Praktikum: Selbstlernende Systeme | Gruppe 2

Aufgabe 3:

1. **Ist A2C ein On- oder Off-policy Lerner? Warum?**

A2C ist eine On-Policy Methode, da sie die Policy, die zur Entscheidungsfindung verwendet wird, versucht zu bewerten bzw. zu verbessern.

2. **Welche Mechanik von A2C bricht die Korrelation der einzelnen Beobachtungen?**

Mehrere Worker, die gleichzeitig verwendet werden und deren State-, Action- & Reward-Paare genutzt werden, um das Netz zu trainieren.

3. **Warum muss die Advantage im Policy-Loss eine Konstante sein?**

Weil wir den Gradienten für den tatsächlichen Wert selbst berechnen wollen und nicht den Advantage. Indem wir es als Konstante behandeln, berechnen wir nur den Gradienten für die Policy.

4. **Warum benötigt A2C keine explizite Explorationsstrategie (bspw. ϵ-greedy) mehr? Wie funktioniert Exploration hier? Erklären Sie anhand eines kleinen (Rechen-)Beispiels.**

Die Actions werden nach jeder Episode aus dem Output des Netzwerks abgetastet. Die Exploration erfolgt nach der On-Policy Methode, bei der die Wahrscheinlichkeiten für die verschiedenen Actions aktualisiert werden. Wir brauchen deshalb keine ϵ-greedy Strategie, um die Umwelt zu explorieren.

Dadurch, dass der policy-value von S immer zwischen 0 und 1 ist, ist der Logarithmus immer negativ, während die policy von S immer positiv ist. Daraus folgt, dass die Summe immer negativ ist. Durch das negative Vorzeichen wird der ganze Ausdruck positiv. Sobald viel Exploration stattfindet, führt das zu einem hohen entropy-value. Gradient Descent minimiert die Loss Funktion, daher muss der Mittelwert der Entropy negiert werden, damit die Entropie durch gradient descent maximiert wird. Dadurch wird implizit die Exploration reduziert.