

## 1. SoftMax parameterisation

a) Show for the tabular softmax parametrisation from Example 5.0.2 that

$$\frac{\partial \log(\pi^\theta(a; s))}{\partial \theta_{s', a'}} = \mathbf{1}_{\{s=s'\}}(\mathbf{1}_{\{a=a'\}} - \pi^\theta(a'; s'))$$

and for the linear softmax with features  $\Phi(s, a)$

$$\nabla \log(\pi^\theta(a; s)) = \Phi(s, a) - \sum_{a'} \pi^\theta(a'; s) \Phi(s, a').$$

*Solution:*

By the definition of the tabular softmax parametrisation ( $\pi^\theta(a; s) = \frac{e^{\theta_{s,a}}}{\sum_{\tilde{a} \in \mathcal{A}} e^{\theta_{s,\tilde{a}}}}$ ) we have

$$\log(\pi^\theta(a; s)) = \theta_{s,a} - \log\left(\sum_{\tilde{a} \in \mathcal{A}} e^{\theta_{s,\tilde{a}}}\right).$$

So for the derivative holds if  $s' \neq s$  then

$$\frac{\partial \log(\pi^\theta(a; s))}{\partial \theta_{s', a'}} = 0.$$

If  $s' = s$  and  $a' = a$  then

$$\frac{\partial \log(\pi^\theta(a; s))}{\partial \theta_{s,a}} = 1 - \frac{1}{\sum_{\tilde{a} \in \mathcal{A}} e^{\theta_{s,\tilde{a}}}} e^{\theta_{s,a}} = 1 - \pi^\theta(a; s)$$

and if  $s' = s$  and  $a' \neq a$  then

$$\frac{\partial \log(\pi^\theta(a; s))}{\partial \theta_{s,a'}} = -\frac{1}{\sum_{\tilde{a} \in \mathcal{A}} e^{\theta_{s,\tilde{a}}}} e^{\theta_{s,a'}} = -\pi^\theta(a'; s).$$

Summing up we get

$$\frac{\partial \log(\pi^\theta(a; s))}{\partial \theta_{s', a'}} = \mathbf{1}_{\{s = s'\}}(\mathbf{1}_{\{a=a'\}} - \pi^\theta(a'; s)).$$

Similarly, for the linear softmax with features  $\Phi(s, a)$  we have

$$\log(\pi^\theta(a; s)) = \theta \cdot \Phi(s, a) - \log\left(\sum_{a' \in \mathcal{A}} e^{\theta \cdot \Phi(s, a')}\right).$$

The derivative can be calculated without considering specific cases, we obtain

$$\begin{aligned} \nabla \log(\pi^\theta(a; s)) &= \Phi(s, a) - \frac{1}{\sum_{a' \in \mathcal{A}} e^{\theta \cdot \Phi(s, a')}} \sum_{a' \in \mathcal{A}} \Phi(s, a') e^{\theta \cdot \Phi(s, a')} \\ &= \Phi(s, a) - \sum_{a' \in \mathcal{A}} \Phi(s, a') \frac{e^{\theta \cdot \Phi(s, a')}}{\sum_{a' \in \mathcal{A}} e^{\theta \cdot \Phi(s, a')}} \\ &= \Phi(s, a) - \sum_{a' \in \mathcal{A}} \Phi(s, a') \pi^\theta(a'; s). \end{aligned}$$

- b) Show that the tabular and linear softmax parametrisation fulfill Assumption 5.1.12, i.e. that  $\log(\pi^\theta(a; s))$  is  $L$ -smooth and  $\nabla \log(\pi^\theta(a; s))$  has bounded norm for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ .

