

Chapter 13 Solutions

Reinforcement Learning: An Introduction

Alireza Azimi

2026-05-24

Table of contents

1 Exercise 13.1	1
2 Exercise 13.2	2
3 Exercise 13.3	3
4 Exercise 13.4	4
5 Exercise 13.5	4

1 Exercise 13.1

Consider the probability of right as p and left as $1-p$. The states are labeled 1,2,3 from left to right:

$$\begin{aligned}v_1 &= p(-1 + v_2) + (1 - p)(-1 + v_1) \\v_2 &= p(-1 + v_1) + (1 - p)(-1 + v_3) \\v_3 &= p(-1 + 0) + (1 - p)(-1 + v_2) \\ \Rightarrow pv_1 &= pv_2 - 1 \\v_2 &= pv_1 + (1 - p)v_3 - 1 \\v_3 &= (1 - p)v_2 - 1 \\ \Rightarrow p(2 - p)v_1 + (2 - p) &= pv_1 + p - 2 \\ \Rightarrow (p - p^2)v_1 &= 2(2 - p) \\ \Rightarrow v_1 &= \frac{2(2 - p)}{p - p^2}\end{aligned}$$

Therefore,

$$\begin{aligned}
p^* &= \operatorname{argmax}_p \left(\frac{2(2-p)}{p-p^2} \right) \\
\frac{d}{dp} \left(\frac{2(2-p)}{p-p^2} \right) &= 0 \\
\Rightarrow \frac{-2p^2 + 8p - 4}{(p-p^2)^2} &= 0 \\
\Rightarrow p^* &= \frac{4 \pm 2\sqrt{2}}{2} = 2 \pm \sqrt{2} \\
\stackrel{p^* < 1}{\longrightarrow} p^* &= 2 - \sqrt{2} \approx 0.59
\end{aligned}$$

2 Exercise 13.2

$$\begin{aligned}
\eta(s) &= h(s) + \gamma \sum_{\bar{s}} \eta(\bar{s}) \sum_a \pi(a|\bar{s}) p(s|\bar{s}, a) \\
\mu(s) &= \frac{\eta(s)}{\sum_{s'} \eta(s')}
\end{aligned}$$

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

$$\begin{aligned}
\nabla J(\theta) &= \nabla v_\pi(s_0) \\
&= \sum_s \left(\sum_{k=0}^{\infty} Pr(s \rightarrow x, k, \pi) \gamma^k \right) \sum_a \nabla \pi(a|s) q_\pi(s, a) \\
&= \sum_s \eta(s) \sum_a \nabla \pi(a|s) q_\pi(s, a) \\
&= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_{s'} \eta(s')} \sum_a \nabla \pi(a|s) q_\pi(s, a) \\
&\propto \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_\pi(s, a) \\
&= \mathbb{E}_\pi \left[\gamma^t \sum_a q_\pi(S_t, a) \nabla \pi(a|S_t, \theta) \right] \quad \text{Expectation under policy } \pi \text{ with termination } \gamma \\
&= \mathbb{E}_\pi \left[\gamma^t \sum_a \pi(a|S_t, \theta) q_\pi(S_t, a) \frac{\nabla \pi(a|S_t, \theta)}{\pi(a|S_t, \theta)} \right] \\
&= \mathbb{E}_\pi \left[\gamma^t q_\pi(S_t, A_t) \frac{\nabla \pi(A_t|S_t, \theta)}{\pi(A_t|S_t, \theta)} \right] \\
&= \mathbb{E}_\pi \left[\gamma^t G_t \frac{\nabla \pi(A_t|S_t, \theta)}{\pi(A_t|S_t, \theta)} \right] \\
&\Rightarrow \theta_{t+1} = \theta_t + \alpha \gamma^t G_t \frac{\nabla \pi(A_t|S_t, \theta_t)}{\pi(A_t|S_t, \theta_t)}
\end{aligned}$$

3 Exercise 13.3

Considering $\pi(a|s, \theta) = \frac{e^{h(s, a, \theta)}}{\sum_b e^{h(s, b, \theta)}}$ and $h(s, a, \theta) = \theta^T \mathbf{x}(s, a)$ or $h(s, b, \theta) = \theta^T \mathbf{x}(s, b)$:

