Exercises - Chapter 13

Carl Fredriksson, c@msp.se

Exercise 13.1

Use your knowledge of the gridworld and its dynamics to determine an exact symbolic expression for the optimal probability of selecting the **right** action in Example 13.1.

My answer:

Let A denote the second state and B the third state (S is the starting state). We can use the Bellman equation for state values

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\big[r + \gamma v_\pi(s')\big], \quad \text{for all} \quad s \in \mathcal{S}$$

to set up a system of equations

$$\begin{cases} v_\pi(S) = \pi(right)\big[-1 + v_\pi(A)\big] + \pi(left)\big[-1 + v_\pi(S)\big] \\ v_\pi(A) = \pi(right)\big[-1 + v_\pi(S)\big] + \pi(left)\big[-1 + v_\pi(B)\big] \\ v_\pi(S) = \pi(right)\big[-1\big] + \pi(left)\big[-1 + v_\pi(A)\big] \end{cases}$$

Let $p = \pi(right)$ and note that $\pi(left) = 1 - p$, we can rewrite the equations as

$$\begin{cases} v_\pi(S) = p\big[-1 + v_\pi(A)\big] + (1-p)\big[-1 + v_\pi(S)\big] \\ v_\pi(A) = p\big[-1 + v_\pi(S)\big] + (1-p)\big[-1 + v_\pi(B)\big] \\ v_\pi(S) = p\big[-1\big] + (1-p)\big[-1 + v_\pi(A)\big] \end{cases}$$

$$\begin{cases} v_\pi(S) = -1 + pv_\pi(A) + (1-p)v_\pi(S) \\ v_\pi(A) = -1 + pv_\pi(S) + (1-p)v_\pi(B) \\ v_\pi(S) = -1 + (1-p)v_\pi(A) \end{cases}$$

The system can be solved to get

$$v_\pi(S) = -2\frac{2-p}{p(1-p)}$$

Our goal is to compute $\arg\max_p \{v_\pi(S) \mid 0 < p < 1\}$ (p is a probability and both $p = 0$ and $p = 1$ means that the agent will never get to the goal state from the starting state). First we find all critical points, i.e. where the derivative is either zero or doesn't exist.

$$\frac{\delta v_\pi(S)}{\delta p} = -2\frac{p(1-p)(-1) - (2-p)(1-2p)}{p^2(1-p)^2}$$

The derivative doesn't exist for $p = 0$ or $p = 1$, but since neither critical point satisfies $0 < p < 1$, we can disregard them. That leaves us with the points where the derivative is zero.

$$\frac{\delta v_\pi(S)}{\delta p} = 0 \implies p(1-p)(-1) = (2-p)(1-2p)$$

This equation can be solved to get

$$p = 2 \pm \sqrt{2}$$

Only $p = 2 - \sqrt{2} \approx 0.5858$ satisfies $0 < p < 1$, which means that we have found our optimal probability p of selecting the **right** action. We can use this probability to check that we get the same value for the start state as given in example 13.1

$$v_\pi(S) = -2\frac{2-p}{p(1-p)} = -2\frac{2-(2-\sqrt{2})}{(2-\sqrt{2})(1-(2-\sqrt{2}))} = -6 - 4\sqrt{2} \approx 11.66$$

Exercise 13.2

Generalize the box on page 199, the policy gradient theorem (13.5), the proof of the policy gradient theorem (page 325), and the steps leading to the REINFORCE update equation (13.8), so that (13.8) ends up with a factor of γ^t and thus aligns with the general algorithm given in the pseudocode.

My answer:

This exercise was tough for me. I asked this question about it: https://ai.stackexchange.com/questions/40894/gammat-in-reinforce-update-sutton-barto-rl-book-exercise-13-2 (no response at the time of writing this).

