tilde{\pi}$ and assume that $\log \pi_\theta(a \mid s)$ is β smooth w.r.t., ℓ_2 norm:*

$$\forall \theta, \theta', s, a : |\log \pi_{\theta'}(a \mid s) - \log \pi_\theta(a \mid s) - \nabla \log \pi_\theta(a \mid s)(\theta' - \theta)| \le \frac{\beta}{2}\|\theta' - \theta\|_2^2. \tag{9}$$

Assume that $\sup_t \left\|w^{(t)}\right\|_2 \le W$ and that err_t is defined as in (8). Then the NPG iterates, given by $\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta w^{(t)}$, satisfy

$$\min_{t<T}\left\{V^{\tilde{\pi}}(\rho) - V^{(t)}(\rho)\right\} \le \frac{1}{1-\gamma}\left(\underbrace{\frac{\log|\mathcal{A}|}{\eta T} + \frac{\eta\beta W^2}{2}}_{MW\ style\ regret\ decomposition} + \frac{1}{T}\sum_{t=0}^{T-1}\mathrm{err}_t\right). \tag{10}$$

Remark 4. *In the solutions document, we sketch how to obtain a complete analysis for NPG, using this regret lemma as a starting point. The final steps highlight how this method relies on the distribution ρ for providing suitable coverage over the state space.*

Exercises for UCB-VI

We will consider the standard finite horizon MDP in this case $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, r, \{P_h\}, \mu_0)$, where $\mu_0 \in \Delta(\mathcal{S})$ is the initial state distribution, $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$, and $P_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$. For simplicity, we assume reward r and initial distribution μ_0 are known, but the transitions $\{P_h\}_{h=0}^{H-1}$ are unknown and need to be learned.

Throughout the section, we denote $V_h^\pi(s)$ as the expected total reward of the policy π starting at state s at time step h. We denote the expected total reward for policy π as $V^\pi := \mathbb{E}_{s\sim\mu_0} V_0^\pi(s)$. We denote $d_h^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$ as the state-action distribution of the policy π at time step h.

1 Proving Simulation Lemma

We start by proving the classic simulation lemma, which concerns the following important question: given a policy π, and two different rewards and transition dynamics $\{r_h, P_h\}_{h=0}^{H-1}$ and $\{\widehat{r}_h, \widehat{P}_h\}_{h=0}^{H-1}$, what is the difference between the policy's value under $\{r_h, P_h\}_{h=0}^{H-1}$ and under $\{\widehat{r}_h, \widehat{P}_h\}_{h=0}^{H-1}$.

Q1: Prove the following lemma.

Lemma 5 (Simulation lemma)**.** *Consider a policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$ and two models $\{r_h, P_h\}_{h=0}^{H-1}$ and $\{\widehat{r}_h, \widehat{P}_h\}_{h=0}^{H-1}$. Let V_h^π and $\widehat{V}_h^\pi$ denote the value function under $\{r_h, P_h\}_{h=0}^{H-1}$ and $\{\widehat{r}_h, \widehat{P}_h\}_{h=0}^{H-1}$ respectively (assume that the starting distribution μ is the same in both models). Then we have:*

$$V_0^\pi - \widehat{V}_0^\pi = \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^\pi} \left[r_h(s,a) + \mathbb{E}_{s'\sim P_h(s,a)} \widehat{V}_{h+1}^\pi(s') - \widehat{r}_h(s,a) - \mathbb{E}_{s'\sim \widehat{P}_h(s,a)} \widehat{V}_{h+1}^\pi(s') \right].$$

2 Optimism

Let us prove the following general result which is not tied to the tabular setting. Suppose have learned transitions based on data, say, $\{\widehat{P}_h\}_{h=0}^{H-1}$, and in addition, we have some uncertainty measure $b_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ for our model satisfying

$$\forall h, s, a : \left| \mathbb{E}_{s'\sim \widehat{P}_h(\cdot|s,a)} V_{h+1}^\star(s') - \mathbb{E}_{s'\sim P_h(\cdot|s,a)} V_{h+1}^\star(s') \right| \le b_h(s,a) \tag{11}$$

Here $V^\star$ is the optimal value function in the true MDP, with dynamics P. Suppose we perform value iteration inside the "bonus augmented MDP" $\widetilde{\mathcal{M}} := (\mathcal{S}, \mathcal{A}, \{r + b_h\}, \{\widehat{P}_h\}, H, \mu_0)$, i.e.,

$$\begin{aligned} \widehat{V}_H(s) &:= 0, \forall s; \\ \widehat{Q}_h(s,a) &:= \min\{H, r(s,a) + b_h(s,a) + \mathbb{E}_{s'\sim \widehat{P}_h(\cdot|s,a)} \widehat{V}_{h+1}(s')\}; \\ \widehat{V}_h(s) &= \max_a \widehat{Q}_h(s,a). \end{aligned}$$

And we define $\widehat{\pi}_h(s) := \operatorname{argmax}_a \widehat{Q}_h(s,a)$.

