Deep Reinforcement Learning

Blatt 01

Sven Ullmann, Valentin Adam

Aufgabe 1.1

Setting:
Wir haben einen zunächst 4-armigen Banditen implementiert, der folgende über die Zeit gleichbleibende Mittelwerte und Standardabweichungen besitzt. Als ε-Wert für den implementierten ε-greedy Algorith-

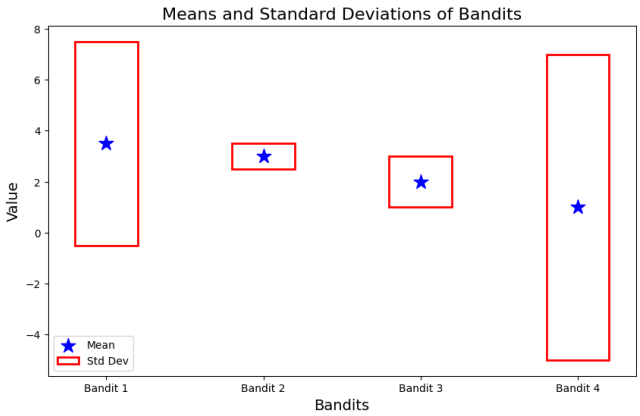


Figure 1: Mittelwerte: 3.5,3,2,1 Standardabweichungen: 4,0.5,1,6

mus haben wir zunächst $\varepsilon = 0.1$ gewählt.
Auswertung:
Wie in Figure 2 zu sehen, erhielt unsere Methode im Schnitt über 1000 Testläufe mit je 10000 Schritten

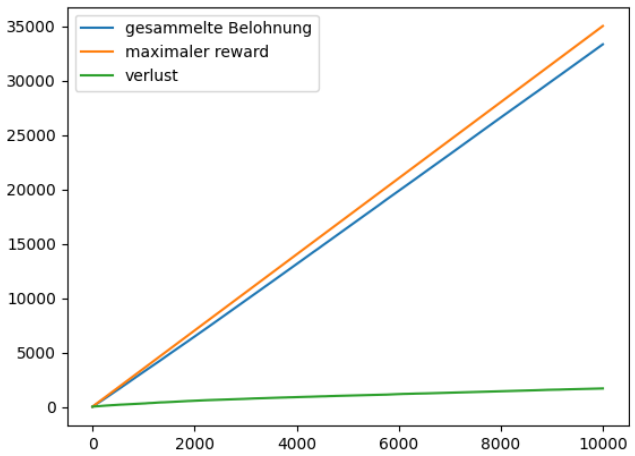


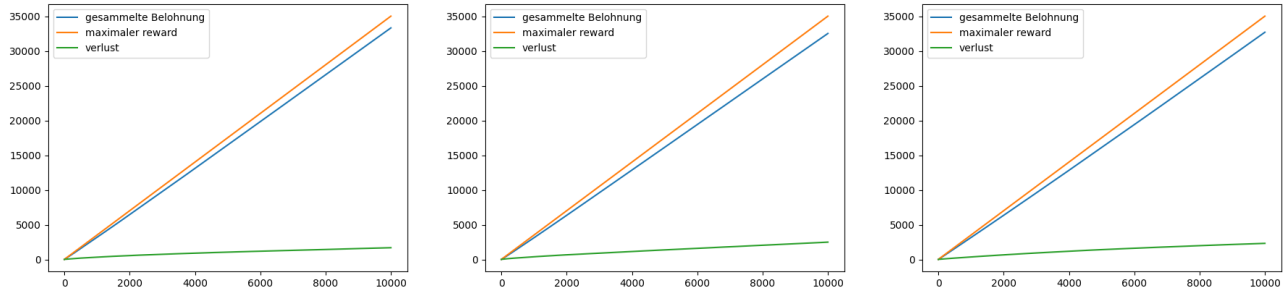
Figure 2: Abbildung des gesammelten Rewards, des maximal möglichen Rewards und des Verlustes der ε-greedy Methode des 4-armigen Banditen. Gemittelt über 1000 Testläufe mit je 10000 Schritten, $\varepsilon = 0.1$.

relativ gute Ergebnisse. Der erzielte Reward liegt nur knapp unter dem maximalen Reward und somit

ist der Verlust relativ gering. Während der Durchführung wählte unserer Algorithmus im Schnitt in ca 81% der Fälle den tatsächlich besten Banditen (Bandit 1) aus.