Generalize box on 199

As explained in the box on 199, include a factor of γ in the second term of (9.2):

$$\eta_\gamma(s) = h(s) + \gamma \sum_{\overline{s}} \eta(\overline{s}) \sum_a \pi(a|\overline{s}) p(s|\overline{s}, a)$$

$$\mu_\gamma(s) = \frac{\eta_\gamma(s)}{\sum_{s'} \eta(s')}$$

I initially thought the definition of $\mu_\gamma(s)$ was

$$\mu_\gamma(s) = \frac{\eta_\gamma(s)}{\sum_{s'} \eta_\gamma(s')}$$

but I could not see how to finish the exercise with this definition. Given that we are assuming a single starting state s_0, we can also write $\eta_\gamma(s)$ as:

$$\eta_\gamma(s) = \sum_{k=0}^{\infty} \gamma^k Pr(s_0 \to s, k, \pi)$$

Generalize Proof of the Policy Gradient Theorem (episodic case)

I followed the steps in the proof on page 325 and added discounting.

$$\begin{aligned}
\nabla v_\pi(s) &= \nabla \left[\sum_a \pi(a|s) q_\pi(s,a) \right], \quad \text{for all } s \in \mathcal{S} \\
&= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla q_\pi(s,a) \right] \\
&= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla \sum_{s',r} p(s',r|s,a)(r + \gamma v_\pi(s')) \right] \\
&= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \gamma \pi(a|s) \sum_{s'} p(s'|s,a) \nabla v_\pi(s') \right] \\
&= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \gamma \pi(a|s) \sum_{s'} p(s'|s,a) \sum_{a'} \left[\nabla \pi(a'|s') q_\pi(s',a') + \gamma \pi(a'|s') \sum_{s''} p(s''|s',a') \nabla v_\pi(s'') \right] \right] \\
&= \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \gamma^k Pr(s \to x, k, \pi) \sum_a \nabla \pi(a|x) q_\pi(x,a)
\end{aligned}$$

$$\begin{aligned}
\nabla J(\theta) &= \nabla v_\pi(s_0) \\
&= \sum_s \left(\sum_{k=0}^{\infty} \gamma^k Pr(s_0 \to s, k, \pi) \right) \sum_a \nabla \pi(a|s) q_\pi(s,a) \\
&= \sum_s \eta_\gamma(s) \sum_a \nabla \pi(a|s) q_\pi(s,a) \\
&= \sum_{s'} \eta(s') \sum_s \frac{\eta_\gamma(s)}{\sum_{s'} \eta(s')} \sum_a \nabla \pi(a|s) q_\pi(s,a) \\
&= \sum_{s'} \eta_\gamma(s') \sum_s \mu_\gamma(s) \sum_a \nabla \pi(a|s) q_\pi(s,a) \\
&\propto \sum_s \mu_\gamma(s) \sum_a \nabla \pi(a|s) q_\pi(s,a)
\end{aligned}$$

Steps leading to the REINFORCE update equation (13.8)

$$
\begin{aligned}
\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu_\gamma(s) \sum_a q_\pi(s,a,\boldsymbol{\theta}) \nabla \pi(a|s) \\
&= \mathbb{E}_\pi \left[\gamma^t \sum_a \pi(a|S_t,\boldsymbol{\theta}) q_\pi(S_t,a) \frac{\nabla \pi(a|S_t,\boldsymbol{\theta})}{\pi(a|S_t,\boldsymbol{\theta})} \right] \\
&= \mathbb{E}_\pi \left[\gamma^t q_\pi(S_t,A_t) \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t,\boldsymbol{\theta})} \right] \\
&= \mathbb{E}_\pi \left[\gamma^t G_t \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t,\boldsymbol{\theta})} \right]
\end{aligned}
$$

Discounted REINFORCE update:

$$
\begin{aligned}
\boldsymbol{\theta}_{t+1} &\doteq \boldsymbol{\theta}_t + \alpha \gamma^t G_t \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t,\boldsymbol{\theta})} \\
&= \boldsymbol{\theta}_t + \alpha \gamma^t G_t \nabla \ln \pi(A_t|S_t,\boldsymbol{\theta})
\end{aligned}
$$

Exercise 13.3

In Section 13.1 we considered policy parameterizations using the soft-max in action preferences (13.2) with linear action preferences (13.3). For this parameterization, prove that the eligibility vector is

$$
\nabla \ln \pi(a|s,\boldsymbol{\theta}) = \mathbf{x}(s,a) - \sum_b \pi(b|s,\boldsymbol{\theta}) \mathbf{x}(s,b)
$$

using the definitions and elementary calculus.

