Solutions for Natural Policy Gradient Exercises

1 Closed form NPG update

First, observe that

\[ \nabla_\theta \log \pi_\theta(a | s) = e_{s,a} - \sum_{a'} e_{s,a'} \pi_\theta(a' | s). \]  

(1)

which implies that

\[ \forall s : \mathbb{E}_{a \sim \pi_\theta(\cdot | s)} [\nabla \log \pi_\theta(a | s)] = 0 \]

Next, by the definition of the Moore-Penrose pseudoinverse \((F_\rho^\theta)^\dagger\nabla V^{\pi_\theta}(\rho)\) is equal to the minimum norm solution of

\[ \min_w \|\nabla V^{\pi_\theta}(\rho) - F_\rho^\theta w\|_2^2 \]

Let us calculate this latter matrix vector product:

\[ F_\rho^\theta w = \mathbb{E}_{s \sim d^\rho_s} \mathbb{E}_{a \sim \pi_\theta(\cdot | s)} \left[ \nabla \log \pi_\theta(a | s) (w^\top \nabla \log \pi_\theta(a | s)) \right] \]

\[ = \mathbb{E}_{s \sim d^\rho_s} \mathbb{E}_{a \sim \pi_\theta(\cdot | s)} \left[ \nabla \log \pi_\theta(a | s) (w_{s,a} - \mathbb{E}_{a' \sim \pi(\cdot | s)} w_{s,a'}) \right] \]

\[ = \mathbb{E}_{s \sim d^\rho_s} \mathbb{E}_{a \sim \pi_\theta(\cdot | s)} [w_{s,a} \cdot \nabla \log \pi_\theta(a | s)] \]

Next, we use the advantage version of the policy gradient theorem to write:

\[ \nabla V^{\pi_\theta}(\rho) = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d^\rho_s} \mathbb{E}_{a \sim \pi_\theta(\cdot | s)} [A^{\pi_\theta}(s, a) \nabla \log \pi_\theta(a | s)] \]

Intuitively, at this point we can see that \(w_{s,a} = A^{\pi_\theta}(s, a)/(1 - \gamma)\) is a solution to the least squares problem, which is quite close to proving the result.

To be more formal, using \((1)\) again, we see that the \((s, a)^{th}\) element of both of these vectors are

\[ [F_\rho^\theta w]_{s,a} = d^\pi_\theta(s, a) \pi_\theta(a | s) \left( w_{s,a} - \sum_{a'} w_{s,a'} \pi_\theta(a' | s) \right), \]

\[ [\nabla V^{\pi_\theta}(\rho)]_{s,a} = d^\pi_\theta(s, a) \pi_\theta(a | s) (A^{\pi_\theta}(s, a)). \]

Here we are using that \(\mathbb{E}_{a \sim \pi_\theta(\cdot | s)} A^{\pi_\theta}(s, a) = 0\). Now we can more clearly see that all solutions with 0 square loss are of the form \(w_{s,a} = A^{\pi_\theta}(s, a) + v_s\) where the second term depends only on the state (it is constant across actions for that state). This proves the claim regarding the update for \(\theta\) and the update for \(\pi_\theta\) follows immediately, since the state-dependent offset can be absorbed into the normalization term.
2 Performance difference lemma

The proof is based on an un-rolling argument. Observe that

\[ V^{\pi_1}(\rho) - V^{\pi_2}(\rho) = \mathbb{E}_{s \sim \rho} \left[ E_{a \sim \pi_1(s)} Q^{\pi_1}(s, a) - V^{\pi_2}(s) \right] \]

\[ = \mathbb{E}_{s \sim \rho} \left[ E_{a \sim \pi_1(s)} \left( Q^{\pi_1}(s, a) - Q^{\pi_2}(s, a) \right) \right] + \mathbb{E}_{s \sim \rho} \left[ E_{a \sim \pi_1(s)} Q^{\pi_2}(s, a) - V^{\pi_2}(s) \right] \]

