Artificial Neural Networks (Gerstner). Solutions for week 5

Policy gradient methods

Exercise 1. Single neuron as an actor[1]

Assume an agent with binary actions $Y \in \{0, 1\}$. Action $y = 1$ is taken with a probability $\pi(Y = 1|\vec{x}; \vec{w}) = g(\vec{w} \cdot \vec{x})$, where $\vec{w}$ are a set of weights and $\vec{x}$ is the input signal that contains the state information. The function g is monotonically increasing and limited by the bounds $0 \leq g \leq 1$.

For each action, the agent receives a reward $R(Y, \vec{x})$.

a. Calculate the gradient of the mean reward $\mathbb{E}[R] = \sum_{Y,\vec{x}} R(Y, \vec{x})\pi(Y|\vec{x}; \vec{w})P(\vec{x})$ with respect to the weight w_j.

   Hint: Insert the policy $\pi(Y = 1|\vec{x}; \vec{w}) = g(\sum_k w_k x_k)$ and $\pi(Y = 0|\vec{x}; \vec{w}) = 1 - g(\sum_k w_k x_k)$. Then take the gradient.

b. The rule derived in (a) is a batch rule. Can you transform this into an 'online rule'?

   Hint: Pay attention to the following question: what is the condition that we can simply 'drop the summation signs'?

Solution:

a. $\frac{\partial}{\partial w_j}\mathbb{E}[R] = \sum_{\vec{x}} P(\vec{x})[R(y = 1, \vec{x}) - R(y = 0, \vec{x})]g'(\vec{w} \cdot \vec{x})x_j$

b. If the online statistics matches the true statistics of the data in the batch, then we can drop the sum-signs. However, here this is not the case because the two outcomes $y = 1$ and $y = 0$ do not have equal probabilities. Therefore, the weight-factors in y need to be added. This can be done by the log-likelihood trick explained in class.

Exercise 2. Policy gradient for binary actions

a. Find an online policy gradient rule for the weights $\vec{w}$ for the same setup as in Exercise 1 by calculating the gradient of the log-likelihood $\log \pi(Y|\vec{x}; \vec{w})$ with respect to the weights.

   Hint: the policy π can be written as $\pi(Y|\vec{x}; \vec{w}) = (1 - \rho)^{1-Y}\rho^Y$ with $\rho = g(\vec{w} \cdot \vec{x})$.

b. Rewrite your update rule for weight w_j in the form

$$\Delta w_j = F(\vec{x}, \vec{w}, R)\,[Y - \mathbb{E}[Y]]\,x_j$$

   and give the expression for the function F.

   Hint: Take your result from part a, use $\mathbb{E}[y] = g(\vec{w} \cdot \vec{x})$ and pull out a factor $\frac{1}{g(1-g)}$.

Solution:

[1] Will be started in class.

a. Let's first calculate the derivative of $\log \pi(Y|\vec{x};\vec{w})$ with respect to w_j, using the hint:

$$\begin{aligned}
\frac{\partial}{\partial w_j}\log \pi(Y|\vec{x};\vec{w}) &= \frac{1}{\pi(Y|\vec{x};\vec{w})}\frac{\partial}{\partial w_j}\pi(Y|\vec{x};\vec{w}) \\
&= \frac{1}{(1-\rho)^{1-Y}\rho^Y}\frac{\partial}{\partial w_j}\left[(1-\rho)^{1-Y}\rho^Y\right] \\
&= \frac{1}{(1-\rho)^{1-Y}\rho^Y}\left[-(1-Y)(1-\rho)^{-Y}\rho^Y + Y(1-\rho)^{1-Y}\rho^{Y-1}\right]\frac{\partial}{\partial w_j}\rho \\
&= \left[-\frac{(1-Y)(1-\rho)^{-Y}}{(1-\rho)^{1-Y}} + \frac{Y\rho^{Y-1}}{\rho^Y}\right] g'(\vec{w}\cdot\vec{x})x_j \\
&= \left[-\frac{(1-Y)}{(1-\rho)} + \frac{Y}{\rho}\right] g'(\vec{w}\cdot\vec{x})x_j.
\end{aligned}$$

