Dokumentation zu:
DRL-Aufgabenblatt "Der K-armige Bandit"

Erik Viere, Daniel Hilfer, Domenic Scholz

October 19, 2024

Aufgabe 1

Aufgabe 1.1

a)

Abbildung 1 zeigt die initiale Auswertung des Banditenproblems. Um eine Mittelung der Ergebnisse zu erhalten, wurde das Experiment über zehn Durchläufe gemittelt. Die Parameter der Normalverteilungen der vier arme sind in Tabelle 1 dargestellt. Die erste Auswahl des Arms

Arm	Erwartungswert	Standardabweichung
1	2.62	2.78
2	1.35	0.93
3	4.62	2.19
4	3.31	1.28

Table 1: Normalverteilungen der vier Arme Aufgabe 1.1 a)

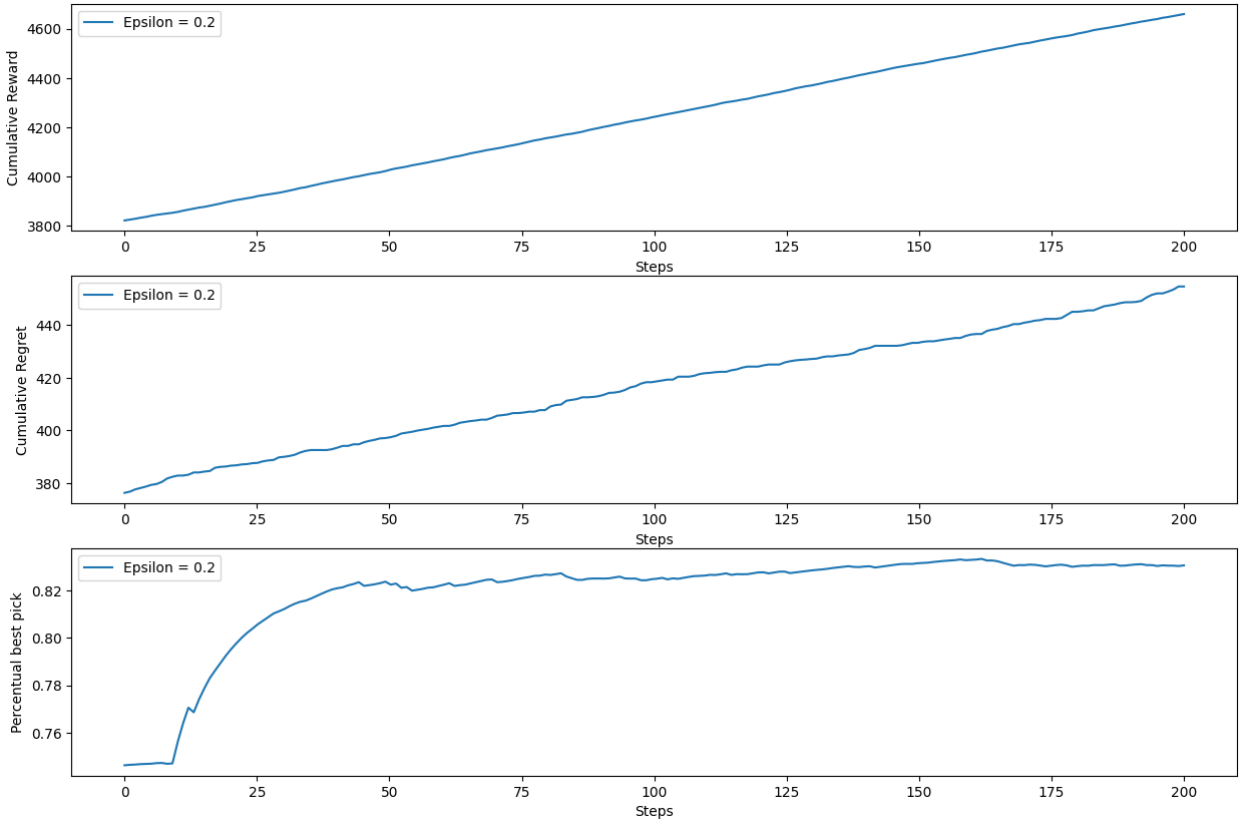

Figure 1: Initiale Auswertung Banditenproblem

erfolgt zufällig, alle darauffolgenden Auswahlen orientieren sich an der ϵ-greedy Methode. Da hier mit einem ϵ-Wert von 20% gearbeitet wird, kann der Anteil der Fälle, in denen der beste Arm gewählt wird, $1 - \epsilon = 80\%$ im Mittel nicht überschreiten. Dies ist im untersten Graphen in Abbildung 1 dargestellt. Außerdem führt dies dazu, dass der kumulierte Regret über den Verlauf der Durchführung des Experiments nicht unbegrenzt lange auf dem gleichen Wert stagnieren kann. Da

