1 Policy Gradient Methods

1.1 Exercise 13.1

Q

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.

A

Define $p = \mathbb{P}(\texttt{right})$, so $\mathbb{P}(\texttt{left}) = 1 - p$. Then, labelling the states 1-3 from right to left (value of terminal state is set to 0), the Bellman equations reduce to

$$\begin{aligned} v_\pi(1) &= pv_\pi(2) - 1 \\ v_\pi(2) &= pv_\pi(1) - 1 + (1-p)v_\pi(3) \\ v_\pi(3) &= (1-p)v_\pi(2) - 1. \end{aligned}$$

Setting $f(p)$ for the value of the initial state and solving this system gives

$$f(p) = \frac{p^2 - 2p + 2}{p(1-p)}$$

which attains its maximum at $p = \sqrt{2}(\sqrt{2} - 1)$. (Note that f is defined only on the open interval $(0, 1)$, the performance becomes infinitely bad as $p \to 0, 1$.)

1.2 *Exercise 13.2

Q

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.

A

- Generalisation the recursion equation that governs expected time in each state:

$$\begin{aligned} \eta(s) &= h(s) + \gamma \sum_{\bar{s}} \eta(\bar{s}) \sum_a \pi(a|\bar{s}) p(s|\bar{s}, a) \\ &= h(s) + \gamma \sum_{\bar{s},a} \pi(a|\bar{s}) p(s|\bar{s}, a) + \gamma^2 \sum_{\bar{s},a} \pi(a|\bar{s}) p(s|\bar{s}, a) \sum_{x,a'} \pi(a'|x) p(\bar{s}|x, a) + \cdots \end{aligned}$$

  This just changes the solution for $\eta(s)$, we still have $\mu(s) = \frac{\eta(s)}{\sum_{s'} \eta(s')}$.

- The generalisation of the proof of the policy gradient theorem comes with the use of the Bellman equation unfolding for the value function. We therefore arrive at the following gradient:

$$\nabla_{\boldsymbol{\theta}} v_\pi(s) = \sum_{x \in \mathcal{S}} \sum_{k=0}^{\infty} \mathbb{P}(s \to x, k, \pi) \gamma^k \sum_a \nabla_{\boldsymbol{\theta}} \pi(a|x) q_\pi(x, a),$$

  and the theorem follows as before.

- To full incorporate discounting, we need to view it as a form of termination. The policy gradient theorem becomes

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_\pi[\gamma_t \sum_a q_\pi(S_t, a) \nabla_{\boldsymbol{\theta}} \pi(a|S_t, \boldsymbol{\theta})].$$

  The factor of γ^t then follows through when we apply SGD. (It's possible to do some rearranging to prove this relation, but it is not done in the book – a little unclear!)

1.3 Exercise 13.3

Q

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

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

using the definitions and elementary calculus.

A

Have softmax policy

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

with linear action preferences

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

The following is then clear:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \log(\pi) &= \boldsymbol{x}(s,a) - \frac{\sum_b \boldsymbol{x}(s,b)\exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))}{\sum_b \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))} \\ &= \boldsymbol{x}(s,a) - \sum_b \boldsymbol{x}(s,b)\pi(b|s,\boldsymbol{\theta}).\end{aligned}$$

1.4 Exercise 13.4

Q

Show that for the gaussian policy parameterization (13.19) the eligibility vector has the following two parts:

$$\begin{aligned}\nabla \log \pi(a|s,\boldsymbol{\theta}_\mu) &= \frac{\nabla \pi(a|s,\boldsymbol{\theta}_\mu)}{\pi(a|s,\boldsymbol{\theta})} = \frac{1}{\sigma(s,\boldsymbol{\theta})^2}(a-\mu(s,\boldsymbol{\theta}))\boldsymbol{x}_\mu(s), \text{ and} \\ \nabla \log \pi(a|s,\boldsymbol{\theta}_\sigma) &= \frac{\nabla \pi(a|s,\boldsymbol{\theta}_\sigma)}{\pi(a|s,\boldsymbol{\theta})} = \left(\frac{(a-\mu(s,\boldsymbol{\theta}))^2}{\sigma(s,\boldsymbol{\theta})^2} - 1\right)\boldsymbol{x}_\sigma(s)\end{aligned}$$