Now the first term gives the difference value starting from the second state visited by \( \pi_1 \):

\[ \mathbb{E}_{s \sim \rho} \left[ E_{a \sim \pi_1(s)} \left( Q^{\pi_1}(s, a) - Q^{\pi_2}(s, a) \right) \right] = \gamma \mathbb{E}_{s, a, s' \sim \rho \pi_{1}} V^{\pi_1}(s') - V^{\pi_2}(s') \]

Applying the same argument as above, we can express this in terms of the advantage function and value starting from the third state. To express this more concisely, let \( P_{\rho}^{\pi_{1}} \) be the distribution over infinitely long trajectories \( \tau = (s_0, a_0, s_1, a_1, \ldots) \) sampled by starting from \( \rho \) and taking actions according to \( \pi_{1} \). With this notation we have

\[ V^{\pi_{1}}(\rho) - V^{\pi_{2}}(\rho) = \mathbb{E}_{\tau \sim P_{\rho}^{\pi_{1}}} \left[ \sum_{t=0}^{\infty} \gamma^{t} A^{\pi_{2}}(s_t, a_t) \right] = \frac{1}{1 - \gamma} \mathbb{E}_{s, a \sim d_{\rho}^{\pi_{1}}} \left[ A^{\pi_{2}}(s, a) \right] \]

3 NPG regret analysis

We use a potential function argument, where our potential is the KL divergence between the comparator \( \tilde{\pi} \) and our iterate \( \hat{\pi}^{(t)} \) on the distribution induced by \( \tilde{\pi} \). Using smoothness, we have

\[ \mathbb{E}_{s \sim d_{\rho}^{\pi}} \left[ \text{KL}(\tilde{\pi}(\cdot \mid s) \parallel \pi^{(t)}(\cdot \mid s)) - \text{KL}(\tilde{\pi}(\cdot \mid s) \parallel \hat{\pi}^{(t+1)}(\cdot \mid s)) \right] \]

\[ = \mathbb{E}_{s, a \sim d_{\rho}^{\pi}} \left[ \log \left( \frac{\hat{\pi}(a \mid s)}{\pi^{(t)}(a \mid s)} \right) - \log \left( \frac{\tilde{\pi}(a \mid s)}{\pi^{(t+1)}(a \mid s)} \right) \right] \]

\[ \geq \mathbb{E}_{s, a \sim d_{\rho}^{\pi}} \left[ \left\langle \nabla \log \pi^{(t)}(a \mid s), \theta^{(t+1)} - \theta^{(t)} \right\rangle - \frac{\beta}{2} \| \theta^{(t+1)} - \theta^{(t)} \|^2 \right] \]

\[ = \mathbb{E}_{s, a \sim d_{\rho}^{\pi}} \left[ \eta \left\langle \nabla \log \pi^{(t)}(a \mid s), w^{(t)} \right\rangle - \frac{\eta^2 \beta}{2} \| w^{(t)} \|^2 \right] \]

\[ = (1 - \gamma) \eta \left( V^{\hat{\pi}}(\rho) - V^{\pi^{(t)}}(\rho) \right) - \frac{\eta^2 \beta}{2} \| w^{(t)} \|^2 - \eta \cdot \text{err}_t \]

The only inequality here is the lower bound implied by our smoothness assumption on \( \pi_{\theta}(a \mid s) \). The last equality is the performance difference lemma.