in 20% der Fälle ein zufälliger anderer Arm gewählt wird, wird die Differenz der Erwartungswerte zum bisher gesammelten Regret addiert, sodass dieser steigt. Dass der Verlauf des kumulierten Reward einer Gerade ähnelt, liegt daran, dass alle Erwartungswerte ähnlich sind und sich der Reward daher auch bei der Wahl des nicht optimalen Arms um einen nahezu gleichbleibenden Wert ändert. Zudem überwiegt der optimale Arm mit 80% der Züge, sodass der geringere Reward durch die nicht optimalen Arme nur wenig ins Gewicht fällt.

b)

Abbildung 2 zeigt das Verhalten des Algorithmus bei verschiedenen ϵ-Werten. Die zugrundeliegenden Normalverteilungen sind auch hier die aus Tabelle 1, außerdem wurde das Experiment für jeden ϵ-Wert auch hier zehn Mal durchgeführt, um Mittelwerte zu erhalten. Es lassen sich zwei markante Dinge erkennen: Der Endwert der relativen Häufigkeit des optimalen Arms und die Dauer bis zum Erreichen des Endwertes. Kleinere ϵ-Werte erlauben ein häufigeres Auswählen des besten Arms, das heißt, dass der Endwert der relativen Häufigkeit höher ist (sie Graph drei in Abbildung 2). Da dies jedoch auch bedeutet, dass der Exploration-Anteil geringer ist, kann es unter Umständen länger dauern, bis das optimale Ergebnis gefunden wird, auch das lässt sich Graph drei in Abbildung 2 entnehmen. Die Durchläufe, in denen der Exploration-Anteil 10% bzw. 20% beträgt, finden schneller das optimale Ergebnis als der Durchlauf mit nur 1% Exploration-Anteil.

