Vorbereitung: Lokales Python Setup

- Für die Installation des lokalen Python-Environments folgenden Sie bitte den Anweisungen in der README.txt

- Installiertes Environment aktivieren:
  - => (on Windows) .venv\scripts\activate
  - => (on Linux) source .venv/bin/activate

- Starten des Notebooks via „jupyter notebook“

Übung 1 – Exploration/Exploitation

- Importieren Sie das Jupyter Notebook nArmedBandit.ipynb
- Implementieren Sie die Incremental Sample Average Rule im zur Verfügung gestellten Python-Skript nArmedBandit.py

```
# recalculate Q value which is the average rewards of the arm
Q[arm] = …
```

- Erweitern Sie das Experiment hinsichtlich einer Untersuchung mehrerer Explorationsparameter epsilon (zu untersuchende Werte: e \in [0, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0]). Die Mittelwerte über jeweils 1000 Experimente sollen in einem Graph dargestellt werden
- Was sind Ihre Beobachtungen?

Übung 2: Softmax + UCB

- Implementieren Sie Softmax + Upper-Confidence Bounds.

  → Hinweis zu UCB: um log(0) und das Teilen durch den Wert 0 zu vermeiden, gehen Sie davon aus, dass jede Aktion bereits einmal ausgeführt wurde (d.h. der Aktionscounter startet mit Wert 1). Das ist ok, da die initialisierte Q-Funktion mit dem Wert 0 als erste Schätzung angesehen werden kann.

- Untersuchen Sie die Parameter-Belegungen:
  - Softmax: [0.01, 0.05, 0.1, 0.2, 0.5., 1.0, 2, 5, 10, 20, 100]
  - UCB: [0, 0.01, 0.05, 0.1, 0.2, 0.5., 1.0, 2, 5, 10, 20, 100]
- Was sind gute Parameter, was sind schlechte Parameter? Belegen Sie Ihre Aussage mit statistischen Auswertungen. Erstellen Sie einen Line-Plot der Ergebnisse aller drei Explorationsverfahren nach folgenden Vorgaben:
  - X-Achse: Explorationsparameter des jeweiligen Explorationsverfahrens
    - E-Greedy: [0, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0]
    - Softmax: [0.01, 0.05, 0.1, 0.2, 0.5, 1.0, 2, 5, 10, 20, 100]
    - UCB: [0, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0, 2, 5, 10, 20, 100]

    -> Verwenden Sie bitte eine kategorische X-Achse (keine numerische), da sonst die Ergebnisse im Intervall [0, 1] schlecht lesbar werden
  - Y-Achse: durchschnittlicher Reward pro Schritt in einer Episode der Länge von 1000 Schritten, gemittelt über 1000 Experiment.
- Woher entstammen die Oszillationen bei UCB(100) ?

Übung 3:

- Implementieren Sie eine beliebige eigene Explorations-Methode, oder Methode aus der Literatur, in der Methode custom_method().

- Diese soll beispielsweise den Explorationsparameter über die Zeit verändern, oder irgendetwas anderes sinnvolles machen. Sie können auch gerne ein an die Biologie angelehntes Verhalten implementieren → Literaturrecherche.

- Vergleichen Sie die Performanz ihrer Methode mit den drei Verfahren aus Übung 1+2. Schaffen Sie die Performanz verglichen zu den drei anderen Verfahren zu verbessern (nicht zwingend notwendig)?

Literatur

- Sutton&Barto: Reinforcement Learning an Introduction
  - Incremental Sample Average: Kapitel 2.5 http://www.incompleteideas.net/book/first/ebook/node19.html
  - e-Greedy + Softmax: Kapitel 2.2 + 2.3 http://www.incompleteideas.net/book/first/ebook/node17.html
  - UCB: Seite 35 http://incompleteideas.net/book/RLbook2018trimmed.pdf