Variation in Parameter ε

Beim Variieren des ε für den greedy-Algorithmus fällt auf, dass verschiedene ε-Werte zu unterschiedlich guten Ergebnissen führt.
Das ε entscheidet im Algorithmus, wie wahrscheinlich es ist, ob der Algorithmus in einem Schritt weiter mit dem bisherig besten Banditen geht oder ob er einen zufällig anderen zieht. Mit der Wahl des ε ist die beste Balance zwischen Erforschung und Ausbeutung zu treffen.
In unseren Beobachtungen ist $\varepsilon = 0.1$ die beste Wahl. In diesem Fall wählte unser Algorithmus im Schnitt (über 1000 Durchläufe) in 81% der Schritte den optimalen Banditen. Ähnlich gut war die Wahl mit $\varepsilon = 0.2$. Hier wurde in $79,97$ % der Fälle der beste Bandit ausgewählt. Mit großem Abstand am schlechtesten ist die Wahl $\varepsilon = 0.01$. Hier wurde offensichtlich zu wenig erforscht. In diesem Fall wählte der Algorithmus nur in 49.69% der Schritte den optimalen Banditen aus (in den anderen ca. 50% wurde größtenteils der zweitbeste Bandit ausgewählt). Große Unterschiede sind in den Gesamtgewinnen/Verlusten trotzdem nicht zu erkenne, da die Mittelwerte der Banditen nicht weit auseinander liegen. (vgl. Figure 3)



(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.2$ (c) $\varepsilon = 0.01$

Figure 3

Variation der Initialisierung der Schätzwerte

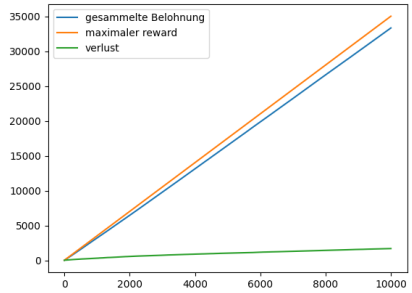
Als nächstes Untersuchen wir die Auswirkungen der Initialisierungen der Schätzwerte. Wir betrachten als Erste Variante, dass alle Mittelwerte der Banditen inital auf 0 geschätzt werden und vergleichen diese mit der Variante, dass die Mittelwerte sehr optimistisch geschätzt werden (Wir schätzen alle auf 0). Wir beobachten, dass bei der ersten Variante ca. 82% der Fälle der optimale Bandit gewählt wurde. Bei der Zeiten ca. 84%. Somit ist die Letztere auch zu bevorzugen. Die Ergebnisse decken sich mit der in der Vorlesung vorgestellten Theorie. Diese besagt, dass Underestimation deutlich schlimmer als Overestimation ist. Dies beruht darauf, das bei Overestimation, der Fehler korriegiert wird und bei Underestimation nicht, da der unterschätzte Bandit im Folgenden selten ausgewählt wird.