Now we can obtain a telescoping sum involving the KL divergences:

\[ \min_{0 \leq t < T} \left( V^{\hat{\pi}}(\rho) - V^{\pi^{(t)}}(\rho) \right) \leq \frac{1}{T} \sum_{t=0}^{T-1} \left( V^{\hat{\pi}}(\rho) - V^{\pi^{(t)}}(\rho) \right) \]

\[ \leq \frac{1}{1 - \gamma} \frac{1}{T} \sum_{t=0}^{T-1} \left[ \frac{1}{\eta} \mathbb{E}_{s \sim d_{\rho}^{\pi}} \left[ \text{KL}(\tilde{\pi}(\cdot \mid s) \parallel \pi^{(t)}(\cdot \mid s)) - \text{KL}(\tilde{\pi}(\cdot \mid s) \parallel \hat{\pi}^{(t+1)}(\cdot \mid s)) \right] + \frac{\eta \beta}{2} \| w^{(t)} \|^2 + \text{err}_t \right] \]

\[ \leq \frac{1}{1 - \gamma} \left( \frac{\log |A|}{T \eta} + \frac{\eta \beta W}{2} + \frac{1}{T} \sum_{t=0}^{T-1} \text{err}_t \right). \]
4 NPG Error Analysis (Sketch)

There are two steps remaining in the analysis of NPG with the softmax policy class in tabular settings. These steps involve controlling the $err_t$ terms above and highlight how the initial/exploratory distribution $\rho$ must ensure sufficient exploration.

The first step is fit $w^{(t)}$ somehow. As discussed previously, we want $w^{(t)}(s, a) \approx Q^{(t)}(s, a)$. Since we can sample from $\rho$, a natural population objective is

$$L_t(w) = \mathbb{E}_{s \sim \rho, a \sim \text{unif}(A)} \left[ (Q^{(t)}(s, a) - w(s, a))^2 \right].$$  \hfill (2)

We will optimize this objective from samples, in the usual way. We can get unbiased estimates of $Q^{(t)}(s, a)$ where $(s, a) \sim \rho \circ \text{unif}(A)$ by (1) sampling the initial state/action and subsequently executing $\pi^{(t)}$, (2) terminating the episode at each time $t$ with probability $1 - \gamma$, and (3) reporting the undiscounted sum of rewards up to termination. Call this random variable $\hat{R}$ and let $t^*$ denote the time step that we terminate. Then for any $(s, a)$ pair sampled from $\rho \circ \text{unif}(A)$, linearity of expectation gives:

$$\mathbb{E}_\pi[\hat{R} | s, a] = \mathbb{E}_\pi \left[ \sum_{t=1}^{\infty} \mathbf{1}\{t^* = T\} \sum_{t=1}^{T} r_t \mid s, a \right] = \mathbb{E}_\pi \left[ \sum_{t=1}^{\infty} r_t \mathbf{1}\{t^* \geq t\} \mid s, a \right] = \mathbb{E}_\pi \left[ \sum_{t=1}^{\infty} \gamma^t r_t \mid s, a \right] = Q^\pi(s, a)$$

With this procedure and since we are using the tabular representation, we can ensure that $L_t(w^{(t)}) \lesssim 1/N$ if we collect $N$ rollouts.

The last step is to use this guarantee to ensure that $err_t$ is small. The main conceptual point is that we have to perform a distribution shift from $\rho \circ \text{unif}$ to $d^\pi_\rho$. With an importance weighting argument, we can obtain

$$err_t := \mathbb{E}_{(s, a) \sim d^\pi_\rho} \left[ A^{(t)}(s, a) - \left\langle \nabla \log \pi^{(t)}(a \mid s), w^{(t)} \right\rangle \right]$$

$$= \mathbb{E}_{(s, a) \sim d^\pi_\rho} \left[ Q^{(t)}(s, a) - \mathbb{E}_{a' \sim \pi^{(t)}(\cdot \mid s)} Q^{(t)}(s, a') - w^{(t)}(s, a) + \mathbb{E}_{a' \sim \pi^{(t)}(\cdot \mid s)} w^{(t)}(s, a') \right]$$

$$\leq 2 \sqrt{|A|} \mathbb{E}_{s \sim d^\rho_\rho, a \sim \text{unif}(A)} \left[ (Q^{(t)}(s, a) - w^{(t)}(s, a))^2 \right]$$

$$\leq 2 |A| \sum_s \frac{d^\rho_\rho(s)}{\rho(s)} \cdot \rho(s) \mathbb{E}_{a \sim \text{unif}(A)} \left[ (Q^{(t)}(s, a) - w^{(t)}(s, a))^2 \right]$$

$$\leq 2 \sqrt{|A|} \cdot \left\| \frac{d^\rho_\rho}{\rho} \right\|_\infty \cdot L_t(w^{(t)}).$$

