December 16, 2022

13 Exercises

13.1 Question

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.

Answer

By equation 13.2, probability selecting right action:
$\pi(a|s,\theta) = \frac{e^{h(s,a,\theta)}}{\sum_b e^{h(s,b,\theta)}}$
$\pi(right|s,\theta) = \frac{e^{h(s,right,\theta)}}{e^{h(s,right,\theta)}+e^{h(s,left,\theta)}}$

Incorporating 13.3 and feature vectors:
$\pi(right|s,\theta) = \frac{e^{\theta^T[1,0]}}{e^{\theta^T[1,0]}+e^{\theta^T[0,1]}} = \frac{e^{\theta_1}}{e^{\theta_1}+e^{\theta_2}}$

From example 3.1 we know that probability of selecting right action is 0.59 then one probable parameter vector can be $\theta = [-0.53, -0.89]$:

13.2 Question

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.

Answer

The text states in the boxed page 199: If there is discounting ($\gamma < 1$) it should be treated as a form of termination, which can be done simply by including a factor of γ in the second term of 9.2.

Equation 9.2 with discounting:
$\eta(s) = h(s) + \gamma \sum_{\bar{s}} \eta(\bar{s}) \sum_a \pi(a|\bar{s}) p(s|\bar{s}, a)$

Proof of policy gradient theorem changes in the way $\nabla v_\pi(s_0)$ is expanded:
$\nabla J(\theta) = \nabla V_\pi(s_0)$
$\nabla J(\theta) = \sum_s (\sum_{k=0}^{\infty} \gamma^k \Pr(s_0 \to s, k, \pi)) \sum_a \nabla \pi(a|s) q_\pi(s, a)$

I cannot show how γ^k is handled from here on. At the end the REINFORCE update 13.8 becomes:
$\theta_{t+1} = \theta t + \alpha \gamma^t G_t \frac{\nabla \pi(A_t|S_t, \theta_t)}{\pi(A_t|S_t, \theta_t)}$

Alternative solution may be found in UCL leacture series.

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) &= \mathbb{E}_{\pi}[G(\tau) \sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)] \\
&= \mathbb{E}_{\pi}[\sum_{t=0}^{T} G(\tau) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)] \\
&= \mathbb{E}_{\pi}[\sum_{t=0}^{T} \left(\sum_{k=0}^{T} \gamma^k R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)] \\
&= \mathbb{E}_{\pi}[\sum_{t=0}^{T} \left(\sum_{k=\boldsymbol{t}}^{T} \gamma^k R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)] \\
&= \mathbb{E}_{\pi}[\sum_{t=0}^{T} \left(\gamma^t \sum_{k=t}^{T} \gamma^{k-t} R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)] \\
&= \mathbb{E}_{\pi}[\sum_{t=0}^{T} \left(\gamma^t G_t \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)] \qquad = \mathbb{E}_{\pi}[\sum_{t=0}^{T} \gamma^t q_{\pi}(S_t, A_t) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)]
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) &= \mathbb{E}_{\pi}[G(\tau) \sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)] \\
&= \mathbb{E}_{\pi}[\sum_{t=0}^{T} G(\tau) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)] \\
&= \mathbb{E}_{\pi}[\sum_{t=0}^{T} \left(\sum_{k=0}^{T} \gamma^k R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)] \\
&= \mathbb{E}_{\pi}[\sum_{t=0}^{T} \left(\sum_{k=\boldsymbol{t}}^{T} \gamma^k R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)] \\
&= \mathbb{E}_{\pi}[\sum_{t=0}^{T} \left(\gamma^t \sum_{k=t}^{T} \gamma^{k-t} R_{k+1} \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)] \\
&= \mathbb{E}_{\pi}[\sum_{t=0}^{T} \left(\gamma^t G_t \right) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)] \qquad = \mathbb{E}_{\pi}[\sum_{t=0}^{T} \gamma^t q_{\pi}(S_t, A_t) \nabla_{\boldsymbol{\theta}} \log \pi(A_t|S_t)]
\end{aligned}
$$

Taken from UCL Lecture 9 by Hado van Hasselt.