My answer:

We have

$$
\pi(a|s,\boldsymbol{\theta}) = \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_b e^{h(s,b,\boldsymbol{\theta})}}
$$

with

$$
h(s,a,\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}(s,a)
$$

Let's start by rewriting the gradient

$$
\begin{aligned}
\nabla \ln \pi(a|s,\boldsymbol{\theta}) &= \nabla \ln \left(\frac{e^{\boldsymbol{\theta}^\top \mathbf{x}(s,a)}}{\sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}} \right) \\
&= \nabla \left(\ln e^{\boldsymbol{\theta}^\top \mathbf{x}(s,a)} - \ln \sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)} \right) \\
&= \nabla \ln e^{\boldsymbol{\theta}^\top \mathbf{x}(s,a)} - \nabla \ln \sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}
\end{aligned}
$$

Let

$$
\frac{\delta}{\delta \boldsymbol{\theta}_i}
$$

denote the i:th element of ∇. We have

$$
\begin{aligned}
\frac{\delta \ln e^{\boldsymbol{\theta}^\top \mathbf{x}(s,a)}}{\delta \boldsymbol{\theta}_i} &= \frac{\delta \ln e^{\boldsymbol{\theta}^\top \mathbf{x}(s,a)}}{\delta e^{\boldsymbol{\theta}^\top \mathbf{x}(s,a)}} \frac{\delta e^{\boldsymbol{\theta}^\top \mathbf{x}(s,a)}}{\delta \boldsymbol{\theta}_i} \\
&= \frac{1}{e^{\boldsymbol{\theta}^\top \mathbf{x}(s,a)}} \frac{\delta e^{\boldsymbol{\theta}^\top \mathbf{x}(s,a)}}{\delta \boldsymbol{\theta}^\top \mathbf{x}(s,a)} \frac{\delta \boldsymbol{\theta}^\top \mathbf{x}(s,a)}{\delta \boldsymbol{\theta}_i} \\
&= \frac{1}{e^{\boldsymbol{\theta}^\top \mathbf{x}(s,a)}} e^{\boldsymbol{\theta}^\top \mathbf{x}(s,a)} \mathbf{x}(s,a)_i \\
&= \mathbf{x}(s,a)_i
\end{aligned}
$$

$$
\begin{aligned}
\frac{\delta \ln \sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}}{\delta \boldsymbol{\theta}_i} &= \frac{\delta \ln \sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}}{\delta \sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}} \frac{\delta \sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}}{\delta \boldsymbol{\theta}_i} \\
&= \frac{1}{\sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}} \sum_b \frac{\delta e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}}{\delta \boldsymbol{\theta}_i} \\
&= \frac{1}{\sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}} \sum_b \frac{\delta e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}}{\delta \boldsymbol{\theta}^\top \mathbf{x}(s,b)} \frac{\delta \boldsymbol{\theta}^\top \mathbf{x}(s,b)}{\delta \boldsymbol{\theta}_i} \\
&= \frac{1}{\sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}} \sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)} \mathbf{x}(s,b)_i \\
&= \sum_b \frac{e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}}{\sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)}} \mathbf{x}(s,b)_i \\
&= \sum_b \pi(b|s,\boldsymbol{\theta}) \mathbf{x}(s,b)_i
\end{aligned}
$$

We can put the pieces together and finish the proof

$$
\begin{aligned}
\nabla \ln \pi(a|s,\boldsymbol{\theta}) &= \nabla \ln e^{\boldsymbol{\theta}^\top \mathbf{x}(s,a)} - \nabla \ln \sum_b e^{\boldsymbol{\theta}^\top \mathbf{x}(s,b)} \\
&= \mathbf{x}(s,a) - \sum_b \pi(b|s,\boldsymbol{\theta}) \mathbf{x}(s,b)
\end{aligned}
$$