*Solution:*

We start with the tabular softmax case. First, similar to the above calculations we can see that

$$\begin{aligned} \frac{\partial \pi^\theta(a; s)}{\partial \theta_{s', a'}} &= \frac{\partial}{\partial \theta_{s', a'}} \frac{\exp(\theta_{s, a})}{\sum_{\tilde{a} \in \mathcal{A}} \exp(\theta_{s, \tilde{a}})} \\ &= \frac{\sum_{\tilde{a} \in \mathcal{A}} \exp(\theta_{s, \tilde{a}}) \frac{\partial}{\partial \theta_{s', a'}} \exp(\theta_{s, a}) - \exp(\theta_{s, a}) \frac{\partial}{\partial \theta_{s', a'}} \sum_{\tilde{a} \in \mathcal{A}} \exp(\theta_{s, \tilde{a}})}{(\sum_{\tilde{a} \in \mathcal{A}} \exp(\theta_{s, \tilde{a}}))^2} \\ &= \mathbf{1}_{\{s=s', a=a'\}} \frac{\exp(\theta_{s, a})}{\sum_{\tilde{a} \in \mathcal{A}} \exp(\theta_{s, \tilde{a}})} - \mathbf{1}_{\{s=s'\}} \frac{\exp(\theta_{s, a})}{\sum_{\tilde{a} \in \mathcal{A}} \exp(\theta_{s, \tilde{a}})} \cdot \frac{\exp(\theta_{s, a'})}{\sum_{\tilde{a} \in \mathcal{A}} \exp(\theta_{s, \tilde{a}})} \\ &= \mathbf{1}_{\{s=s', a=a'\}} \underbrace{\pi^\theta(a; s)}_{\in [0, 1]} - \mathbf{1}_{\{s=s'\}} \underbrace{\pi^\theta(a; s) \pi^\theta(a', s)}_{\in [0, 1]} \in [-1, 1]. \end{aligned}$$

Therefore, with the mean value theorem we obtain that  $\pi^\theta(a; s)$  is Lipschitz continuous for fixed  $(s, a) \in \mathcal{S} \times \mathcal{A}$  with Lipschitz-constant 1:

$$\|\pi^{\theta_1}(a; s) - \pi^{\theta_2}(a; s)\|_\infty = \|\nabla_\theta \pi^{\theta^*}(a; s)\|_\infty \|\theta_1 - \theta_2\|_\infty \leq \|\theta_1 - \theta_2\|_\infty.$$

This and part a) allow us to compute that for all  $\theta_{s', a'}$

$$\left| \frac{\partial \log(\pi^{\theta_1}(a; s))}{\partial \theta_{s', a'}} - \frac{\partial \log(\pi^{\theta_2}(a; s))}{\partial \theta_{s', a'}} \right| = \mathbf{1}_{\{s=s'\}} |\pi^{\theta_1}(a'; s') - \pi^{\theta_2}(a'; s')| \leq \|\theta_1 - \theta_2\|_\infty$$

and thus the 1-smoothness of  $\log(\pi^\theta(a; s))$ . Furthermore, since  $\pi^\theta$  is a measure and its values therefore below 1 we can directly compute for any  $\theta_{s', a'}$ :

$$\left| \frac{\partial \log(\pi^\theta(a; s))}{\partial \theta_{s', a'}} \right| \leq 1,$$

which finishes the proof in the tabular setting. Now let us consider the linear softmax parametrisation. Similarly to above we obtain Lipschitz-continuity of the gradient:

$$\begin{aligned} \nabla_\theta \pi^\theta(a; s) &= \nabla_\theta \frac{\exp(\theta \cdot \phi(s, a))}{\sum_{\tilde{a} \in \mathcal{A}} \exp(\theta \cdot \phi(s, \tilde{a}))} \\ &= \frac{\sum_{\tilde{a} \in \mathcal{A}} \exp(\theta \cdot \phi(s, \tilde{a})) \nabla_\theta \exp(\theta \cdot \phi(s, a)) - \exp(\theta \cdot \phi(s, a)) \nabla_\theta \sum_{\tilde{a} \in \mathcal{A}} \exp(\theta \cdot \phi(s, \tilde{a}))}{(\sum_{\tilde{a} \in \mathcal{A}} \exp(\theta \cdot \phi(s, \tilde{a}))^2} \\ &= \underbrace{\pi^\theta(a; s)}_{\in [0, 1]} \underbrace{\phi(s, a)}_{\in [0, 1]} - \underbrace{\pi^\theta(a; s)}_{\in [0, 1]} \sum_{\tilde{a} \in \mathcal{A}} \underbrace{\pi^\theta(\tilde{a}; s)}_{\sum_a=1} \phi(s, \tilde{a}) \in \left[ -\max_{(s, a) \in \mathcal{S} \times \mathcal{A}} \phi(s, a), \max_{(s, a) \in \mathcal{S} \times \mathcal{A}} \phi(s, a) \right] \end{aligned}$$

