FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Vorlesung
Grundlagen adaptiver Wissenssysteme

Prof. Dr. Thomas Gabel
Frankfurt University of Applied Sciences
Faculty of Computer Science and Engineering
tgabel@fb2.fra-uas.de

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Vorlesungseinheit 8

Modellfreies Lernen

Modellfreies Lernen

Lernziele

- vollständige Ersetzung des Modells
- gierige Wertfunktionsauswertung ohne Modell

Modellfreies Lernen

Überblick

1. Motivation

Modellfreies Lernen

Überblick

1. Motivation

2. Zustands-Aktions-Wertfunktionen

Modellfreies Lernen

Überblick

1. Motivation
2. Zustands-Aktions-Wertfunktionen
3. Exploration

Modellfreies Lernen

Überblick

1. Motivation
2. Zustands-Aktions-Wertfunktionen
3. Exploration
4. Optimistische Strategieiteration

Modellfreies Lernen

Überblick

1. Motivation
2. Zustands-Aktions-Wertfunktionen
3. Exploration
4. Optimistische Strategieiteration

Von der Bewertung zur Strategie

Strategieiteration

- Bewertung der Pfadkosten: $TD(\lambda) \Rightarrow V^{\pi}$
- Idee: Anwendung Policy Iteration:
  - Strategie – Bewertung – Greedy-Auswertung – neue Strategie – Bewertung – ...

Erinnerung: Die Begriffe Pfadkostenvektor und (Zustands-)Bewertungsfunktion (auch Wertfunktion, Value Function) sind gleichbedeutend.

Modellabhängigkeit

- Natürlich braucht man für die Greedy-Auswertung einer Bewertungsfunktion ein Modell.
- Eine Möglichkeit besteht darin, das Modell der Umgebung zu schätzen, also auf Basis der getätigten Interaktionen mit der Umwelt zu approximieren.

Modellabhängigkeit

- Natürlich braucht man für die Greedy-Auswertung einer Bewertungsfunktion ein Modell.
- Eine Möglichkeit besteht darin, das Modell der Umgebung zu schätzen, also auf Basis der getätigten Interaktionen mit der Umwelt zu approximieren.
- Wir lernen in dieser Vorlesung eine zweite Möglichkeit kennen: die Verwendung sogenannter Zustands-Aktions-Wertfunktionen.
- Der später vorgestellte Q-LEARNING Algorithmus erlaubt das Lernen komplett ohne Modell; die dahinterstehenden Prinzipien sind dieselben.

