Prof. Dr. Leif Döring
André Ferdinand, Sara Klein

Reinforcement Learning

10. Solution Sheet

1. SoftMax parameterisation

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*

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

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*

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

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

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

2. Policy Gradient Theorems

For episodic MDPs (the MDP terminates almost surely under all policies π_θ), we can get rid of the assumption of the existence of $\nabla J_s(\theta)$. Go through the proof of Theorem 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 Theorem 5.3.6 (Policy Gradient Theorem in infinite time horizon). The first step of the proof was to show by induction that

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

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 π^θ is differentiable in θ, then

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

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 $p(s \to \hat{s}; T, \pi^\theta) = 1$ and $J_{\hat{s}}(\theta) = 0$. Thus,

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

exists almost surely. Reading the equations in the proof of Theorem 5.3.6 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

Show that the constant baseline b in Theorem 5.2.1 can be replaced by any deterministic state-dependent baseline $b : \mathcal{S} \to \mathbb{R}$, i.e.

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

Solution:

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

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

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.