The density ratio term measures how well $\rho$ covers the states visited by the comparator policy $\tilde{\pi}$. This highlights the role of the reset distribution in providing good coverage for policy gradient methods.

**Remark 1 (Omitted details).** There are some details we are glossing over here. First, while the sample-based estimates of $Q^{(t)}(s, a)$ are bounded with high probability, they are technically unbounded, so the least squares generalization argument needs to account for this. Second, we need to set the smoothness parameter $\beta$ and the norm bound $W$ appropriately. For $W$, we can add a constraint in the least squares problem, since $Q^{(t)}(s, a) \in [0, \frac{1}{1-\gamma}]$. For $\beta$, since we are in the tabular representation, one can verify that $\beta = 1$ suffices.
Solutions for for UCB-VI Exercises

1 Prove Simulation lemma

By definition, we have:

\[ V_0^\pi - \hat{V}_0^\pi = \mathbb{E}_{s \sim \mu, a \sim \pi(\cdot|s_0)} \left[ Q_0^\pi(s, a) - \hat{Q}_0^\pi(s, a) \right] \]

\[ = \mathbb{E}_{s \sim \mu, a \sim \pi(\cdot|s_0)} \left[ \tau_0(s, a) - \hat{\tau}_0(s, a) + \mathbb{E}_{s' \sim P_h(s, a)} \hat{V}_1^\pi(s') - \mathbb{E}_{s' \sim P_h(s, a)} \hat{V}_1^\pi(s') \right] \]

\[ = \mathbb{E}_{s \sim \mu, a \sim \pi(\cdot|s_0)} \left[ \tau_0(s, a) - \hat{\tau}_0(s, a) + \mathbb{E}_{s' \sim P_h(s, a)} \hat{V}_1^\pi(s') - \mathbb{E}_{s' \sim P_h(s, a)} \hat{V}_1^\pi(s') \right] \]

Now recursively apply the same argument on the second term of the RHS of the above equation \( H \) times, and we can conclude the lemma.

2 Prove Optimism

We prove optimism via induction. In the base case, we consider the fictitious step \( H \) where we have \( \hat{V}_H(s) = 0 \) and \( V_0(s) = 0 \) for all \( s \). Thus, the base case holds.

The inductive hypothesis is that at time step \( h+1 \) with \( h \leq H-1 \), we have optimism, i.e., \( \hat{V}_{h+1}(s) \geq V^*_h(s), \forall s \).

We will prove that at time step \( h \), we have optimism as well.

Recall the update procedure for \( \hat{V}_h \). When \( \hat{V}_h(s) = H \), we have \( \hat{V}_h(s) = H \geq V^*_h(s) \) (since \( \| V^*_h \| \leq H \)). Thus, we have:

\[ \hat{Q}_h(s, a) - Q^*_h(s, a) = r(s, a) + b_h(s, a) + \mathbb{E}_{s' \sim P_h(s, a)} \hat{V}_{h+1}(s') - (r(s, a) + \mathbb{E}_{s' \sim P_h(s, a)} V^*_h(s')) \]

\[ = b_h(s, a) + \mathbb{E}_{s' \sim P_h(s, a)} \hat{V}_{h+1}(s') - \mathbb{E}_{s' \sim P_h(s, a)} V^*_h(s') \]

