Universität Bielefeld

Technische Fakultät
AG Neuroinformatik
Robert Haschke

Sekretariat: P. Udelhoven
Raum M7-113
Tel. 106-6891

Übungen zu “Robotik II”
WS 2003/04 Blatt 11

Ausgabe: 15.1.2004, **Abgabe:** 22.1.2004

Aufgabe 11.1, Nicht-deterministische Rewards:
Bei nicht-deterministischen MDP's ist es sinnvoll den Reward auch in Abhängigkeit vom Folgezustand s' zu beschreiben. Man erhält dann eine Funktion $r(s, a, s') : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$.

- Wie ändert sich in diesem Falle die Definition der Q-Funktion

$$Q^\pi(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) \cdot V^\pi(s') \ ?$$

- Wie sehen die Bellman'schen Optimalitätsgleichungen in diesem Falle aus?

Aufgabe 11.2, Stochastische Policies: Oft ist es auch sinnvoll nicht-deterministische Policies zu betrachten. Die Policy ist dann keine Auswahlfunktion $\pi : \mathcal{S} \to \mathcal{A}$, sondern gibt die Wahrscheinlichkeit $\pi(s, a)$ an, im Zustand s die Aktion a auszuwählen. Wie sehen in diesem Fall die Bellman'schen Optimalitätsgleichungen der Q-Funktion aus?

Aufgabe 11.3, Delayed Rewards:

- Wann kommt es beim Q-Lernen zu Updates der Q-Funktion? Beginnen Sie bei einer Q-Funktion $Q \equiv 0$.
- Würden Sie eher delayed Rewards verwenden, d.h. eine Rückmeldung erst bei Abschluss/Misserfolg einer Aufgabe geben, oder sofortige Rewards? Wie könnten diese aussehen?