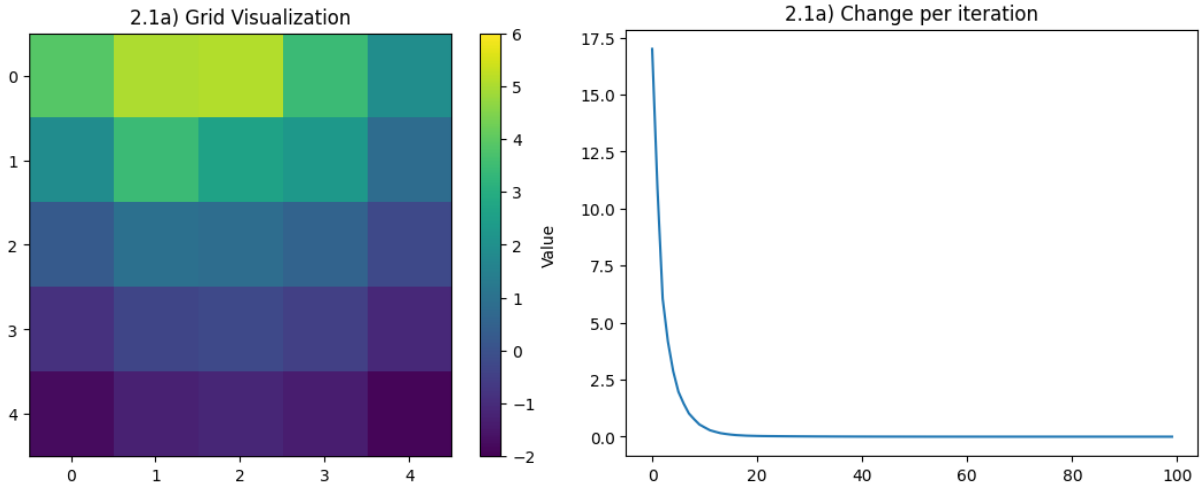
Task 2.1 a)

Die Visualisierung zeigt die Bewertung der Zustände nach 100 Iterationen. Mit einer Zufälligen Auswahl, ist der Zustand (2, 0) optimal, da in 50% der Fällen ein Reward erzielt wird. Die beiden Zustände (1,0) und (3,0) erhalten ebenfalls eine Bewertung, können jedoch niemals vom Algorithmus erreicht werden, da dieser auf ein anderes Feld springt.

Wie zu erwarten, erhalten Felder, die weit von A und B entfernt sind eine niedrige Bewertung. Das Problem mit der Random Strategy ist, dass jede mögliche Aktion für die Bewertung berücksichtigt wird, anstatt aus der Erfahrung Schlüsse zu ziehen. Demnach gibt es keine sehr guten Bewertungen, weil schlechte Aktionen gewählt werden.

Der zweite Graph zeigt die absolute Änderung der Bewertung je Epoche. Es ist klar zu erkennen, dass der Algorithmus konvergiert und die Matrix sehr schnell nichtmehr verändert wird.



Task 2.1b)

Diesmal wird die Bewertung je nach möglicher Aktion festgehalten (statt dem Durchschnitt wie in 2.1a). Für den ersten Graph wurde für jeden Zustand die optimale Aktion ausgewählt, was einer sehr guten Bewertung entspricht. Hierbei „lernt“ der Algorithmus die Distanz zu A und B und gewichtet die Zustände entsprechend (Sehr viele identische Werte je nach Abstand). Die Random Policy bedeutet, dass alle Aktionen gleich oft getestet werden, doch am Ergebnis sind die besten Zustände klar zu erkennen. Zu Beginn gibt es sehr starke Änderungen in der Bewertung, nach 100 Epochen bleibt diese jedoch auch stabil. Durch die 4 Aktionen gibt es 4-mal so viele Werte, entsprechend ist die gesamte Änderung als Summe hier höher.