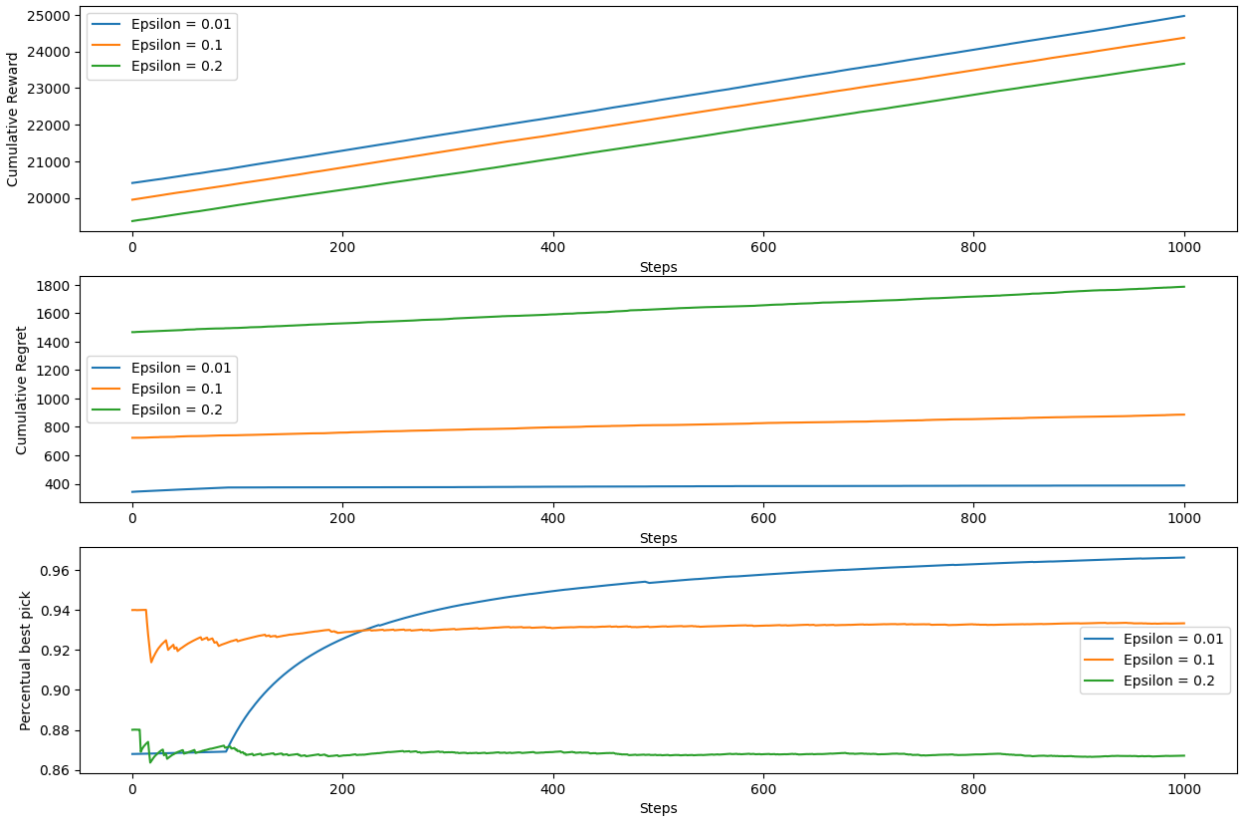

Figure 2: Banditenproblem mit verschiedenen ϵ-Werten

c)

Auch für diesen Aufgabenteil sind die zugrundeliegenden Verteilungen in Tabelle 1 dargestellt und die Durchgänge wurden zehnmal durchgeführt, um eine Mittelung der Ergebnisse zu erreichen. Q wurde nun mit 5.0 anstatt 0.0 initialisiert, die Ergebnisse sind in Abbildung 3 zu sehen. Das