Again, using part a) we can calculate that

$$\begin{aligned} \|\nabla \log(\pi^{\theta_1}(a; s)) - \nabla \log(\pi^{\theta_2}(a; s))\|_\infty &\leq \sum_{a' \in \mathcal{A}} \phi(s, a') \|\pi^{\theta_1}(a', s) - \pi^{\theta_2}(a', s)\|_\infty \\ &\leq |\mathcal{A}| \max_{(s, a) \in \mathcal{S} \times \mathcal{A}} \phi(s, a)^2 \|\theta_1 - \theta_2\|_\infty \end{aligned}$$

and thus the  $L$ -smoothness of  $\log(\pi^\theta(a; s))$ . Analogously we obtain the boundedness of its gradient, finishing the proof:

$$|\nabla_\theta \pi^\theta(a; s)| = |\phi(s, a) - \sum_{a' \in \mathcal{A}} \pi^\theta(a'; s) \phi(s, a')| \leq \max_{(s, a) \in \mathcal{S} \times \mathcal{A}} |\phi(s, a)| - \min_{(s, a) \in \mathcal{S} \times \mathcal{A}} |\phi(s, a)|.$$

## 2. Policy Gradient Theorems

For episodic MDPs (the MDP terminates almost surely under all policies  $\pi_\theta$ ), we can get rid of the assumption of the existence of  $\nabla J_s(\theta)$ . Go through the proof of Lemma 5.1.6 and argue why it is enough to assume the existence of  $\nabla \pi_\theta(\cdot; s)$  for all  $s \in \mathcal{S}$ .

*Solution:*

Recall the proof of Lemma 5.1.6 (Policy Gradient Theorem in infinite time horizon). The first step of the proof was to show by induction that

$$\begin{aligned} \nabla J_s(\theta) &= \sum_{t=0}^n \sum_{s' \in \mathcal{S}} \gamma^t p(s \rightarrow s'; t, \pi^\theta) \sum_{a \in \mathcal{A}_{s'}} \nabla \pi^\theta(a; s') Q^{\pi^\theta}(s', a) \\ &\quad + \sum_{s'} \gamma^{n+1} p(s \rightarrow s'; t, \pi^\theta) \nabla J_{s'}(\theta) \end{aligned}$$

for all  $n \in \mathbb{N}$ . Now assume that the MDP is terminating. Then there exists a random time  $T$ , which is almost surely finite, such that  $p(\hat{s}; \hat{s}, a) = 1$  and  $R(\hat{s}, a) = 0$  for all  $a \in \mathcal{A}_{\hat{s}}$ . Intuitively, we want to argue that the RHS regarding the claim proven by induction stated above exists because  $J_{\hat{s}}(\theta)$  is zero after the terminating time  $T$ . If we assume that  $\pi^\theta$  is differentiable in  $\theta$ , then

$$\sum_{t=0}^{T-1} \sum_{s' \in \mathcal{S}} \gamma^t p(s \rightarrow s'; t, \pi^\theta) \sum_{a \in \mathcal{A}_{s'}} \nabla \pi^\theta(a; s') Q^{\pi^\theta}(s', a)$$

exists almost surely. It remains to show that this is equal to the derivative of  $\nabla J_s(\theta)$ . By the termination we know that  $p(s \rightarrow \hat{s}; T, \pi^\theta) = 1$  and  $J_{\hat{s}}(\theta) = 0$ . Thus,