Now let's consider the term $\frac{\partial}{\partial w_j}\mathbb{E}[R]$ again. We can write

$$\begin{aligned}
\frac{\partial}{\partial w_j}\mathbb{E}[R] &= \sum_{Y,\vec{x}} R(Y,\vec{x})\frac{\partial}{\partial w_j}\pi(Y|\vec{x};\vec{w})P(\vec{x}) \\
&= \sum_{Y,\vec{x}} R(Y,\vec{x})\pi(Y|\vec{x};\vec{w}) \underbrace{\frac{1}{\pi(Y|\vec{x};\vec{w})}\frac{\partial}{\partial w_j}\pi(Y|\vec{x};\vec{w})}_{\frac{\partial}{\partial w_j}\log \pi(Y|\vec{x};\vec{w})} P(\vec{x}) \\
&= \mathbb{E}\left[R\frac{\partial}{\partial w_j}(\log \pi)\right],
\end{aligned}$$

where we multiplied by $\pi(\cdot)/\pi(\cdot) = 1$ and identified the derivative of the log. This suggest an online rule with an update term:

$$\Delta w_j = R\frac{\partial}{\partial w_j}\log \pi(Y|\vec{x};\vec{w}) = R\left[\frac{Y}{\rho} - \frac{(1-Y)}{(1-\rho)}\right] g'(\vec{w}\cdot\vec{x})x_j. \tag{1}$$

b. Equation 1 can be simplified as

$$\Delta w_j = R\left[\frac{Y-\rho}{\rho(1-\rho)}\right] g'(\vec{w}\cdot\vec{x})x_j = \frac{Rg'}{g(1-g)}\left[Y - \mathbb{E}[Y]\right]x_j, \tag{2}$$

which has the form of $\Delta w_j = F(\vec{x},\vec{w},R)\left[Y-\mathbb{E}[Y]\right]x_j$ with

$$F(\vec{x},\vec{w},R) = \frac{Rg'(\vec{w}\cdot\vec{x})}{g(\vec{w}\cdot\vec{x})\left(1-g(\vec{w}\cdot\vec{x})\right)}.$$

Exercise 3. Policy gradient

a. **Other parameterizations of Exercise 2**: Consider your solution to Exercise 2. What happens to the policy gradient rule if the likelihood ρ of action 1 is parameterized not by the weights $\vec{w}$ but by other parameters: $\rho = \rho(\theta)$? Derive a learning rule for θ.

b. **Generalization to the natural exponential family**: The natural exponential family is a family of probability distributions that is widely used in statistics because of its favorable properties. These distributions can be written in the form

$$p(Y) = h(Y)\exp\left(\theta Y - A(\theta)\right).$$

This family includes many of the standard probability distributions. The Bernoulli, the Poisson and the Gaussian distribution are all member of this family. A nice property of these distributions is that the mean can easily be calculated from the function $A(\theta)$:

$$\mathbb{E}[Y] = A'(\theta) := \frac{dA}{d\theta}(\theta)\,.$$

Assume that the policy $\pi(Y|\vec{x};\theta)$ is an element of the natural exponential family. Show that the online rule for the policy gradient has the shape:

$$\Delta\theta = R(Y - \mathbb{E}[Y])\,.$$

Can you give an intuitive interpretation of this learning rule?

c. **The Bernoulli distribution**: Apply your result from (b) to the case of Exercise 2.

Solution:

a. **Other parameterizations**: Replacing $\vec{w}\cdot\vec{x}$ by θ, we can follow the same steps as in Exercise 2. The only difference comes in the expression of $\frac{d\rho}{d\theta}$, for which we don't have an explicit expression anymore. The learning rule is:

$$\Delta\theta = R\left[\frac{Y}{\rho} - \frac{(1-Y)}{(1-\rho)}\right]\rho'(\theta). \tag{3}$$

b. **Generalization to the natural exponential family**: Let's calculate $\frac{\partial}{\partial\theta}\log p(Y)$:

$$\begin{aligned}\frac{\partial}{\partial\theta}\log p(Y) &= \frac{\partial}{\partial\theta}\log\left[h(Y)\exp\left(\theta Y - A(\theta)\right)\right]\\ &= \frac{1}{h(Y)\exp\left(\theta Y - A(\theta)\right)}\cdot h(Y)\exp\left(\theta Y - A(\theta)\right)\cdot\big(Y - A'(\theta)\big)\\ &= Y - A'(\theta) = (Y - \mathbb{E}[Y]).\end{aligned}$$

Therefore:

$$\Delta\theta = R\frac{\partial}{\partial\theta}\log P(y) = R\left(Y - \mathbb{E}[Y]\right)\,.$$

This learning rule will look for correlation between the reward and the deviations of Y from its expectation value. If R is systematically positive when Y is higher than its expectation value, θ will increase, leading to higher probabilities of higher Y. Inversely, if R is systematically negative when Y is higher than its expectation value, theta will decrease and the probability of lower Y will decrease.

c. For the Bernoulli distribution with $Y \in \{0, 1\}$ and $p(Y = 1) = \rho$, we have

$$\begin{aligned}p(Y) &= \rho^Y(1-\rho)^{1-Y} = \exp\left(Y\log\frac{\rho}{1-\rho} - \log\frac{1}{1-\rho}\right)\\ &= h(Y)\exp\left(\theta Y - A(\theta)\right)\,,\end{aligned}$$

where

$$\begin{aligned}h(Y) &= 1\\ \theta &= \log\frac{\rho}{1-\rho} \Leftrightarrow \rho = \frac{1}{1+e^{-\theta}}\\ A(\theta) &= \log\frac{1}{1-\rho} = \log\left(1+e^{\theta}\right)\,.\end{aligned}$$

