Artificial Neural Networks (Gerstner). Solutions for week 13

Intrinsically motivated exploration

Exercise 1. Information-gain, surprise, and the number of observations

Consider an environment with a finite set of states $\mathcal{S}$ and a finite set of actions $\mathcal{A}$. At each time $t > 0$, we assume that the agent uses its past experiences (i.e., $s_0, a_0, \cdots, s_{t-1}, a_{t-1}, s_t$) and estimates the environment transition probabilities as

$$\hat{p}^{(t)}(s'|s,a) = \frac{T^{(t)}_{s,a,s'} + \epsilon}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon},$$

where $T^{(t)}_{s,a}$ is the number of times that the agent has taken action a in state s until time t, $T^{(t)}_{s,a,s'}$ is the number of times that taking action a in state s took the agent to state s', $|\mathcal{S}|$ is the total number of states, and $\epsilon > 0$ is a small constant to avoid division by zero.

Consider $s_t = s$ and $a_t = a$. In this exercise, we study the Information Gain (IG) of the transition $(s,a) \to s_{t+1}$ and its link to surprise and number of observations.

a. Given the transition $(s,a) \to s_{t+1}$ at $t+1$, what is the updated $\hat{p}^{(t+1)}(s'|s,a)$ for all $s' \in \mathcal{S}$? Write your answer as a function of $T^{(t)}_{s,a,s'}$, $T^{(t)}_{s,a}$, $|\mathcal{S}|$ and ϵ.

b. Show that

$$\hat{p}^{(t+1)}(s'|s,a) - \hat{p}^{(t)}(s'|s,a) = \frac{1}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1}\left(\delta_{s',s_{t+1}} - \hat{p}^{(t)}(s'|s,a)\right),$$

where δ is the Kronecker delta function.

c. One appraoch to define information gain is the L1 norm of the difference between $\hat{p}^{(t+1)}(.|s,a)$ and $\hat{p}^{(t)}(.|s,a)$:

$$\mathrm{IG}_{t+1} = \sum_{s' \in \mathcal{S}} \left|\hat{p}^{(t+1)}(s'|s,a) - \hat{p}^{(t)}(s'|s,a)\right|.$$

Find IG_{t+1} as a function of $T^{(t)}_{s,a}$, $|\mathcal{S}|$, ϵ, and $\hat{p}^{(t)}(s_{t+1}|s,a)$.

How does increasing the number of observation $T^{(t)}_{s,a}$ influence the information gain IG_{t+1}?

d. One of the many ways to define the surprise of the transition $(s,a) \to s_{t+1}$ is to use the notion of 'State Prediction Error' (see Modirshanechi et al. 2022 for its link to other definitions of surprise):

$$\mathrm{SPE}_{t+1} = 1 - \hat{p}^{(t)}(s_{t+1}|s,a).$$

Rewrite IG_{t+1} as a function of $T^{(t)}_{s,a}$, $|\mathcal{S}|$, ϵ, and SPE_{t+1}.

How does the information gain IG_{t+1} relate to the state prediction error SPE_{t+1}?

e. Assume that we know the true transition probabilities $p(.|s,a)$ and that $\lim_{t\to\infty} T^{(t)}_{s,a} = \infty$ (i.e., agents choose each action inifnitely many times). For a given next state $s_{t+1} = s'$, find the limits

$$\lim_{t\to\infty} \mathrm{SPE}_{t+1} \quad \text{and} \quad \lim_{t\to\infty} \mathrm{IG}_{t+1}.$$

What do these results imply about seeking SPE or IG as intrinsic rewards in the presence of stochasticity?

Which intrinsic reward is less prone to the noisy-TV problem?