Exercise 13.4

Show that for the Gaussian policy parameterization (Equations 13.19 and 13.20) the eligibility vector has the following two parts:

$$
\nabla \ln \pi(a|s,\boldsymbol{\theta}_\mu) = \frac{\nabla \pi(a|s,\boldsymbol{\theta}_\mu)}{\pi(a|s,\boldsymbol{\theta})} = \frac{1}{\sigma(s,\boldsymbol{\theta})^2}\big(a - \mu(s,\boldsymbol{\theta})\big)\mathbf{x}_\mu(s), \text{ and}
$$

$$
\nabla \ln \pi(a|s,\boldsymbol{\theta}_\sigma) = \frac{\nabla \pi(a|s,\boldsymbol{\theta}_\sigma)}{\pi(a|s,\boldsymbol{\theta})} = \left(\frac{\big(a - \mu(s,\boldsymbol{\theta})\big)^2}{\sigma(s,\boldsymbol{\theta})^2} - 1\right)\mathbf{x}_\sigma(s)
$$

My answer:

Let

$$
\frac{\delta}{\delta \boldsymbol{\theta}_{\mu,i}}
$$

denote the i:th element of $\nabla_{\boldsymbol{\theta}_\mu}$, and

$$
\frac{\delta}{\delta \boldsymbol{\theta}_{\sigma,i}}
$$

the i:th element of $\nabla_{\boldsymbol{\theta}_\sigma}$.

Proof for the first part:

$$
\begin{aligned}
\frac{\frac{\delta}{\delta \boldsymbol{\theta}_{\mu,i}}\pi(a|s,\boldsymbol{\theta}_\mu)}{\pi(a|s,\boldsymbol{\theta})} &= \frac{1}{\pi(a|s,\boldsymbol{\theta})}\frac{\delta}{\delta \boldsymbol{\theta}_{\mu,i}}\left[\frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2}}\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\right] \\
&= \frac{1}{\pi(a|s,\boldsymbol{\theta})}\left[\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\frac{\delta}{\delta \boldsymbol{\theta}_{\mu,i}}\frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2}} + \frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2}}\frac{\delta}{\delta \boldsymbol{\theta}_{\mu,i}}\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\right] \\
&= \frac{1}{\pi(a|s,\boldsymbol{\theta})}\left[\frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2}}\frac{\delta}{\delta \boldsymbol{\theta}_{\mu,i}}\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\right] \\
&= \frac{1}{\pi(a|s,\boldsymbol{\theta})\sigma(s,\boldsymbol{\theta})\sqrt{2}}\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\frac{\delta}{\delta \boldsymbol{\theta}_{\mu,i}}\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right) \\
&= \frac{1}{\pi(a|s,\boldsymbol{\theta})\sigma(s,\boldsymbol{\theta})\sqrt{2}}\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\frac{-1}{2}\left(\frac{\sigma(s,\boldsymbol{\theta})^2\frac{\delta}{\delta \boldsymbol{\theta}_{\mu,i}}\big(a-\mu(s,\boldsymbol{\theta})\big)^2 + \big(a-\mu(s,\boldsymbol{\theta})\big)^2\frac{\delta}{\delta \boldsymbol{\theta}_{\mu,i}}\sigma(s,\boldsymbol{\theta})^2}{\sigma(s,\boldsymbol{\theta})^4}\right) \\
&= \frac{1}{\pi(a|s,\boldsymbol{\theta})\sigma(s,\boldsymbol{\theta})\sqrt{2}}\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\frac{-1}{2}\frac{-2\big(a-\mu(s,\boldsymbol{\theta})\big)\frac{\delta}{\delta \boldsymbol{\theta}_{\mu,i}}\mu(s,\boldsymbol{\theta})}{\sigma(s,\boldsymbol{\theta})^2} \\
&= \frac{1}{\pi(a|s,\boldsymbol{\theta})\sigma(s,\boldsymbol{\theta})\sqrt{2}}\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)\mathbf{x}_\mu(s)_i}{\sigma(s,\boldsymbol{\theta})^2} \\
&= \frac{1}{\left(\frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2}}\exp\left(-\frac{\left(a-\mu(s,\boldsymbol{\theta})\right)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\right)\sigma(s,\boldsymbol{\theta})\sqrt{2}}\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)\mathbf{x}_\mu(s)_i}{\sigma(s,\boldsymbol{\theta})^2} \\
&= \frac{1}{\sigma(s,\boldsymbol{\theta})^2}\big(a-\mu(s,\boldsymbol{\theta})\big)\mathbf{x}_\mu(s)_i
\end{aligned}
$$

