CMPUT 609 Assignment 6

Shaurya Seth

March 14, 2023

Question 1

Not attempted.

Question 2

Not attempted.

Question 3

Using (13.2) and (13.3) we get:

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

Now the eligibility vector can be written as:

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

Question 4

The Gaussian policy parameterization is given by:

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

where $\mu(s,\boldsymbol{\theta}) = \boldsymbol{\theta}_\mu^\top \mathbf{x}_\mu(s)$ and $\sigma(s,\boldsymbol{\theta}) = \exp(\boldsymbol{\theta}_\sigma^\top \mathbf{x}_\sigma(s))$.

Taking log on both sides:

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

The gradient with respect to $\boldsymbol{\theta}_\mu$ becomes:

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

The gradient with respect to $\boldsymbol{\theta}_\sigma$ becomes:

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

Question 5

(a) Using (13.2) with a=1:

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

(b) The Monte Carlo REINFORCE update is:

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

(c) The eligibility vector for the Bernoulli-logistic unit can be given by:

$$\begin{aligned}
\nabla \ln \pi(a|s,\boldsymbol{\theta}_t) &= \nabla \ln \left[\frac{1}{1 + e^{(-1)^a \boldsymbol{\theta}_t^\top \mathbf{x}(s)}} \right] \qquad \text{where } a \in \{0,1\} \\
&= \nabla \left[\ln(1) - \ln(1 + e^{(-1)^a \boldsymbol{\theta}_t^\top \mathbf{x}(s)}) \right] \\
&= -\nabla \ln(1 + e^{(-1)^a \boldsymbol{\theta}_t^\top \mathbf{x}(s)}) \\
&= \frac{e^{(-1)^a \boldsymbol{\theta}_t^\top \mathbf{x}(s)}}{1 + e^{(-1)^a \boldsymbol{\theta}_t^\top \mathbf{x}(s)}} (-1)^a \mathbf{x}(s) \\
&= (1 - \pi(a|s,\boldsymbol{\theta}_t))(-1)^a \mathbf{x}(s)
\end{aligned}$$

Question 6

The policy gradient theorem states that the gradient of performance is proportional to the policy parameters which makes gradient ascent possible. The policy improvement theorem states that if a policy is changed using its action-value at any state, the new policy is equal to or better than the original policy and is what makes policy iteration work.

Both on these theorems are important for the main task of reinforcement learning which is finding a good policy. However, the two are not entirely analogous because the policy improvement theorem holds for the tabular case and is with respect to value-based methods while the policy gradient theorem is for the function-approximation setting and is in context of policy-based methods.

Question 7

The advantages of policy-based methods over value-based methods are:

- The approximate policy can approach a deterministic policy.
- The selection of actions with arbitrary probability is enabled.
- The policy may be a simpler function to approximate.
- Allows injecting prior knowledge about desired form of policy.
- Action probabilities change smoothly with learned parameters.
- Allow continuous action spaces with infinite number of actions.

Question 8

It is better to minimize the projected Bellman error because the Bellman error is not learnable. The projected Bellman error can be learned directly from the data distribution and will have a unique minimizer.