Exercises for Natural Policy Gradient

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

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

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

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

\[
\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in A} \exp(\theta_{s,a'})},
\]

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

\[
\theta^{(t+1)} = \theta^{(t)} + \eta F_\rho(\theta^{(t)}) \nabla_\theta V^{(t)}(\rho),
\]

and \(V^{(t)}(\rho)\) is the value of policy \(\pi_\theta^{(t)}\) from initial distribution \(\rho\). 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

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

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 \(\mathbb{R}^{|S||A|}\) and instead show that

\[
\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1 - \gamma} A^{(t)}(\cdot, \cdot) + \eta v
\]

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 \( \pi_1, \pi_2 \) be arbitrary policies. Then
\[
V^{\pi_1}(\rho) - V^{\pi_2}(\rho) = \frac{1}{1 - \gamma} \mathbb{E}_{(s,a) \sim d^\pi} [A^{\pi_2}(s,a)].
\] (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 \(w \in \mathbb{R}^{|S||A|}\). Informally, we want
\[
A(t)(s,a) \approx \langle w(t), \nabla \theta \log \pi(t)(a | s) \rangle.
\]

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 \(e_{s,a} - \sum_{a'} e_{s,a'} \pi(t)(a' | s)\). 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
\[
\text{err}_t := \mathbb{E}_{s \sim \rho} \mathbb{E}_{a \sim \pi(\cdot|s)} \left[ A(t)(s,a) - \langle w(t), \nabla \theta \log \pi(t)(a | s) \rangle \right].
\] (8)

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

Q3: Prove the following regret lemma using Lemma 2.
Lemma 3 (NPG Regret Lemma). Fix comparison policy \(\tilde{\pi}\) and assume that \(\log \pi_\theta(a | s)\) is \(\beta\) smooth w.r.t., \(\ell_2\) norm:
\[
\forall \theta, \theta', s, a : | \log \pi_\theta(a | s) - \log \pi_\theta(a | s) - \nabla \log \pi_\theta(a | s)(\theta' - \theta) | \leq \frac{\beta}{2} \| \theta' - \theta \|^2_2.
\] (9)

Assume that \(\sup_t \| w(t) \|_2 \leq W \) and that \(\text{err}_t\) is defined as in (8). Then the NPG iterates, given by \(\theta(t+1) \leftarrow \theta(t) + \eta w(t)\), satisfy
\[
\min_{t < T} \left\{ V^{\pi}(\rho) - V^{(t)}(\rho) \right\} \leq \frac{1}{1 - \gamma} \left( \frac{\log |A|}{\eta T} + \frac{\eta \beta W^2}{2} + \frac{1}{T} \sum_{t=0}^{T-1} \text{err}_t \right). \] (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 \(\rho\) for providing suitable coverage over the state space.
Exercises for UCB-VI

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

Throughout the section, we denote $V^\pi_h(s)$ as the expected total reward of the policy $\pi$ starting at state $s$ at time step $h$. We denote the expected total reward for policy $\pi$ as $V^\pi := \mathbb{E}_{s \sim \mu_0} V^\pi_0(s)$. We denote $d^\pi_h \in \Delta(S \times A)$ as the state-action distribution of the policy $\pi$ 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 $\pi$, and two different rewards and transition dynamics $\{r_h, P_h\}_{h=0}^{H-1}$ and $\{\tilde{r}_h, \tilde{P}_h\}_{h=0}^{H-1}$, what is the difference between the policy’s value under $\{r_h, P_h\}_{h=0}^{H-1}$ and under $\{\tilde{r}_h, \tilde{P}_h\}_{h=0}^{H-1}$.

Q1: Prove the following lemma.

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

$$V^\pi_0 - \tilde{V}^\pi_0 = \sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim d^\pi_h} \left[ r_h(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \tilde{V}^\pi_{h+1}(s') - \tilde{r}_h(s,a) - \mathbb{E}_{s' \sim \tilde{P}_h(s,a)} \tilde{V}^\pi_{h+1}(s') \right].$$

2 Optimism

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

$$\forall h, s, a : \left| \mathbb{E}_{s' \sim \tilde{P}_h(\cdot|s,a)} V^*_{h+1}(s') - \mathbb{E}_{s' \sim \tilde{P}_h(\cdot|s,a)} \tilde{V}^\pi_{h+1}(s') \right| \leq b_h(s,a) \tag{11}$$

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

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

And we define $\pi_h(s) := \arg\max_a \tilde{Q}_h(s,a)$.

Q2: Prove the following statement.

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

3 Regret Decomposition

Next, we will condition on the event in (11) being true and consider the regret of the policy $\pi$ computed by value iteration in the bonus-augmented model $\mathcal{M}$.
Q3: Using the fact that $\hat{V}_h(s)$ is an optimistic estimate, prove the following statement.

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

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

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 $b_{t,h}(s,a) = \min\{ H, 2H \sqrt{\frac{\ln(SAHN/\delta)}{N_{t,h}(s,a)}} \}$. And recall that we estimate the transition operator $\hat{P}_{t,h}(s'|s,a)$ 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 $s, a \in S \times A$, and for all $h \in [H]$ we must have:

$$\left| \mathbb{E}_{s' \sim \hat{P}_{t,h}(.|s,a)} V^*_{h+1}(s') - \mathbb{E}_{s' \sim P_{h}(.|s,a)} V^*_{h+1}(s') \right| \leq b_{t,h}(s,a)$$

$$\left\| \hat{P}_{t,h}(.|s,a) - P_{h}(.|s,a) \right\|_1 \leq 2 \sqrt{\frac{S \ln(SAHN/\delta)}{N_{t,h}(s,a)}}.$$  

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 $\hat{V}_{t,0}(s) \geq V^*_t(s)$, we immediately have the following upper bound for the total regret across $N$ iterations,

$$\text{Regret}_N = \sum_{t=0}^{N-1} V^* - V^\pi \leq \sum_{t=0}^{N-1} \sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim d^{\pi}_h} \left[ \frac{\ln(SAHN/\delta)}{N_{t,h}(s,a)} + H \sqrt{\frac{S \ln(SAHN/\delta)}{N_{t,h}(s,a)}} \right]$$

$$\leq H \sum_{t=0}^{N-1} \sum_{h=0}^{H-1} \mathbb{E}_{s,a \sim d^{\pi}_h} \left[ \frac{\ln(SAHN/\delta)}{N_{t,h}(s,a)} \right]$$

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

Lemma 9 (Confidence sum). We have:

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

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 $d^{\pi}_h$. However, the difference between can be bounded by a standard martingale difference argument, which we omit from this exercise.