Multi-Agent Systems

Homework Assignment 4

MSc AI, VU

E.J. Pauwels

Version: November 29, 2023 — Deadline: Wednesday, December 6, 2023 (23h59)

4 Fictitious Play

Consider the following pay-off matrix for a 2-player simultaneous game (Capitals indicate the actions, small letters the probabilities with which the corresponding action is played in a mixed eq.):

	$W(w)$	$X(x)$	$Y(y)$	$Z(z)$
$A(a)$	$1,5$	$2,2$	$3,4$	$3,1$
$B(b)$	$3,0$	$4,1$	$2,5$	$4,2$
$C(c)$	$1,3$	$2,6$	$5,2$	$2,3$

Since this game has no pure Nash equilibrium (check this), it must have at least one mixed Nash equilibrium. Recall that $a + b + c = 1$ and $w + x + y + z = 1$.

1. Program the fictitious play algorithm to find a mixed Nash equilibrium. Do the results make sense to you, i.e. can you – *post hoc* – theoreticallly explain the experimental result? Provide a brief discussion.

5 Monte Carlo simulation

5.1 Recap

Recall that Monte Carlo sampling allows us to estimate the expectation of a random function by sampling from the corresponding probability distribution. More precisely, if $f(x)$ is a 1-dim (continuous) probability density, and $X \sim f$ is a stochastic variable distributed according to this density f, then the expected value of some function φ can be estimated using Monte Carlo sampling by:

$$E_f(\varphi(X)) \equiv \int \varphi(x)f(x)\,dx \approx \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) \qquad \text{for sample of independent } X_1, X_2, \ldots, X_n \sim f.$$

Simulated p-value In the same vein, if you've observed a specific value for φ_{obs} and you need to decide whether this value is *exceptional* (in some sense) rather than typical, you can compute the *simulated p-value* which quantifies how exceptional that observed value φ_{obs} is in the simulated sample $\varphi(X_1), \varphi(X_2), \ldots, \varphi(X_n)$.

5.2 Warming up ...

1. Assume that $X \sim N(0,1)$ is standard normal. Estimate the mean value $E(\cos^2(X))$. Quantify the uncertainty on your result.

5.3 Quantifying the significance of an observed correlation

2. Suppose you're designing a deep neural network that needs to maximize some score function S. The actual design of the network depends on some hyperparameter A. Training the networks is computationally very demanding and time consuming, and as a consequence you have only been able to perform ten experiments to date. Based on these ten data points you observe a slight positive correlation of 0.3 between the value of the hyperparameter A and the score S. If this result is genuine, it suggests to increase A in the next experiment in order to improve the score. But if the correlation is not significant, increasing A could lead you astray. How would you use MC to decide whether the correlation is significant?
Hint: Compute the simulated p-value of the observed result, under the assumption of independence.

5.4 Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence quantifies the similarity (or more precisely, the dissimilarity) of two probability densities. More specifically, given two continuous (1-dim) probability densities f, g, the KL-divergence is defined as:

$$KL(f||g) = \int_{-\infty}^{\infty} f(x) \log\left(\frac{f(x)}{g(x)}\right) dx \quad \equiv \quad E_f\left[\log\left(\frac{f(X)}{g(X)}\right)\right] \tag{1}$$

3. Let $f \sim N(\mu, \sigma^2)$ and $g \sim N(\nu, \tau^2)$ both be normal distributions. Express $KL(f||g)$ as a function of the means and variances of f and g. We mention in passing that the KL expression in eq.1 is called a **divergence** rather than a **distance** because it's not symmetric. Use the expression obtained above to convince yourself of this fact.

4. Check your theoretical result in (3) by computing a sample-based estimate of the KL-divergence (Monte Carlo simulation). Pick an appropriate sample size. Compare the MC estimate to the theoretical result.

6 Exploitation versus Exploration

6.1 UCB versus ϵ-greedy for k-bandit problem

Write a programme to experiment with the exploration/exploitation for the k-bandit problem (take $k = 2$ or some larger value if you're feeling lucky :-). Assume that the arms (a) generate normally distributed rewards with unit standard deviation, but different means $q(a)$ (e.g. randomly generated). Assume that in every single experiment the agent can take a total of $T = 1000$ actions (i.e. arm pulls). Let $L(t)$ be the expected total regret at time t in a sample history of T pulls: , defined as:

$$L(t) = \sum_{i=1}^{t} (q^* - q(a_i)) \qquad t = 1, 2, \ldots, T$$

with corresponing mean (over all histories):

$$\ell(t) = E(L(t)) = E\left(\sum_{i=1}^{t} (q^* - q(a_i))\right)$$

1. Compute the experimental $\ell(t)$ curves for different strategies (ϵ-greedy for different values of ϵ, UCB). Compare to the theoretical lower bound found by Lai-Robbins:

$$\ell(t) \geq A \log(t) \qquad \text{where} \quad A = \sum_{a:\Delta_a \neq 0} \frac{\Delta_a}{KL(f_a || f_a^*)} \quad \text{and} \quad \Delta_a = q^* - q(a).$$

2. Compute and compare the percentage correct decisions (selection of correct arm) under the different strategies (i.e. ϵ-greedy for different values of ϵ, UCB). What is the influence of the UCB hyper-parameter c?

PS: *No need to submit code, only the results.*