A

Gaussian policy

$$\pi(a|s,\boldsymbol{\theta}) = \frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2\pi}} \exp\left(-\frac{(a-\mu(s,\boldsymbol{\theta}))^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)$$

with the models $\mu(s,\boldsymbol{\theta}_\mu) = \boldsymbol{\theta}_\mu^\top \boldsymbol{x}_\mu(s)$ and $\sigma(s,\boldsymbol{\theta}_\sigma) = \exp(\boldsymbol{\theta}_\sigma^\top \boldsymbol{x}_\sigma(s))$. First,

$$\log \pi(a|s,\boldsymbol{\theta}) = -\log\sqrt{2\pi} - \log\sigma - \frac{(a-\mu)^2}{2\sigma^2}$$

so we have

$$\nabla_{\boldsymbol{\theta}_\mu} \log \pi(a|s,\theta) = \frac{a-\mu}{\sigma^2}\nabla_{\boldsymbol{\theta}_\mu}\mu(s,\boldsymbol{\theta}_\mu) = \frac{a-\mu}{\sigma^2}\boldsymbol{x}_\mu(s)$$

and

$$\nabla_{\boldsymbol{\theta}_\sigma} \log \pi(a|s,\boldsymbol{\theta}) = -\frac{\nabla_{\boldsymbol{\theta}_\sigma}\sigma}{\sigma} + \frac{(a-\mu)^2}{\sigma^2}\nabla_{\boldsymbol{\theta}_\sigma}\sigma = \left(\frac{(a-\mu)^2}{\sigma^2} - 1\right)\boldsymbol{x}_\sigma(s)$$

because $\nabla_{\boldsymbol{\theta}_\sigma}\sigma = \boldsymbol{x}_\sigma(s)\sigma$.

1.5 Exercise 13.5

Q

A *Bernoulli-logistic unit* is a stochastic neuron-like unit used in some ANNs (Section 9.6). Its input at time t is a feature vector $\boldsymbol{x}(S_t)$; 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 $h(s, 1, \boldsymbol{\theta}) - h(s, 0, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{x}(s)$, 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 $P_t = \pi(1|S_t, \boldsymbol{\theta}_t) = 1/(1 + \exp(-\boldsymbol{\theta}_t^\top \boldsymbol{x}(S_t)))$(the logistic function).

(b) What is the Monte-Carlo REINFORCE update of $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}_t$ upon receipt of return G_t?

(c) Express the eligibility $\nabla \log \pi(a|s, \boldsymbol{\theta})$ for a Bernoulli-logistic unit, in terms of a, $\boldsymbol{x}(s)$ and $\pi(a|s, \boldsymbol{\theta})$ by calculating the gradient.

Hint: separately for each action compute the derivative of the logarithm first with respect to $P_t = \pi(1|S_t, \boldsymbol{\theta}_t)$, combine the two results into one expression that depends on a and P_t, and then use the chain rule, noting that the derivative of the logistic function $f(x)$ is $f(x)(1 - f(x))$.

A

a) $\pi(1|S_t, \boldsymbol{\theta}_t) = \frac{e^{h(s,1,\boldsymbol{\theta}_t)}}{e^{h(s,1,\boldsymbol{\theta}_t)} + e^{h(s,0,\boldsymbol{\theta}_t)}} = \frac{1}{1+e^{-\boldsymbol{\theta}^\top \boldsymbol{x}(s)}}$

b) $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \gamma^t G_t \nabla_{\boldsymbol{\theta}_t} \log \pi(a|S_t, \boldsymbol{\theta}_t)$

c) Write $\pi(a|S_t, \boldsymbol{\theta}_t) = g((-1)^a \boldsymbol{\theta}_t^\top \boldsymbol{x}(s))$ where $a \in \{0, 1\}$ and g is the sigmoid function $g(t) = 1/(1 + e^{-t})$. It's then quite easy to see that

$$\frac{\mathrm{d}}{\mathrm{d}t} \log g(t) = 1 - g(t)$$

which leads to

$$\nabla_{\boldsymbol{\theta}} \pi(a|s, \boldsymbol{\theta}) = (-1)^a \boldsymbol{x}(s)(1 - \pi(a|s, \boldsymbol{\theta})).$$