Adaptive Erkundung:

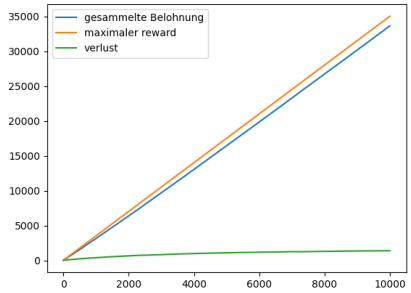
Wir vergleichen drei Strategien zur Erkundung:

1. Die Erkundungsrate bleibt stabil -¿ ε bleibt über alle Schritte konstant
2. Die Erkundungsrate nimmt linear ab -¿ ε nimmt linear ab, sodass es beim letzten Schritt gleich 0 ist
3. Die Erkundungsrate nimmt exponentiell ab -¿ ε nimmt exponentiell ab

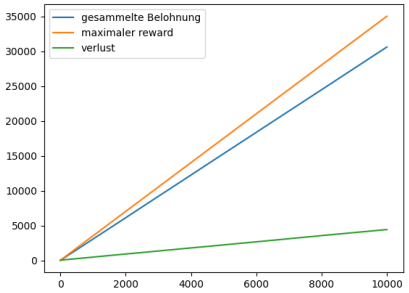
Wir beobachten, dass der lineare Abnahme der Erkundungsrate am Besten performed. Ab einer bestimmten Iteration bleibt der Verlust nahezu konstant. Dies kommt daher, dass mit der Zeit die Erkundung abnimmt und damit immer mehr der bisher beste Bandit ausgenutzt wird. Bei der exponentiellen Abnahme von ε nimmt die Erkundungsrate jedoch so schnell ab, dass die Wahrscheinlichkeit den falschen Banditen als "optimalen" Banditen auszuwählen größer ist. Beim konstanten ε wird auch



(a) Konstantes ε



(b) Lineare Abnahme von ε



(c) Exponentielle Abnahme von ε

Figure 4

dann noch unnötiger Weise erkundet, obwohl schon genug Wissen angehäuft wurde um den "optimalen Banditen zu finden".

Aufgabe 1.2

Im Vergleich zum allgemeinen Banditen Problems, geht es bei dem nicht stationären k-armigen Banditenproblems darum, dass die tatsächlichen erwarteten Belohnungen nicht stationär im Verlaufe des "Spiels" sind. Nach jeder Ziehung kann sich die erwartende Belohnung unabhängig von den anderen Banditen und unabhängig von der vorherigen erwartenden Belohnung verändern. Zum Beispiel kann dies, wie in (b) durch eine Ziehung aus einer Normalverteilung mit $\mu = 0.0$ realisiert werden.
Zur Lösung des nicht-stationären Problems können folgende Anpassungen des ε- Greedy Algorithmus vorgenommen werden:

- Eine ε-Abnahme darf nicht verwendet werden, da man sich niemals auf einen Banditen beschränken sollte, da diese sich offensichtlich im Laufe der Zeit verändern. Somit wird über den ganzen Zeitraum in gleichem Maße erforscht!
- Zur Berechnung der geschätzten erwartetem Belohnung sollten nur die letzten k-Werte in Betracht gezogen werden. Oder die beobachteten gezogenen Werte werden proportional dazu gewichtet, wie lange es her ist, dass sie gezogen wurden. Somit wird verhindert, dass nicht mehr aktuelle "alte" Werte die Entscheidung zu stark beeinflussen, da diese schon nicht mehr aktuell sein können.