$$
\implies \nabla \ln \pi(a|s,\boldsymbol{\theta}_\mu) = \frac{\nabla \pi(a|s,\boldsymbol{\theta}_\mu)}{\pi(a|s,\boldsymbol{\theta})} = \frac{1}{\sigma(s,\boldsymbol{\theta})^2}\big(a-\mu(s,\boldsymbol{\theta})\big)\mathbf{x}_\mu(s)
$$

Proof for the second part:

$$
\begin{aligned}
\frac{\frac{\delta}{\delta \boldsymbol{\theta}_{\sigma,i}}\pi(a|s,\boldsymbol{\theta}_\sigma)}{\pi(a|s,\boldsymbol{\theta})} &= \frac{1}{\pi(a|s,\boldsymbol{\theta})}\frac{\delta}{\delta \boldsymbol{\theta}_{\sigma,i}}\left[\frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2}}\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\right] \\
&= \frac{1}{\pi(a|s,\boldsymbol{\theta})}\left[\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\frac{\delta}{\delta \boldsymbol{\theta}_{\sigma,i}}\frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2}} + \frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2}}\frac{\delta}{\delta \boldsymbol{\theta}_{\sigma,i}}\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\right] \\
&= \frac{1}{\pi(a|s,\boldsymbol{\theta})}\left[\frac{-\exp\left(-\frac{\left(a-\mu(s,\boldsymbol{\theta})\right)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)}{\sigma(s,\boldsymbol{\theta})\sqrt{2}}\mathbf{x}_\sigma(s)_i + \frac{\exp\left(-\frac{\left(a-\mu(s,\boldsymbol{\theta})\right)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)}{\sigma(s,\boldsymbol{\theta})\sqrt{2}}\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{\sigma(s,\boldsymbol{\theta})^2}\mathbf{x}_\sigma(s)_i\right] \\
&= \frac{1}{\left(\frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2}}\exp\left(-\frac{\left(a-\mu(s,\boldsymbol{\theta})\right)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\right)}\left[\frac{-\exp\left(-\frac{\left(a-\mu(s,\boldsymbol{\theta})\right)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)}{\sigma(s,\boldsymbol{\theta})\sqrt{2}} + \frac{\exp\left(-\frac{\left(a-\mu(s,\boldsymbol{\theta})\right)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)}{\sigma(s,\boldsymbol{\theta})\sqrt{2}}\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{\sigma(s,\boldsymbol{\theta})^2}\right]\mathbf{x}_\sigma(s)_i \\
&= \left(\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{\sigma(s,\boldsymbol{\theta})^2} - 1\right)\mathbf{x}_\sigma(s)_i
\end{aligned}
$$

$$
\implies \nabla \ln \pi(a|s,\boldsymbol{\theta}_\sigma) = \frac{\nabla \pi(a|s,\boldsymbol{\theta}_\sigma)}{\pi(a|s,\boldsymbol{\theta})} = \left(\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{\sigma(s,\boldsymbol{\theta})^2} - 1\right)\mathbf{x}_\sigma(s)
$$