Q2: Prove the following statement.

Lemma 6 (Optimism)**.** *Assume* (11) *holds. Let $Q_h^\star(s,a)$ be the optimal Q function of the original MDP $\mathcal{M}$. Then $(\widehat{Q}_h, \widehat{V}_h)$ are pointwise optimistic, that is $\widehat{Q}_h(s,a) \ge Q_h^\star(s,a), \forall s, a,$ and $\widehat{V}_h(s) \ge V_h^\star(s), \forall s.$*

3 Regret Decomposition

Next, we will condition on the event in (11) being true and consider the regret of the policy $\widehat{\pi}$ computed by value iteration in the bonus-augmented model $\widetilde{\mathcal{M}}$.

Q3: Using the fact that $\widehat{V}_h(s)$ is an optimistic estimate, prove the following statement.

Lemma 7 (Regret Decomposition)**.** *The regret is upper bounded as:*

$$V^\star - V^{\widehat{\pi}} \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\widehat{\pi}}} \left[b_h(s,a) + H \|\widehat{P}_h(s,a) - P_h(s,a)\|_1 \right].$$

Observe that the proof is quite similar to that of the simulation lemma.

4 Proving UCB-VI has valid bonus

Let us consider a particular iteration t. Recall that in UCB-VI, we set the reward bonus $b_{t,h}(s,a) = \min\{H, 2H\sqrt{\frac{\ln(SAHT/\delta)}{N_{t,h}(s,a)}}\}$. And recall that we estimate the transition operator $\widehat{P}_{t,h}(s'|s,a)$ using the observed frequencies.

Q4: Prove the following result regarding the estimated model's error.

Lemma 8. *With probability at least $1-\delta$, for all $t \in [N]$, for all $s, a \in \mathcal{S} \times \mathcal{A}$, and for all $h \in [H]$ we must have:*

$$\left| \mathbb{E}_{s'\sim \widehat{P}_{t,h}(\cdot|s,a)} V^\star_{h+1}(s') - \mathbb{E}_{s'\sim P_h(\cdot|s,a)} V^\star_{h+1}(s') \right| \leq b_{t,h}(s,a)$$
$$\left\| \widehat{P}_{t,h}(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right\|_1 \leq 2\sqrt{\frac{S \ln(SAHN/\delta)}{N_{t,h}(s,a)}}.$$

Note that the first inequality in the above lemma indicates that with $b_{t,h}(s,a)$ as above, performing VI inside the bonus augmented model gives us an optimistic policy, via Lemma 6.

5 Concluding the proof

Now conditioned on the event in Lemma 8 being true, we can proceed to conclude the proof as follows. Using optimism and the fact that $\widehat{V}_{t,0}(s) \geq V_0^\star(s)$, we immediately have the following upper bound for the total regret across N iterations,

$$\text{Regret}_N = \sum_{t=0}^{N-1} V^\star - V^{\pi_t} \lesssim \sum_{t=0}^{N-1}\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi_t}} \left[\sqrt{\frac{\ln(SAHN/\delta)}{N_{t,h}(s,a)}} + H\sqrt{\frac{S\ln(SAHN/\delta)}{N_{t,h}(s,a)}} \right] \tag{12}$$
$$\lesssim H \sum_{t=0}^{N-1}\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi_t}} \left[\sqrt{\frac{S\ln(SAHN/\delta)}{N_{t,h}(s,a)}} \right] \tag{13}$$

Q5: The last step to conclude the proof is to prove the following lemma

Lemma 9 (Confidence sum)**.** *We have:*

$$\sum_{t=0}^{T-1}\sum_{h=0}^{H-1} \sqrt{\frac{1}{N_{t,h}(s_{t,h}, a_{t,h})}} \leq H\sqrt{SAN}.$$

Hint: Use the fact that $N_{t+1,h}(s_{t,h}, a_{t,h}) = N_{t,h}(s_{t,h}, a_{t,h}) + 1$, since $(s_{t,h}, a_{t,h})$ is visited at time step h of the t^{th} episode.

Note that we cannot directly plug in the above result into the regret formulation yet, as the regret involves expectations under $d_h^{\pi_t}$. However, the difference between can be bounded by a standard martingale difference argument, which we omit from this exercise.