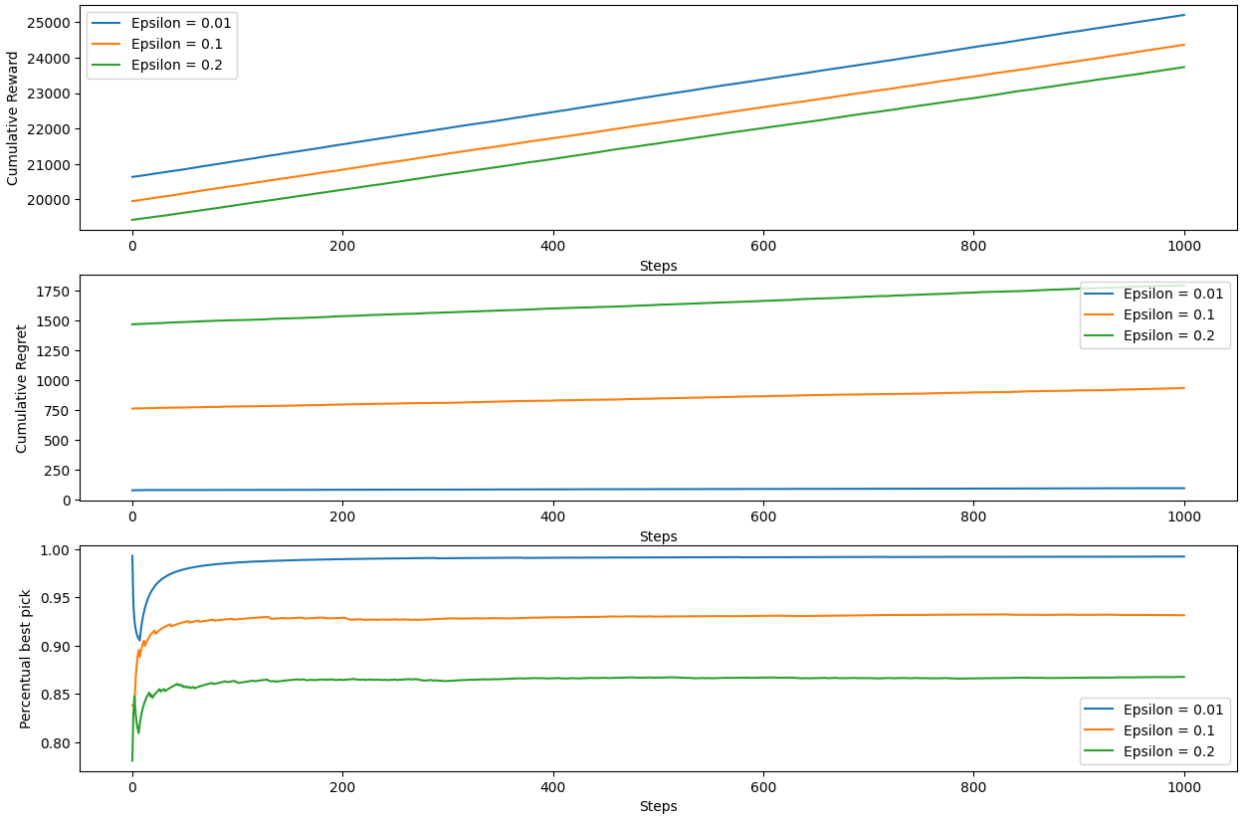

Figure 3: Banditenproblem mit überschätzen Anfangswerten

Überschätzen der Anfangswerte führt dazu, dass die erste Auswahl eines jeden Arms dazu führt, dass sich der Erwartungswert für diesen Arm verringert. Dadurch wird der Bandit zunächst gezwungen, jeden Arm mindestens einmal zu wählen. Arme mit einem signifikant niedrigeren (echten) Erwartungswert, reduzieren den geschätzen Erwartungswert weiter, sodass die Wahrscheinlichkeit höher ist, den optimalen Arm schneller zu finden. Dies äußert sich in den Ergebnissen dadurch, dass nun auch der Bandit mit einem ϵ-Wert von nur 1% schnell den optimalen Arm findet (siehe Graph drei, Abbildung 3), da er durch das Überschätzen sämtlicher Anfangswerte zunächst zur Exploration gezwungen wird.
Auch hier ist wieder zu sehen, dass die Endwerte der relativen Häufigkeit des besten Arms $1 - \epsilon$ nicht überschreiten.

d)

Es werden wieder die in Tabelle 1 dargestellten Verteilungen verwendet, als Initialwerte wird jeweils 0.0 zugrunde gelegt. Die Ergebnisse sind in Abbildung 4 und Abbildung 5 dargestellt.

Aus den Ergebnissen lassen sich drei hauptsächliche Unterschiede erkennen: Der Verlauf des kumulierten Regrets, der Endwert der relativen Häufigkeit des besten Arms und die Dauer bis zur