I computed the derivatives in the second equality separately:

$$
\begin{aligned}
\frac{\delta}{\delta \boldsymbol{\theta}_{\sigma,i}}\frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2}} &= \frac{1}{\sqrt{2}}\frac{-1}{\sigma(s,\boldsymbol{\theta})^2}\frac{\delta}{\delta \boldsymbol{\theta}_{\sigma,i}}\sigma(s,\boldsymbol{\theta}) \\
&= \frac{1}{\sqrt{2}}\frac{-1}{\sigma(s,\boldsymbol{\theta})^2}\sigma(s,\boldsymbol{\theta})\mathbf{x}_\sigma(s)_i \\
&= \frac{-1}{\sigma(s,\boldsymbol{\theta})\sqrt{2}}\mathbf{x}_\sigma(s)_i
\end{aligned}
$$

$$
\begin{aligned}
\frac{\delta}{\delta\boldsymbol{\theta}_{\sigma,i}}\exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right) &= \exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\frac{-1}{2}\frac{\delta}{\delta\boldsymbol{\theta}_{\sigma,i}}\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{\sigma(s,\boldsymbol{\theta})^2} \\
&= \exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\frac{-1}{2}\frac{\sigma(s,\boldsymbol{\theta})^2\frac{\delta}{\delta\boldsymbol{\theta}_{\sigma,i}}\big(a-\mu(s,\boldsymbol{\theta})\big)^2-\big(a-\mu(s,\boldsymbol{\theta})\big)^2\frac{\delta}{\delta\boldsymbol{\theta}_{\sigma,i}}\sigma(s,\boldsymbol{\theta})^2}{\sigma(s,\boldsymbol{\theta})^4} \\
&= \exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2\sigma(s,\boldsymbol{\theta})\frac{\delta}{\delta\boldsymbol{\theta}_{\sigma,i}}\sigma(s,\boldsymbol{\theta})}{\sigma(s,\boldsymbol{\theta})^4} \\
&= \exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2\sigma(s,\boldsymbol{\theta})^2\mathbf{x}_\sigma(s)_i}{\sigma(s,\boldsymbol{\theta})^4} \\
&= \exp\left(-\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)\frac{\big(a-\mu(s,\boldsymbol{\theta})\big)^2}{\sigma(s,\boldsymbol{\theta})^2}\mathbf{x}_\sigma(s)_i
\end{aligned}
$$

Exercise 13.5

A *Bernoulli-logistic unit* is a stochastic neuron-like unit used in some ANNs (Section 9.7). Its input at time t is a feature vector $\mathbf{x}(S_t)$; its output, A_t, is a random variable having two values, 0 and 1, with $Pr\{A_t=1\}=P_t$ and $Pr\{A_t=0\}=1-P_t$ (the Bernoulli distribution). Let $h(s,0,\boldsymbol{\theta})$ and $h(s,1,\boldsymbol{\theta})$ be the preferences in state s for the unit's two actions given policy parameter $\boldsymbol{\theta}$. Assume that the difference between the action preferences is given by a weighted sum of the unit's input vector, that is, assume that $h(s,1,\boldsymbol{\theta})-h(s,0,\boldsymbol{\theta})=\boldsymbol{\theta}^\top\mathbf{x}(s)$, where $\boldsymbol{\theta}$ is the unit's weight vector.

- (a) Show that if the exponential soft-max distribution (13.2) is used to convert action preferences to policies, then $P_t=\pi(1|S_t,\boldsymbol{\theta}_t)=1/(1+\exp(-\boldsymbol{\theta}^\top\mathbf{x}(S_t)))$ (the logistic function).
- (b) What is the Monte-Carlo REINFORCE update of $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}_{t+1}$ upon receipt of return G_t?
- (c) Express the eligibility $\nabla\ln\pi(a|s,\boldsymbol{\theta})$ for a Bernoulli-logistic unit, in terms of a, $\mathbf{x}(s)$, and $\pi(a|s,\boldsymbol{\theta})$ by calculating the gradient.

Hint for part (c): Define $P=\pi(1|s,\boldsymbol{\theta})$ and compute the derivative of the logarithm, for each action, using the chain rule on P. Combine the two results into one expression that depends on a and P, and then use the chain rule again, this time on $\boldsymbol{\theta}^\top\mathbf{x}(S_t)$, noting that the derivative of the logistic function $f(x)=1/(1+e^{-x})$ is $f(x)(1-f(x))$.