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

4 Exercise 13.4

Note: $\nabla_{\theta_\mu} \mu(s, \theta) = \mathbf{x}_\mu(s)$ and $\nabla_{\theta_\sigma} \sigma(s, \theta) = \mathbf{x}_\sigma(s)$.

$$\begin{aligned}\ln \pi(a|s, \theta) &= -\frac{(a - \mu(s, \theta))^2}{2\sigma^2(s, \theta)} - \ln \sigma(s, \theta) - \ln \sqrt{2\pi} \\ \nabla_{\theta_\mu} \ln \pi(a|s, \theta_\mu) &= \nabla_{\theta_\mu} \left(-\frac{(a - \mu(s, \theta))^2}{2\sigma^2(s, \theta)} - \ln \sigma(s, \theta) - \ln \sqrt{2\pi} \right) \\ &= \frac{2(a - \mu(s, \theta))\mathbf{x}_\mu(s)}{2\sigma^2(s, \theta)} = \frac{(a - \mu(s, \theta))\mathbf{x}_\mu(s)}{\sigma^2(s, \theta)}\end{aligned}$$

$$\begin{aligned}\nabla_{\theta_\sigma} \ln \pi(a|s, \theta_\sigma) &= \nabla_{\theta_\sigma} \left(-\frac{(a - \mu(s, \theta))^2}{2\sigma^2(s, \theta)} - \ln \sigma(s, \theta) - \ln \sqrt{2\pi} \right) \\ &= \frac{(a - \mu(s, \theta))^2 \mathbf{x}_\sigma(s) \sigma(s, \theta)}{\sigma^3(s, \theta)} - \mathbf{x}_\sigma(s) \\ &= \frac{(a - \mu(s, \theta))^2 \mathbf{x}_\sigma(s)}{\sigma^2(s, \theta)} - \mathbf{x}_\sigma(s) \\ &= \left(\frac{(a - \mu(s, \theta))^2}{\sigma^2(s, \theta)} - 1 \right) \mathbf{x}_\sigma(s)\end{aligned}$$

5 Exercise 13.5

a

Note that: $h(s, 1, \theta) = \theta^T \mathbf{x}(s) + h(s, 0, \theta)$.

$$\begin{aligned}P_t &= \pi(1|S_t, \theta_t) = \frac{e^{h(S_t, 1, \theta)}}{e^{h(S_t, 1, \theta)} + e^{h(S_t, 0, \theta)}} \\ &= \frac{e^{\theta^T \mathbf{x}(S_t) + h(S_t, 0, \theta)}}{e^{\theta^T \mathbf{x}(S_t) + h(S_t, 0, \theta)} + e^{h(S_t, 0, \theta)}} \\ &= \frac{1}{1 + \frac{e^{h(S_t, 0, \theta)}}{e^{\theta^T \mathbf{x}(S_t) + h(S_t, 0, \theta)}}} \\ &= \frac{1}{1 + e^{-\theta^T \mathbf{x}(S_t)}}\end{aligned}$$

b

$$\theta_{t+1} = \theta_t + \alpha \gamma^t G_t \nabla \ln \pi(a|S_t, \theta_t)$$

c

$$\begin{aligned}P &= \pi(1|s, \theta) = \frac{1}{1 + e^{-\theta^T \mathbf{x}(s)}} \\ \Rightarrow \nabla P &= \mathbf{x}(s)P(1 - P) \\ \nabla \ln(P) &= \frac{\nabla P}{P} = \mathbf{x}(s)(1 - P) \\ \nabla \ln(1 - P) &= \frac{-\nabla P}{1 - P} = \mathbf{x}(s)P \\ \Rightarrow \nabla \ln(\pi(a|s, \theta)) &= \frac{(a - P)\nabla P}{P(1 - P)} \\ &= \frac{(a - P)\mathbf{x}(s)P(1 - P)}{P(1 - P)} \\ &= (a - P)\mathbf{x}(s) = (a - \pi(1|s, \theta))\mathbf{x}(s)\end{aligned}$$