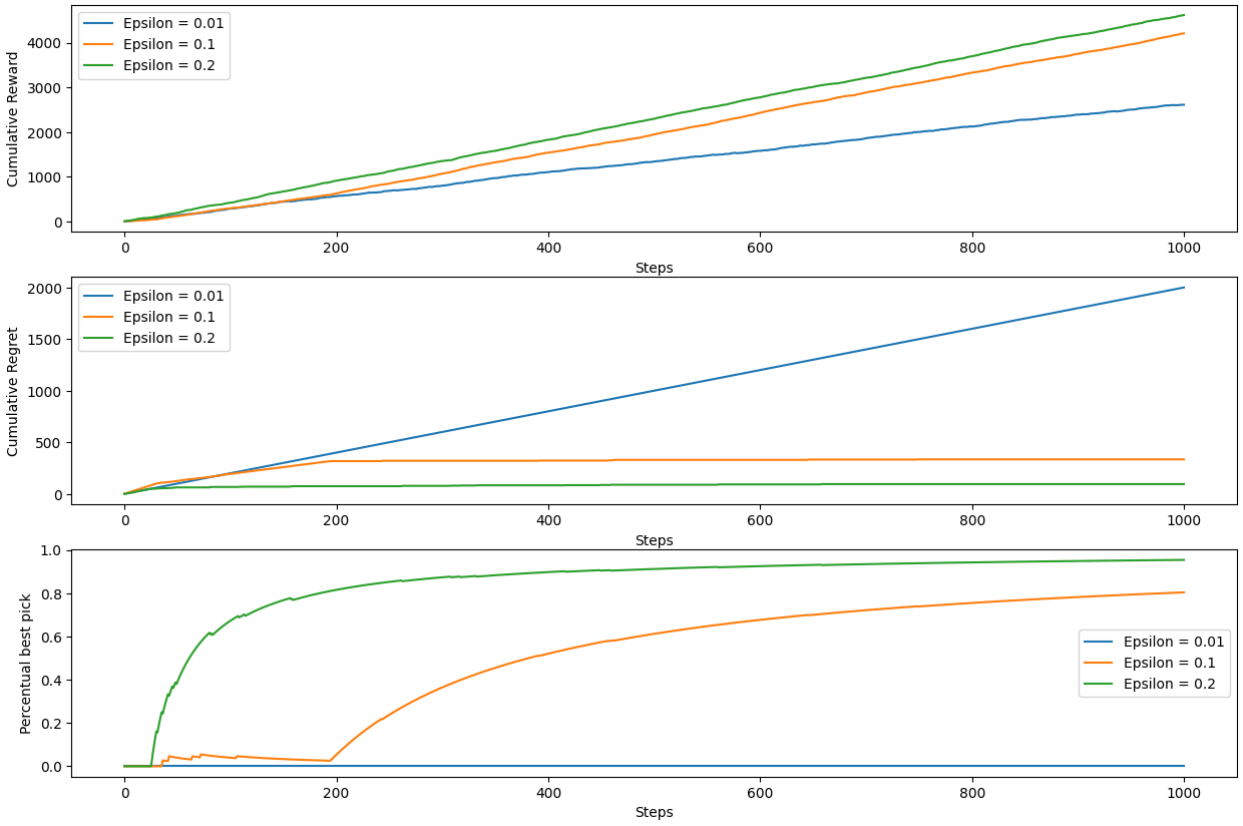

Figure 4: Banditenproblem mit Verringerung von ϵ gemäß $\epsilon - Startwert * \exp(-0.005x)$

konstanten Wahl des besten Arms.
Die zeitliche Verringerung von ϵ führt dazu, dass der Bandit von Zeit zu Zeit weniger zur Exploration gezwungen wird, sondern sich auf die Exploitation des optimalen Arms fokussieren kann. Das führt dazu, dass der Regret stationär bleiben kann, da kein Arm ausgewählt wird (oder nur noch sehr selten), der nicht optimal ist. Dieser Effekt lässt sich zwar auch in Abbildung 5 erkennen, in der der ϵ-Wert über die Zeit nicht verringert wurde, liegt dort jedoch an dem ohnehin schon geringen ϵ-Wert von 1%. In diesem konkreten Fall führt dieser Wert dazu, dass etwa ab dem 400. Durchlauf durch Zufall kein nicht optimaler Arm mehr gewählt wird.
Außerdem hat die Verringerung des ϵ-Wertes Einfluss auf den Endwert der relativen Häufigkeit des optimalen Arms. Da sich der Wert auf nahezu 0 verringert, ist nun als Endwert der relativen Häufigkeit theoretisch $1 - 0 = 1 = 100\%$ möglich.
Abschließend lässt sich ein Unterschied in der Lernphase erkennen. Der verringerte ϵ-Wert kann dazu führen, dass der optimale Arm erst später oder gar nicht erkannt wird. Letzteres kann besonders häufig auftreten, wenn der initiale ϵ-Wert bereits sehr gering ist (siehe $\epsilon = 0.01$ in Abbildung 4).

Aufgabe 1.2

a)

Das Banditenproblem beschreibt ein Problem, bei dem es darum geht, aus k verschiedenen Möglichkeiten die beste zu erkennen und diese zu verfolgen, die konkreten Anwendungsfälle definieren dabei