My answer:

(a):

$$
\begin{aligned}
\pi(1|S_t,\boldsymbol{\theta}_t) &= \frac{e^{h(S_t,1,\boldsymbol{\theta})}}{\sum_b e^{h(S_t,b,\boldsymbol{\theta})}} \\
&= \frac{e^{h(S_t,1,\boldsymbol{\theta})}}{e^{h(S_t,0,\boldsymbol{\theta})}+e^{h(S_t,1,\boldsymbol{\theta})}} \\
&= \frac{e^{\boldsymbol{\theta}_t^\top\mathbf{x}(S_t)+h(S_t,0,\boldsymbol{\theta})}}{e^{h(S_t,0,\boldsymbol{\theta})}+e^{\boldsymbol{\theta}_t^\top\mathbf{x}(S_t)+h(S_t,0,\boldsymbol{\theta})}} \\
&= \frac{e^{\boldsymbol{\theta}_t^\top\mathbf{x}(S_t)}}{1+e^{\boldsymbol{\theta}_t^\top\mathbf{x}(S_t)}} \\
&= \frac{e^{\boldsymbol{\theta}_t^\top\mathbf{x}(S_t)}e^{-\boldsymbol{\theta}_t^\top\mathbf{x}(S_t)}}{e^{-\boldsymbol{\theta}_t^\top\mathbf{x}(S_t)}(1+e^{\boldsymbol{\theta}_t^\top\mathbf{x}(S_t)})} \\
&= \frac{1}{1+e^{-\boldsymbol{\theta}_t^\top\mathbf{x}(S_t)}}
\end{aligned}
$$

(b):

From the box on page 328:

$$
\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}_t+\alpha\gamma^t G_t\nabla\ln\pi(A_t|S_t,\boldsymbol{\theta}_t)
$$

(c):

Let

$$
\frac{\delta}{\delta\boldsymbol{\theta}_i}
$$

denote the i:th element of ∇. We have

$$
\begin{aligned}
\frac{\delta \ln \pi(1|s,\boldsymbol{\theta})}{\delta \boldsymbol{\theta}_i} &= \frac{\delta \ln P}{\delta \boldsymbol{\theta}_i} \\
&= \frac{\delta \ln P}{\delta P}\frac{\delta P}{\delta \boldsymbol{\theta}_i} \\
&= \frac{1}{P}\frac{\delta P}{\delta \boldsymbol{\theta}^\top \mathbf{x}(s)}\frac{\delta \boldsymbol{\theta}^\top \mathbf{x}(s)}{\delta \boldsymbol{\theta}_i} \\
&= \frac{1}{P}P(1-P)\mathbf{x}(s)_i \\
&= (1-P)\mathbf{x}(s)_i
\end{aligned}
$$

$$
\begin{aligned}
\frac{\delta \ln \pi(0|s,\boldsymbol{\theta})}{\delta \boldsymbol{\theta}_i} &= \frac{\delta \ln(1-P)}{\delta \boldsymbol{\theta}_i} \\
&= \frac{\delta \ln(1-P)}{\delta(1-P)}\frac{\delta(1-P)}{\delta P}\frac{\delta P}{\delta \boldsymbol{\theta}_i} \\
&= \frac{1}{1-P}(-1)P(1-P)\mathbf{x}(s)_i \\
&= -P\mathbf{x}(s)_i
\end{aligned}
$$

Combining these expressions as a function of a (where $a = 0$ or $a = 1$) yields:

$$
\begin{aligned}
\frac{\delta \ln \pi(a|s,\boldsymbol{\theta})}{\delta \boldsymbol{\theta}_i} &= \big(a - P\big)\mathbf{x}(s)_i = \big(a - \pi(1|s,\boldsymbol{\theta})\big)\mathbf{x}(s)_i \\
\implies \nabla \ln \pi(a|s,\boldsymbol{\theta}) &= \big(a - \pi(1|s,\boldsymbol{\theta})\big)\mathbf{x}(s)
\end{aligned}
$$