\[ \geq b_h(s, a) + \mathbb{E}_{s' \sim P_h(s, a)} V^*_h(s') - \mathbb{E}_{s' \sim P_h(s, a)} V^*_h(s') \]

\[ \geq b_h(s, a) - \mathbb{E}_{s' \sim P_h(s, a)} V^*_h(s') - \mathbb{E}_{s' \sim P_h(s, a)} V^*_h(s') \geq 0, \]

where the first inequality uses the inductive hypothesis that \( \hat{V}_{h+1}(s) \geq V^*_h(s), \forall s \) and the fact that \( P(s'|s, a) \geq 0 \) for all \( s, a, s' \), and the last inequality uses the given assumption on the bonuses \( b_h(s, a) \). Note that the above derivation holds for any \( (s, a) \), so we have proved that \( \hat{Q}_h \) is optimistic. This immediately implies that \( \hat{V}_h \) is also optimistic, since, if we let \( a^*(s) = \arg\max_a Q_h^*(s, a) \) we have

\[ V^*_h(s) = Q^*_h(s, a^*(s)) \leq \hat{Q}_h(s, a^*(s)) \leq \max_a \hat{Q}_h(s, a) \leq \hat{V}_h(s), \]

which verifies the inductive claim.

3 Regret Decomposition

Since \( \hat{V}_h(s) \geq V^*_h(s) \), we can upper bound the regret as follows:

\[ V^* - \hat{V}^\pi \leq \mathbb{E}_{s \sim \mu} \left[ \hat{V}_0(s) - V^*_0(s) \right] \]

Note here that \( \hat{V}_0(s) \) is the value of policy \( \hat{\pi} \) in the bonus augmented MDP \( \hat{M} \), while \( V^*_0(s) \) is the value of policy \( \hat{\pi} \) in the true MDP \( M \). Thus, we are in a position to apply an argument similar to the simulation lemma.

\[ \mathbb{E}_{s \sim \mu} \left[ \hat{V}_0(s) - V^*_0(s) \right] = \mathbb{E}_{s \sim \mu} \left[ \hat{Q}_0(s, \hat{\pi}(s)) - Q_0^\pi(s, \hat{\pi}(s)) \right] \]

\[ \leq \mathbb{E}_{s \sim \mu} \left[ b_0(s, \hat{\pi}(s)) + \mathbb{E}_{s' \sim P_h(s, \hat{\pi}(s))} \hat{V}_1(s') - \mathbb{E}_{s' \sim P_h(s, \hat{\pi}(s))} \hat{V}_1^\pi(s') \right] \]

\[ = \mathbb{E}_{s \sim \mu} \left[ b_0(s, \hat{\pi}(s)) + \mathbb{E}_{s' \sim P_h(s, \hat{\pi}(s))} \hat{V}_1(s') - \mathbb{E}_{s' \sim P_h(s, \hat{\pi}(s))} \hat{V}_1^\pi(s') + \mathbb{E}_{s' \sim P_h(s, \hat{\pi}(s))} \hat{V}_1(s') - \mathbb{E}_{s' \sim P_h(s, \hat{\pi}(s))} \hat{V}_1^\pi(s') \right] \]

\[ \leq \mathbb{E}_{s \sim \mu} \left[ b_0(s, \hat{\pi}(s)) + H \left\| \hat{P}_0(s, \hat{\pi}(s)) - P_0(s, \hat{\pi}(s)) \right\|_1 \right] + \mathbb{E}_{s \sim \mu} \left[ \hat{V}_1(s) - \hat{V}_1^\pi(s) \right] \]
Now recursively apply the same procedure on the second term of RHS of the above inequality \( H \) times, and we can conclude the proof. Note that this is basically the simulation lemma derivation except that in the first inequality above, we use the fact that \( \hat{Q}_h(s,a) = \min \left\{ H, b_h(s,a) + r_h(s,a) + E_{s' \sim P_h(s,a)} \hat{V}_{h+1}(s') \right\} \) and the inequality that \( \min\{a,b\} \leq b \).