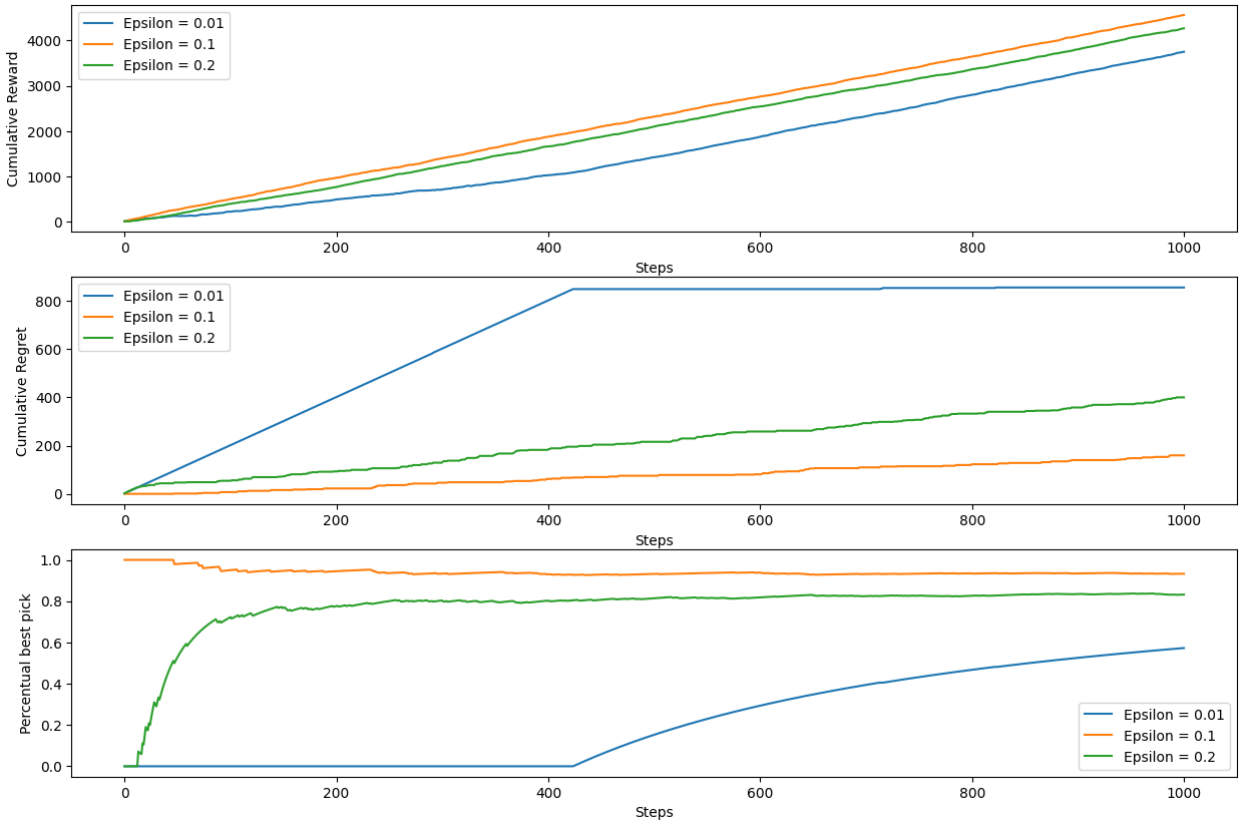

Figure 5: Banditenproblem ohne Verringerung des ϵ-Wertes

"beste Möglichkeit" und "verfolgen" selbst. Nicht-Stationärität beschreibt, dass sich die Güte einer Möglichkeit im Lauf der Zeit verändert (konkreter: Es verändert sich beispielsweise die zu erwartende Belohnung bei der Wahl einer Möglichkeit nach oben oder unten). Dies kann so modelliert werden, indem die wahre erwartete Belohnung einer Möglichkeit alle X Schritte um einen zufälligen oder bestimmten Wert in- bzw. dekrementiert wird, der Erwartungswert der einzelnen Möglichkeiten und somit auch die beste der Möglichkeiten ist also nicht stationär, sondern verändert sich mit der Zeit. Ein Algorithmus zur Lösung des Problems kann im Prinzip ähnlich wie der Algorithmus zur Lösung des stationären Banditenproblems arbeiten. Der Unterschied ist, dass sich der Algorithmus nun nicht mehr sicher sein kann, dass die beste Möglichkeit die beste bleibt. Es sollte also mehr Wert auf die Erkundung gelegt werden, sodass Veränderungen in den Möglichkeiten erkannt werden. Auch ist es möglich, den Algorithmus zwischendurch (teilsweise) zu "resetten", sodass ihm das bisherige Wissen genommen wird und er die Möglichkeiten unvoreingenommen erneut erkunden kann. Letzeres ist jedoch nur sinnvoll, wenn starke Schwankungen in den Möglichkeiten den Banditen zu erwarten sind.Eine bestehende Implementierung könnte a) nach dem Regret oder b) nach ihrer Trägheit beurteilt werden. Trägheit meint dabei, wie schnell oder langsam ein Algorithmus auf eine Veränderung in den Möglichkeiten reagiert und die Schätzwerte und die Wahl der optimalen Möglichkeit anpasst.