13.3 Question

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:

$$\nabla \ln \pi(a|s,\theta) = x(s,a) - \textstyle\sum_b \pi(b|s,\theta)x(s,b) \tag{13.9}$$

using the definitions and elementary calculus.

Answer

Equation 13.2 is:
$\pi(a|s,\theta) = \frac{e^{h(s,a,\theta)}}{\sum_b e^{h(s,b,\theta)}}$

Equation 13.3 is:
$h(s,a,\theta) = \theta^T x(s,a)$

Let's start with 13.2 taking logarithm of both sides :
$\ln \pi(a|s,\theta) = \ln \frac{e^{h(s,a,\theta)}}{\sum_b e^{h(s,b,\theta)}}$
$\ln \pi(a|s,\theta) = \ln e^{h(s,a,\theta)} - \ln \sum_b e^{h(s,b,\theta)}$
$\ln \pi(a|s,\theta) = h(s,a,\theta) - \ln \sum_b e^{h(s,b,\theta)}$

Now take derivative both sides :
$\nabla \ln \pi(a|s,\theta) = \nabla h(s,a,\theta) - \nabla \ln \sum_b e^{h(s,b,\theta)}$
$\nabla \ln \pi(a|s,\theta) = \nabla h(s,a,\theta) - \frac{\nabla \sum_b e^{h(s,b,\theta)}}{\sum_b e^{h(s,b,\theta)}}$
$\nabla \ln \pi(a|s,\theta) = \nabla h(s,a,\theta) - \frac{\sum_b \nabla e^{h(s,b,\theta)}}{\sum_b e^{h(s,b,\theta)}}$
$\nabla \ln \pi(a|s,\theta) = \nabla h(s,a,\theta) - \frac{\sum_b \nabla h(s,b,\theta) e^{h(s,b,\theta)}}{\sum_b e^{h(s,b,\theta)}}$
$\nabla \ln \pi(a|s,\theta) = \nabla h(s,a,\theta) - \sum_b \nabla h(s,b,\theta)\pi(b|s,\theta)$ from equation 13.2
$\nabla \ln \pi(a|s,\theta) = x(s,a) - \sum_b x(s,b)\pi(b|s,\theta)$ from equation 13.3

13.4 Question

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

Answer

Relevant equations are:

$\mu(s,\theta) = \theta_\mu x_\mu(s)$
$\sigma(s,\theta) = e^{\theta_\sigma^T x_\sigma(s)}$

Taking logarithm of the policy function:
$\ln \pi(a|s,\theta) = \ln \frac{e^{-\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}}}{\sigma(s,\theta)\sqrt{2\pi}}$
$\ln \pi(a|s,\theta) = \ln e^{-\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}} - \ln \sigma(s,\theta)\sqrt{2\pi}$
$\ln \pi(a|s,\theta) = -\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2} - \ln \sigma(s,\theta) - \ln \sqrt{2\pi}$
$\ln \pi(a|s,\theta) = -\frac{(a-\mu(s,\theta_\mu))^2}{2\sigma(s,\theta_\sigma)^2} - \ln \sigma(s,\theta_\sigma) - \ln \sqrt{2\pi}$ (1)

Using (1) and taking derivative wrt. θ_μ
$\nabla\theta_\mu \ln \pi(a|s,\theta_\mu) = -\nabla \frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2} = -\frac{1}{2\sigma(s,\theta)^2}\nabla(a-\mu(s,\theta))^2\ \nabla\theta_\mu \ln \pi(a|s,\theta_\mu) =$
$-\frac{1}{2\sigma(s,\theta)^2}2(a-\mu(s,\theta))(-x_\mu(s))\ \nabla\theta_\mu \ln \pi(a|s,\theta_\mu) = \frac{1}{\sigma(s,\theta)^2}(a-\mu(s,\theta))x_\mu(s)$