From part (b), we know that $\Delta\theta = R\,(Y - \mathbb{E}[Y])$. To apply apply this update rule to the case of Exercise 2, we first use the fact that $\rho = g(\vec{w}\cdot\vec{x})$ and write

$$\theta = \log\frac{\rho}{1-\rho} = \log\frac{g(\vec{w}\cdot\vec{x})}{1-g(\vec{w}\cdot\vec{x})}.$$

We can use this and write

$$\Delta w_j = \frac{\partial}{\partial w_j}\mathbb{E}[R] = \frac{\partial}{\partial\theta}\mathbb{E}[R]\;\frac{\partial\theta}{\partial w_j} = \Delta\theta\,\left(\frac{\partial}{\partial w_j}\log\frac{g(\vec{w}\cdot\vec{x})}{1-g(\vec{w}\cdot\vec{x})}\right),$$

where

$$\frac{\partial}{\partial w_j}\log\frac{g(\vec{w}\cdot\vec{x})}{1-g(\vec{w}\cdot\vec{x})} = \left(\frac{g'}{g}+\frac{g'}{1-g}\right)x_j = \frac{g'}{g(1-g)}x_j.$$

Putting everything together, we have

$$\Delta w_j = R\,(Y-\mathbb{E}[Y])\,\frac{g'}{g(1-g)}x_j$$

which is the same as Δw_j in Equation 2.

Exercise 4. Subtracting the mean

You have two stochastic variables, x and y with means $\mathbb{E}[x]$ and $\mathbb{E}[y]$. Angles denote expectations. We are interested in the product $z = (x-b)(y-\mathbb{E}[y])$ with a fixed parameter b.

a. Show that $\mathbb{E}[z]$ is independent of the choice of the parameter b.

b. Show that $\mathbb{E}[z^2]$ is minimal if $b = \frac{\mathbb{E}[xf(y)]}{\mathbb{E}[f(y)]}$, where $f(y) = (y-\mathbb{E}[y])^2$.

   Hint: write $\mathbb{E}[z^2] = F(b)$ and set $\frac{dF}{db} = 0$.

c. What is the optimal b, if x and $f(y)$ are approximately independent?

d. Make the connection to policy gradient rules.

   Hint: take $x = r$ (reward) and y the action taken in state s. Compare with the policy gradient formula of the simple 1-neuron actor. What can you conclude for the best value of b? Consider different states s. Why should b depend on s?

Solution:

a.

$$\begin{aligned}
\mathbb{E}[z] &= \mathbb{E}[(x-b)(y-\mathbb{E}[y])]\\
&= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] - b\mathbb{E}[y] + b\mathbb{E}[y]\\
&= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]
\end{aligned}$$

b.

$$\begin{aligned}
F(b) &= \mathbb{E}\left[(x-b)^2 f(y)\right]\\
\Rightarrow 0 = \frac{d}{db}F(b) &= -2\mathbb{E}\left[(x-b)f(y)\right]\\
\Rightarrow 0 &= \mathbb{E}[xf(y)] - b\mathbb{E}[f(y)]\\
\Rightarrow b &= \frac{\mathbb{E}[xf(y)]}{\mathbb{E}[f(y)]}
\end{aligned}$$

c. If x and $f(y)$ are approximately independent, $\mathbb{E}[xf(y)] \approx \mathbb{E}[x]\mathbb{E}[f(y)]$ and we find $b \approx \mathbb{E}[x]$.

d. If we set $r = x$ and introduce states s as a further stochastic variable, we see that $y - \mathbb{E}[y]$ appears in the derivative of the log-policy (e.g. for a Gaussian policy $\frac{\partial}{\partial w} \log\left((1/\sqrt{2\pi})\exp(-(y-ws)^2/2)\right) = (y - ws)s$ with $ws = \mathbb{E}[y]$; see also next exercise), and thus $(r-b)(y-\mathbb{E}[y]) \propto (r-b)\frac{\partial}{\partial w}\log\pi(y|s;w) = \frac{\partial}{\partial w}R(y,s)$. Since r and y are now state dependent, the optimal baseline should also be state-dependent.

Exercise 5. Computer exercises: Environment 2 (part 1)[1]

Download the Jupyter notebook of the 2nd computer exercise and complete it until the end of Section 1.3.4 (Reinforce with Baseline).

[1] Start this exercise in the second exercise session of week 5.