### 4 Proving UCB-VI has valid bonus

Starting from this section, to simplify, we will use \( \lesssim \) to suppress absolute constants. First, via standard concentration on discrete distributions (e.g., Proposition A.4 in the AJKS monograph), for a fixed \((s,a,h,t)\) tuple, we must have that with probability at least \(1 - \delta\),

\[
\| \hat{P}_{t,h}(s,a) - P_h(s,a) \|_1 \lesssim \sqrt{\frac{SAH}{N_{t,h}(s,a)}}.
\]

Via a union bound over all \( s \in S, a \in A, h \in [H], t \in [N] \), we must have that with probability at least \(1 - \delta\):

\[
\| \hat{P}_{t,h}(s,a) - P_h(s,a) \|_1 \lesssim \sqrt{\frac{S \ln(1/\delta)}{N_{t,h}(s,a)}}.
\]

Now let us consider bounding \( \hat{P}_{t,h}(s,a)^\top V_{h+1} - P_h(s,a)^\top V_{h+1} \) (note that here we abuse notation a bit and treat \( P_h(s,a) \) and \( V_{h}^* \) as vectors in \( \mathbb{R}^{|S|} \)). We have:

\[
\hat{P}_{t,h}(s,a)^\top V_{h+1} - P_h(s,a)^\top V_{h+1} = \frac{1}{N_{t,h}(s,a)} \sum_{i=1}^{t-1} \mathbf{1}\{ (s_{i,h}, a_{i,h}) = (s,a) \} V_{h+1}^* - P_h(s,a)^\top V_{h+1},
\]

and note that for iteration \( i \) where \( s_{i,h}, a_{i,h} = s, a \), we have \( V_{h+1}^*(s_{i,h+1}) \) being an unbiased estimate of \( P_h(s,a)^\top V_{h+1}^* \). Also note that \( \| V_{h+1}^* \|_\infty \lesssim H \) which implies that our random variables are bounded in \([0,H]\). Thus, we can apply the standard Hoeffding’s inequality. For a fixed \((s,a,h,t)\) tuple, with probability at least \(1 - \delta\), we have:

\[
\left| \hat{P}_{t,h}(s,a)^\top V_{h+1} - P_h(s,a)^\top V_{h+1} \right| \lesssim H \sqrt{\frac{\ln(1/\delta)}{N_{t,h}(s,a)}}.
\]

Again with union bound over all \( s, a, h, t \) tuples, we immediately have that with probability at least \(1 - \delta\),

\[
\left| \hat{P}_{t,h}(s,a)^\top V_{h+1} - P_h(s,a)^\top V_{h+1} \right| \lesssim H \sqrt{\frac{\ln(SAH/\delta)}{N_{t,h}(s,a)}}.
\]

This concludes the proof.

### 5 Concluding the proof: bounding the confidence sum

We proceed to upper bound the confidence sum as follows:

\[
\sum_{t=1}^{N} \sum_{h=0}^{H-1} \frac{1}{\sqrt{\tilde{N}_{t,h}(s_{t,h}, a_{t,h})}} = \sum_{h=0}^{H-1} \sum_{s,a}^{N_{N,h}(s,a)} \frac{1}{\sqrt{i}} \lesssim \sum_{h=0}^{H-1} \sum_{s,a}^{N_{N,h}(s,a)} \frac{\sqrt{N_{N,h}(s,a)}}{\sqrt{i}} \leq \sum_{h=0}^{H-1} \sqrt{SA} \sum_{s,a} N_{N,h}(s,a) = \sum_{h=0}^{H-1} \sqrt{SAN} = H \sqrt{SAN}.
\]

Here in the first inequality, we use that \( \sum_{i=1}^{n} 1/\sqrt{i} \leq 2\sqrt{n} \), and the second inequality is by Cauchy-Schwarz. This concludes the proof.