b)

Siehe Jupyter-Notebook.

c)

Verwendet wurden auch hier die Verteilungen gemäß Tabelle 1 als Startverteilungen. Zu deren Erwartungswert und Standardabweichung wurde nun jedoch alle 5 Schritte ein zufälliger Wert aus der Verteilung $\mathcal{N}(0.0, 2)$ zum Erwartungswert und ein Wert aus der Verteilung $\mathcal{N}(0.0, 0.5)$ zur Standardabweichung addiert. Das führt dazu, dass sich die Erwartungswerte der Verteilungen über die Zeit verändern aber auch dazu, dass der Index der optimalen Wahl wechselt, nämlich nach etwa 550 Schritten (konstant) von 3 auf 1 (siehe Graph vier in Abbildung 6). Zu sehen ist, dass der Algorithmus auf diese Veränderung nicht reagieren kann und den Index der aus seiner Sicht besten Wahl nicht verändert. Dass der Algorithmus nun eine nicht optimale Wahl trifft, spiegelt sich auch in den gesammelten Belohnungen wieder. In den kumulierten Belohnungen sieht man, dass diese am Ende sinken, heißt, dass negativer Reward gesammelt wird (siehe Graph zwei in Abbildung 6), da weiterhin Möglichkeit drei gewählt wird, die inzwischen einen negativen Erwartungswert hat (siehe Graph vier Abbildung 7), anstatt Möglichkeit eins, welche die optimale mit einem positiven Erwartungswert ist (Graph zwei, Abbildung 7).