$$\begin{aligned} &\sum_{t=0}^{T-1} \sum_{s' \in \mathcal{S}} \gamma^t p(s \rightarrow s'; t, \pi^\theta) \sum_{a \in \mathcal{A}_{s'}} \nabla \pi^\theta(a; s') Q^{\pi^\theta}(s', a) \\ &= \sum_{t=0}^{T-1} \sum_{s' \in \mathcal{S}} p(s \rightarrow s'; t, \pi^\theta) \sum_{a \in \mathcal{A}_{s'}} \nabla \pi^\theta(a; s') Q^{\pi^\theta}(s', a) + \sum_{s'} \gamma^{T+1} p(s \rightarrow s'; T, \pi^\theta) \nabla J_{s'}(\theta) \end{aligned}$$

exists almost surely. Reading the equations in the proof of Lemma 5.1.6 backwards yields that this is equal to  $\nabla J_s(\theta)$ . We are allowed to interchange the derivative and the sums as stated there, because we know that the RHS exists.

## 3. Baseline Trick

- a) Show that the constant baseline  $b$  in Theorem 5.1.17 can be replaced by any deterministic state-dependent baseline  $b: \mathcal{S} \rightarrow \mathbb{R}$ , i.e.

$$\nabla_\theta J(\theta) = \mathbb{E}_s^{\pi^\theta} \left[ \sum_{t=0}^{T-1} \nabla_\theta (\log \pi^\theta(A_t; S_t)) (Q_t^{\pi^\theta}(S_t, A_t) - b(S_t)) \right].$$

*Solution:*

The computation is very similar to the computations in the lecture notes. Assume that  $b : \mathcal{S} \rightarrow \mathbb{R}$ , then

$$\begin{aligned} \mathbb{E}_s^{\pi^\theta} [\nabla_\theta (\log \pi^\theta(A_t; S_t)) b(S_t)] &= \sum_{s_t \in \mathcal{S}} \sum_{a_t \in \mathcal{A}_s} \mathbb{P}_s^{\pi^\theta}(S_t = s_t) \pi^\theta(a_t; s_t) \nabla_\theta (\log \pi^\theta(a_t; s_t)) b(s_t) \\ &= \sum_{s_t \in \mathcal{S}} \mathbb{P}_s^{\pi^\theta}(S_t = s_t) b(s_t) \sum_{a_t \in \mathcal{A}_s} \nabla_\theta \pi^\theta(a_t; s_t) \\ &= \sum_{s_t \in \mathcal{S}} \mathbb{P}_s^{\pi^\theta}(S_t = s_t) b(s_t) \nabla_\theta \underbrace{\sum_{a_t \in \mathcal{A}} \pi^\theta(a_t; s_t)}_{=1} = 0. \end{aligned}$$

If the baseline remains unaffected by the action, we can express the baseline separately from the summation over  $a$ . This condition is sufficient for the trick to be effective.

- b) Write down and prove the baseline gradient representation with baseline  $b : \mathcal{S} \rightarrow \mathbb{R}$  for infinite discounted MDPs.

*Solution:*

We aim to prove

$$\nabla J_s(\theta) = \sum_{s' \in \mathcal{S}} \rho_s^{\pi^\theta}(s') \sum_{a \in \mathcal{A}_{s'}} \nabla \pi^\theta(a; s') (Q^{\pi^\theta}(s', a) - b(s')),$$

for some  $b : \mathcal{S} \rightarrow \mathbb{R}$ . By the finiteness of the state and action space we have that

$$\begin{aligned} &\sum_{s' \in \mathcal{S}} \rho_s^{\pi^\theta}(s') \sum_{a \in \mathcal{A}_{s'}} \nabla \pi^\theta(a; s') b(s') \\ &= \sum_{s' \in \mathcal{S}} \rho_s^{\pi^\theta}(s') b(s') \nabla \underbrace{\sum_{a \in \mathcal{A}_{s'}} \pi^\theta(a; s')}_{=1} \\ &= 0. \end{aligned}$$

Hence, the claim follows from the policy gradient theorem for discounted MDPs (5.1.6) in the lecture.