Solution:

a. Using the definition of the estimated transition probabilities, we have

$$\hat{p}^{(t+1)}(s'|s,a) = \frac{T^{(t+1)}_{s,a,s'} + \epsilon}{T^{(t+1)}_{s,a} + |\mathcal{S}|\epsilon} = \frac{T^{(t)}_{s,a,s'} + \epsilon + \delta_{s',s_{t+1}}}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1}.$$

b. Using part a, we can write

$$\hat{p}^{(t+1)}(s'|s,a) - \hat{p}^{(t)}(s'|s,a) = \frac{T^{(t)}_{s,a,s'} + \epsilon + \delta_{s',s_{t+1}}}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1} - \frac{T^{(t)}_{s,a,s'} + \epsilon}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon}$$
$$= \frac{1}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1}\Big(T^{(t)}_{s,a,s'} + \epsilon + \delta_{s',s_{t+1}} - \Big(T^{(t)}_{s,a,s'} + \epsilon\Big) \underbrace{\frac{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon}}_{=1+\frac{1}{T^{(t)}_{s,a}+|\mathcal{S}|\epsilon}}\Big).$$

By simplifying the term in the parantheses and reusing the definition of $\hat{p}^{(t)}(s'|s,a)$, we have

$$\hat{p}^{(t+1)}(s'|s,a) - \hat{p}^{(t)}(s'|s,a) = \frac{1}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1}\Big(\delta_{s',s_{t+1}} - \hat{p}^{(t)}(s'|s,a)\Big).$$

c. We can use the result of part b and write

$$\mathrm{IG}_{t+1} = \frac{1}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1}\sum_{s'\in\mathcal{S}}\Big|\delta_{s',s_{t+1}} - \hat{p}^{(t)}(s'|s,a)\Big|$$
$$= \frac{1}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1}\Big(1 - \hat{p}^{(t)}(s_{t+1}|s,a) + \underbrace{\sum_{s'\in\mathcal{S}-\{s_{t+1}\}}\hat{p}^{(t)}(s'|s,a)}_{=1-\hat{p}^{(t)}(s_{t+1}|s,a)}\Big) = \frac{2\,\big(1-\hat{p}^{(t)}(s_{t+1}|s,a)\big)}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1}.$$

IG_{t+1} is a strictly decreasing function $T^{(t)}_{s,a}$.

d. We can use the result of part c and write

$$\mathrm{IG}_{t+1} = \frac{2\ \mathrm{SPE}_{t+1}}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1}.$$

This means that the information-gain as defined in this exercise is equal to a scaled version of the state prediction error, while the scaling factor decreases over time. Hence, the same SPE will lead to less changes in the estimate of transition probabilities with accumulation of observations.

e. We have

$$\lim_{t\to\infty}\mathrm{SPE}_{t+1} = 1 - \lim_{t\to\infty}\frac{T^{(t)}_{s,a,s'} + \epsilon}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon} = 1 - p(s'|s,a)$$

and

$$\lim_{t\to\infty}\mathrm{IG}_{t+1} = \lim_{t\to\infty}\frac{2\ \mathrm{SPE}_{t+1}}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1} = \lim_{t\to\infty}\frac{2\ \lim_{t\to\infty}\mathrm{SPE}_{t+1}}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1} = \lim_{t\to\infty}\frac{2\ (1-p(s'|s,a))}{T^{(t)}_{s,a} + |\mathcal{S}|\epsilon + 1} = 0.$$

This implies that the limit of IG is always equal to zero independently of the level of the stochasticity in the environment. However, SPE can have high values even in the limit of $t \to \infty$ if the environment is stochastic (if $0 < p(s'|s,a) < 1$). Hence, SPE is prone to the Noisy-TV problem but IG is not.

Exercise 2. Disagreement and information-gain

Consider an environment with a finite set of states $\mathcal{S}$ and a finite set of actions $\mathcal{A}$. At each time $t > 0$, we assume that the agent uses its past experiences (i.e., $s_0, a_0, \cdots, s_{t-1}, a_{t-1}, s_t$) and estimates the environment transition probabilities with K different and parallel models, i.e., for $k \in \{1, \ldots, K\}$, we have

$$\hat{p}^{(t)}_k(s'|s,a) = \frac{T^{(t)}_{s,a,s'} + \hat{p}^{(0)}_k(s'|s,a)}{T^{(t)}_{s,a} + 1},$$

where $T^{(t)}_{s,a}$ is the number of times that the agent has taken action a in state s until time t, $T^{(t)}_{s,a,s'}$ is the number of times that taking action a in state s took the agent to state s', and $\hat{p}^{(0)}_k(s'|s,a)$ is the random initialization of model $k \in \{1, \ldots, K\}$. The agent uses

$$\hat{p}^{(t)}(s'|s,a) = \frac{1}{K}\sum_{k=1}^{K}\hat{p}^{(t)}_k(s'|s,a) = \frac{T^{(t)}_{s,a,s'} + \hat{p}^{(0)}(s'|s,a)}{T^{(t)}_{s,a} + 1},$$

as its final estimate.