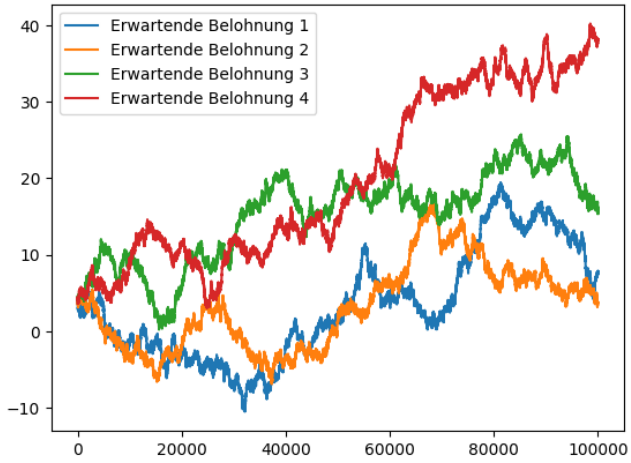


Figure 5: Visualisierung der Veränderung der zu erwartendenden Belohnung

Der ε-greedy Algorithmus wählt bei der Entwicklung der erwarteten Belohnung (vgl. Figure 5) die Banditen aus wie dargestellt in Figure 6. Dabei wurden aus Gründen der Darstellung jeweils 10 konsekutive ausgewählte Banditen zu einem gemittelt. In Figure 6 sieht man ganz gut, dass richtigerweise ca. ab dem 20000sten Schritt nur noch der 3. und 4. Bandit ausgewählt wird. Diese Entwickeln sich auch tatsächlich zu den beiden bestlichen Banditen. Die generelle schlechte Performance des ε-greedy Algorithmus ist in Figure 7 zu sehen.
Die oben genannten Modifikationen des ε-greedy Algorithmus könnten zu einer Verbesserung führen.

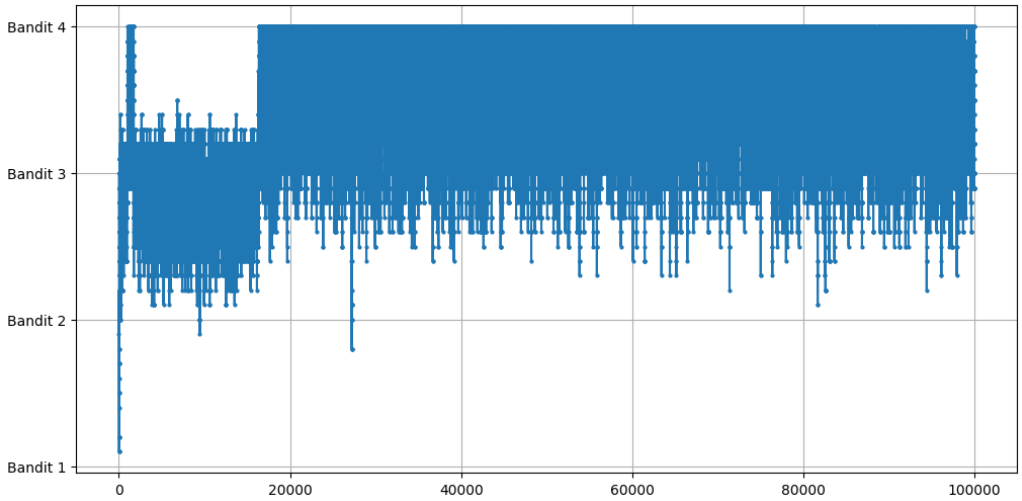


Figure 6: Visualisierung, wann welcher Bandit gewählt wurde.

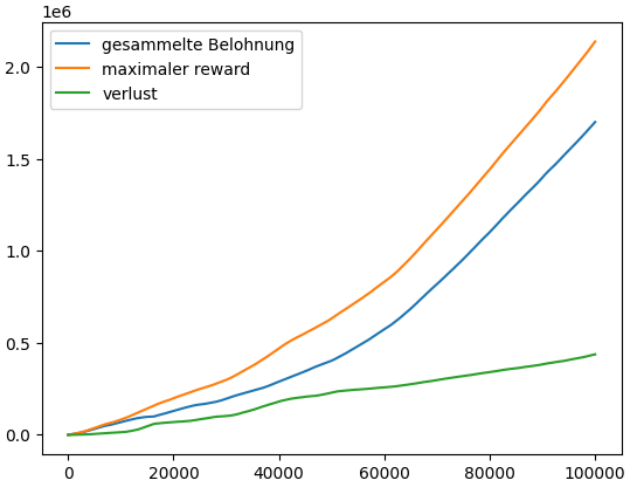


Figure 7: Performance des ε-greedy Algorithmus bei nicht stationärem k-ären Banditen.