Using (1) and taking derivative wrt. θ_σ
$\nabla\theta_\sigma \ln \pi(a|s,\theta_\mu) = \nabla - \frac{(a-\mu(s,\theta_\mu))^2}{2\sigma(s,\theta)^2} - \nabla \ln \sigma(s,\theta)$
$\nabla\theta_\sigma \ln \pi(a|s,\theta_\mu) = -(a-\mu(s,\theta_\mu))^2 \nabla \frac{1}{2\sigma(s,\theta)^2} - \nabla \ln \sigma(s,\theta)$
$\nabla\theta_\sigma \ln \pi(a|s,\theta_\mu) = -(a-\mu(s,\theta_\mu))^2 \frac{-4\sigma(s,\theta)\sigma(s,\theta)x_\sigma(s)}{4\sigma(s,\theta)^4} - x_\sigma(s)$
$\nabla\theta_\sigma \ln \pi(a|s,\theta_\mu) = (a-\mu(s,\theta_\mu))^2 \frac{x_\sigma(s)}{\sigma(s,\theta)^2} - x_\sigma(s)$
$\nabla\theta_\sigma \ln \pi(a|s,\theta_\mu) = (\frac{(a-\mu(s,\theta_\mu))^2}{\sigma(s,\theta)^2} - 1)x_\sigma(s)$

13.5 Question

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 $\mathbf{x}(S_t)$; 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 $P_t = \pi(1|S_t, \boldsymbol{\theta}_t) = 1/(1 + \exp(-\boldsymbol{\theta}_t^\top \mathbf{x}(S_t)))$ (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: separately for each action compute the derivative of the logarithm first with respect to $P_t = \pi(a|s, \boldsymbol{\theta}_t)$, 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 $f(x)(1 - f(x))$. □

Answer

a

$$\pi(s, 1, \theta) = \frac{e^{h(s,1,\theta)}}{e^{h(s,1,\theta)} + e^{h(s,0,\theta)}}$$
$$\pi(s, 1, \theta) = \frac{e^{h(s,1,\theta)}}{e^{h(s,1,\theta)} + e^{h(s,1,\theta) - \theta x(s)}}$$
$$\pi(s, 1, \theta) = \frac{1}{1 + e^{-\theta x(s)}}$$

b

$$\theta_{t+1} = \theta_t + \alpha G_t \nabla \ln \pi(A_t|S_t, \theta_t) \qquad \text{From equation 13.8}$$

c

I am not sure about the result. I followed the steps hoping to reach a reasonable result. The result I found seems be reasonable but it does not include the x(s) term.

Separately for each action computing the derivative of the logarithm with respect to P

$\pi(1|S_t, \theta_t) = P_t$
$\ln \pi(1|S_t, \theta_t) = \ln P_t$
$\nabla \ln \pi(1|S_t, \theta_t) = \frac{P_t^{'}}{P_t} = \frac{P_t(1-P_t)}{P_t} = 1 - P_t$

$\pi(0|S_t, \theta_t) = 1 - P_t$
$\ln \pi(0|S_t, \theta_t) = \ln(1 - P_t)$
$\nabla \ln \pi(0|S_t, \theta_t) = \frac{P_t^{'}}{1-P_t} = \frac{P_t(1-P_t)}{1-P_t} = P_t$

Combining both results into one expressing that depends on a and P
$\nabla \ln \pi(a|S_t, \theta_t) = a\nabla \ln \pi(1|S_t, \theta_t) + (1 - a)\nabla \ln \pi(0|S_t, \theta_t)$
$\nabla \ln \pi(a|S_t, \theta_t) = a(1 - P_t) + (1 - a)P_t$

Now replacing with P with π and rearranging:
$\nabla \ln \pi(a|S_t, \theta_t) = a\pi(0|S_t, \theta_t) + (1 - a)\pi(1|S_t, \theta_t)$
$\nabla \ln \pi(a|S_t, \theta_t) = a(1 - \pi(1|S_t, \theta_t)) + (1 - a)(1 - \pi(0|S_t, \theta_t))$

The first part is non-zero only when a is 1. The second part is only non-zero only when a is 0. Thus we can replace selected actions with a:
$\nabla \ln \pi(a|S_t, \theta_t) = a(1 - \pi(a|S_t, \theta_t)) + (1 - a)(1 - \pi(a|S_t, \theta_t))$
$\nabla \ln \pi(a|S_t, \theta_t) = 1 - \pi(a|S_t, \theta_t)$