Consider $s_t = s$ and $a_t = a$. In this exercise, we study how the disagreement of the different K models relate to the information-gain of the transition $(s, a) \to s_{t+1}$.

a. Repeat what you did in Exercise 1 to calculate the information gain defined as

$$\mathrm{IG}_{t+1} = \sum_{s' \in \mathcal{S}} \left| \hat{p}^{(t+1)}(s'|s,a) - \hat{p}^{(t)}(s'|s,a) \right|$$

as a function of $T_{s,a}^{(t)}$ and $\mathrm{SPE}_{t+1} = 1 - \hat{p}^{(t)}(s_{t+1}|s,a)$.

b. We define the disagreement at time t as

$$\mathrm{D}_t = \frac{1}{K} \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \left(\hat{p}^{(t)}(s'|s,a) - \hat{p}_k^{(t)}(s'|s,a) \right)^2 .$$

Find D_t as a function of $T_{s,a}^{(t)}$ and the initial disagreement D_0 at $t = 0$.

c. We now compare three different intrinsic rewards: the State Prediction Error SPE, the Information Gain IG, and the Disagreement D. Let us assume that we know the true transition probabilities $p(.|s,a)$ and that $\lim_{t\to\infty} T_{s,a}^{(t)} = \infty$ (i.e., agents choose each action infinitely many times). For a given next state $s_{t+1} = s'$, compare the limits

$$\lim_{t\to\infty} \mathrm{SPE}_{t+1}, \quad \lim_{t\to\infty} \mathrm{IG}_{t+1}, \quad \text{and} \quad \lim_{t\to\infty} \mathrm{D}_t.$$

What do these results imply about seeking these different intrinsic rewards in the presence of stochasticity? Which intrinsic reward is less prone to the noisy-TV problem?

Solution:

a. We can repeat the exact same procedure as in Exercise 1 and find

$$\mathrm{IG}_{t+1} = \frac{2\ \mathrm{SPE}_{t+1}}{T_{s,a}^{(t)} + 2}.$$

b. Using the definition of D_t, we have

$$\begin{aligned} \mathrm{D}_t &= \frac{1}{K} \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \left(\frac{T_{s,a,s'}^{(t)} + \hat{p}^{(0)}(s'|s,a)}{T_{s,a}^{(t)} + 1} - \frac{T_{s,a,s'}^{(t)} + \hat{p}_k^{(0)}(s'|s,a)}{T_{s,a}^{(t)} + 1} \right)^2 \\ &= \frac{1}{\left(T_{s,a}^{(t)} + 1\right)^2} \frac{1}{K} \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \left(\hat{p}^{(0)}(s'|s,a) - \hat{p}_k^{(0)}(s'|s,a) \right)^2 = \frac{\mathrm{D}_0}{\left(T_{s,a}^{(t)} + 1\right)^2}. \end{aligned}$$

c. We have (similar to Exercise 1)

$$\lim_{t\to\infty} \mathrm{SPE}_{t+1} = 1 - p(s'|s,a), \quad \lim_{t\to\infty} \mathrm{IG}_{t+1} = 0, \quad \text{and} \quad \lim_{t\to\infty} \mathrm{D}_t = 0.$$

This implies that the limits of disagreement and IG are always equal to zero independently of the level of the stochasticity in the environment. However, SPE can have high values even in the limit of $t \to \infty$ if the environment is stochastic (if $0 < p(s'|s,a) < 1$). Hence, disagreement and IG are robust with respect to the noisy-TV problem.