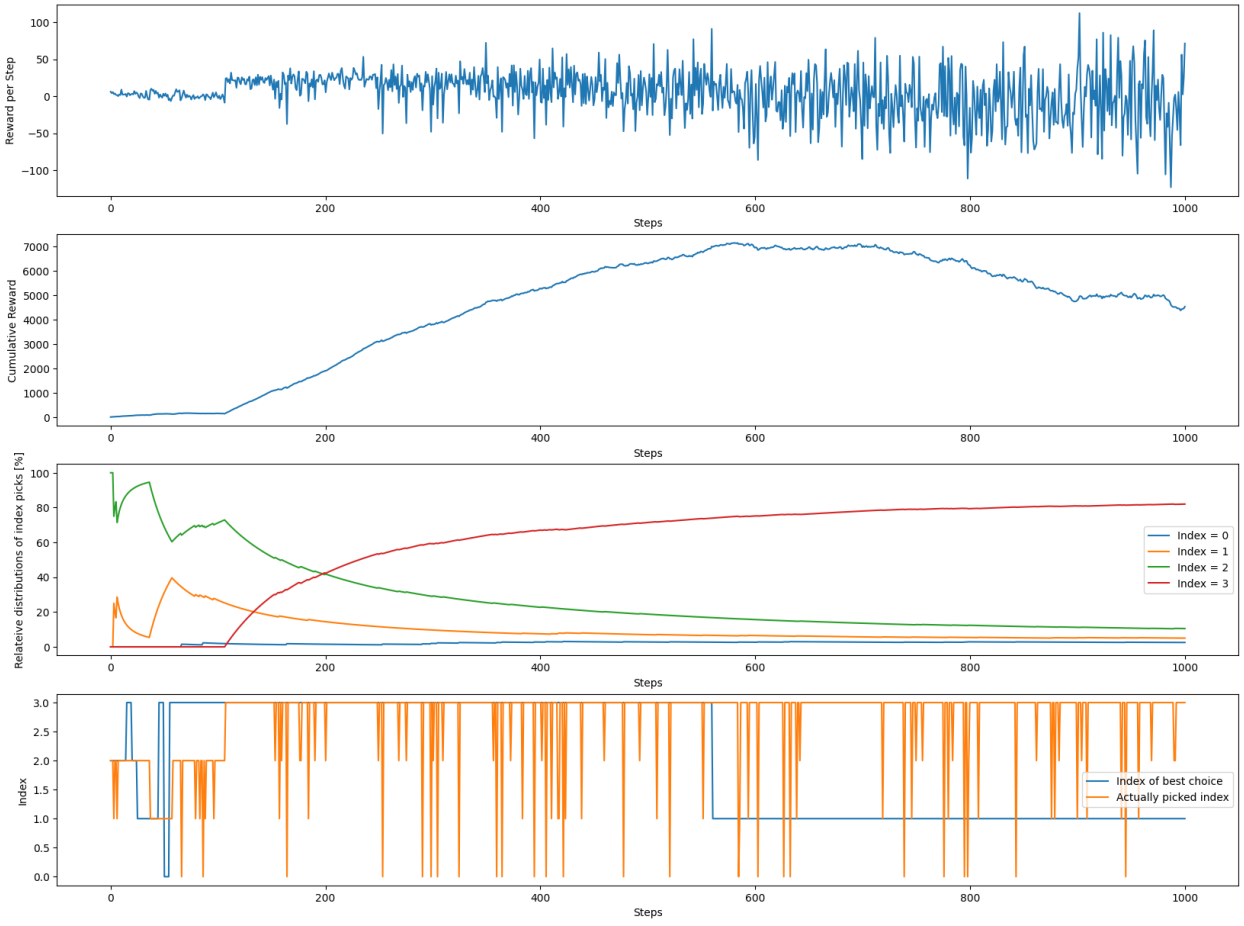

Figure 6: Auswertung des nicht-stationären Banditenproblems

Des Weiteren zeigt Abbildung 7 für jeden Index, wie die Schätzung des Erwartungswertes auf die Veränderung des echten Erwartungswertes reagiert. Zu sehen ist, dass nur die Schätzung für Möglichkeit drei annähernd dem echten Wert folgt, da der Algorithmus diese Möglichkeit noch immer für die beste hält und häufig Werte aus ihr wählt, die die Schätzung beeinflussen. Die

anderen Schätzungen bleiben nahezu konstant, sodass der Algorithmus erst viel zu spät auf die neue optimale Möglichkeit reagieren kann.

Figure 7: Nicht-Stationäres Banditenproblem: Schätzung vs. echter Wert für jede Wahlmöglichkeit

d)

In Aufgabenteil 1.2c) wurde gezeigt, dass die aktuelle Implementierung des Algorithmus nicht auf die Veränderung des optimalen Index reagieren kann. In der Beschreibung zu 1.2a) wurden bereits mögliche Verbesserungsvorschläge angeschnitten. Diese Vorschläge beziehen sich primär auf die Gewichtung des Erkundungs-Anteils. Da die realen Erwartungswerte der Verteilungen nicht statisch sind, ist ein geringer Erkundungs-Anteil oder gar ein zeitabhängiges Verschwinden dieses Anteils hinderlich für den Algorithmus. Das heißt im Umkehrschluss, dass ein höherer Erkundungs-Anteil förderlich sein könnte, indem so auf den Wechsel der optimalen Wahl reagiert werden kann. Außerdem kann ein "resetten" des Algorithmus von Zeit zu Zeit sinnvoll sein. Dies kann zum einen bei der Detektion der Änderung der optimalen Wahl helfen und zum anderen den Algorithmus dazu befähigen, die Erwartungsschätzung anzupassen. Der Grund für letzteren Vorteil liegt darin, dass die Schätzung des Erwartungswertes auf Durchschnittsbildung basiert. Das heißt, dass vom Durchschnitt abweichende Werte im zeitlich späteren Teil der Durchführung keinen großen Einfluss mehr auf den Durchschnittswert haben, da die vorherigen Werte diese Schätzung bereits gefestigt haben und mengenmäßig (stark) überwiegen. So kann nur schlecht auf Veränderung der Erwartungswerte

reagiert werden, selbst wenn sich der Index der optimalen Wahl nicht ändert. Die Konsequenz für eine Verbesserung der Implementierung ist, dass mit einem “Moving Average” über ein kleineres Fenster gearbetet werden könnte, um der Schätzung die Möglichkeit zu geben, auch im späteren Verlauf der Durchführung auf abweichende Werte zu reagieren.