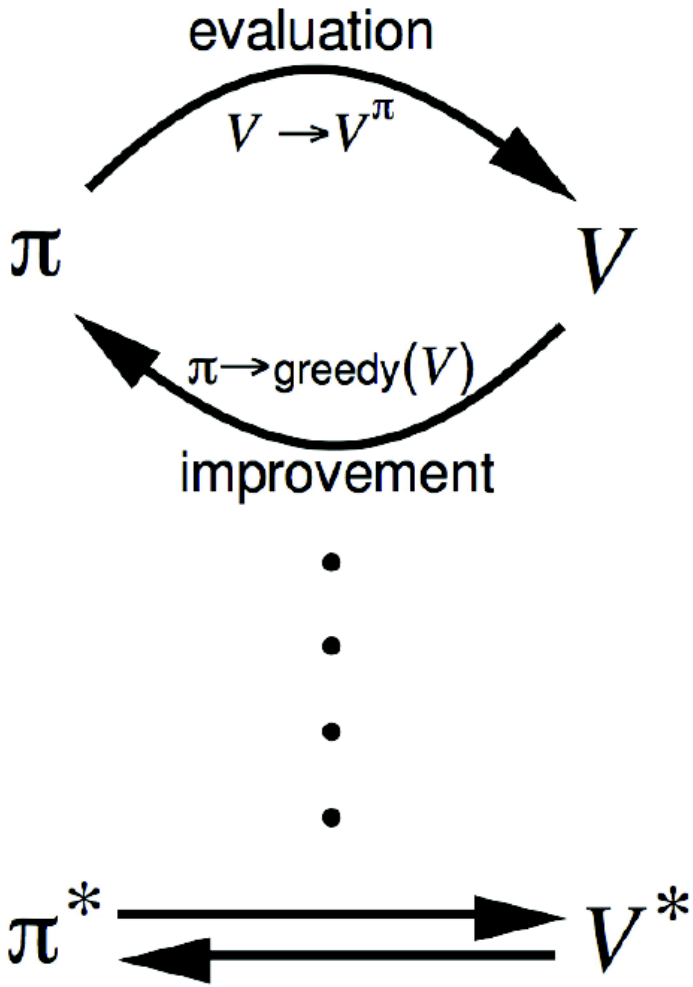
Modellabhängigkeit

- Natürlich braucht man für die Greedy-Auswertung einer Bewertungsfunktion ein Modell.
- Eine Möglichkeit besteht darin, das Modell der Umgebung zu schätzen, also auf Basis der getätigten Interaktionen mit der Umwelt zu approximieren.
- Wir lernen in dieser Vorlesung eine zweite Möglichkeit kennen: die Verwendung sogenannter Zustands-Aktions-Wertfunktionen.
- Der später vorgestellte Q-LEARNING Algorithmus erlaubt das Lernen komplett ohne Modell; die dahinterstehenden Prinzipien sind dieselben.
- Beim Einsatz im Rahmen eines Lernsystems soll i.d.R. kein Modell bekannt sein, sonst kann man eigentlich konventionelles dynamisches Programmieren ausführen.
  - Ein Grund, warum Lernen auch bei bekanntem Modell sehr sinnvoll sein kann, ist die Idee, DP nur auf den Teilen des Zustandsraums anzuwenden, die auf relevanten Trajektorien ausgehend von einer Menge Startzustände liegen.

Verallgemeinerte Strategieiteration (1)

Zweiphasiges Verfahren:

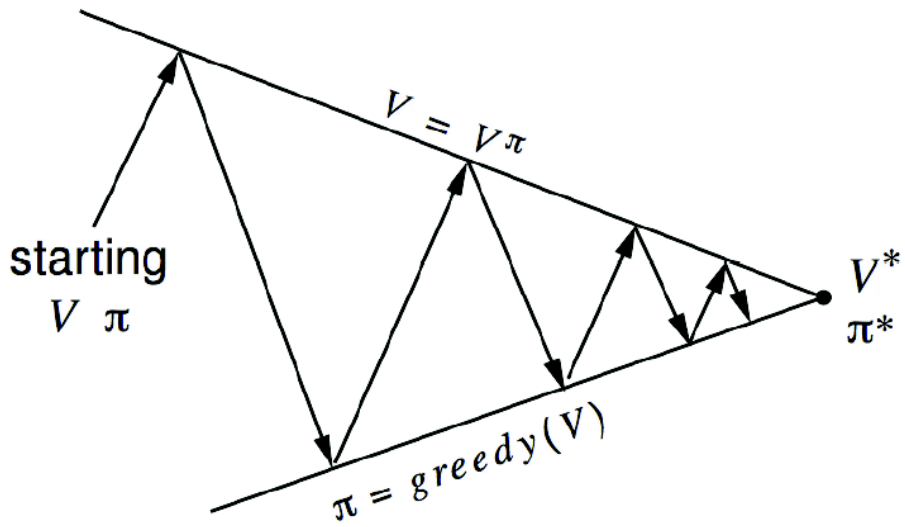
- Phasen der Strategiebewertung und Strategieverbesserung wechseln sich ab.
- Strategiebewertung ermittelt V^π für aktuelle Strategie π
- Strategieverbesserung ermittelt (“gierig”) verbesserte Strategie π' mit $\pi' \geq \pi$



Verallgemeinerte Strategieiteration (2)

Zweiphasiges Verfahren:

- Phasen der Strategiebewertung und Strategieverbesserung wechseln sich ab.
- Problem: Die Zwischenbewertungen werden eigentlich gar nicht (zumindest nicht 100% exakt) benötigt.
  - Hier wird eventuell zu viel Zeit auf eine zu exakte Lösung verschwendet.



Verallgemeinerte Strategieiteration (3)

Erkenntnisse:

1. Wir müssen eine Möglichkeit finden, eine gierige Strategieverbesserung durchzuführen (auch ohne Modell).
2. Wir müssen uns die Frage stellen, wie häufig wir die Strategie anpassen, also durch gierige Strategieverbesserung modifizieren.
   - Soll wirklich gewartet werden, bis durch die Strategiebewertung die perfekte V^{π} vorliegt?

Modellfreies Lernen

Überblick

1. Motivation
2. Zustands-Aktions-Wertfunktionen
3. Exploration
4. Optimistische Strategieiteration

Die Q-Funktion (1)

Bekannt:

- Eine Wertfunktion $V^\pi : S \to \mathbb{R}$ bewertet für $i \in S$ mit $V^\pi(i)$, die zu erwartenden Kosten, die ab Zustand i für den Agenten anfallen, wenn er fortan seine Aktionen ausschließlich gemäß Strategie π wählt.

Die Q-Funktion (1)

Bekannt:

- Eine Wertfunktion $V^\pi : S \to \mathbb{R}$ bewertet für $i \in S$ mit $V^\pi(i)$, die zu erwartenden Kosten, die ab Zustand i für den Agenten anfallen, wenn er fortan seine Aktionen ausschließlich gemäß Strategie π wählt.

Definition (Zustands-Aktions-Wertfunktion Q)

Für eine gegebene Strategie π ist eine Zustands-Aktions-Wertfunktion $Q^\pi : S \times A \to \mathbb{R}$ definiert als

$$Q^\pi(i,a) := \sum_{j=0}^{n} p_{ij}(a)\,(c(i,a) + Q^\pi(j,\pi(j)))$$

für alle $i \in S = \{0,\ldots,n\}$ und alle $a \in A$.

Die Q-Funktion (2)

Erläuterung:

- Eine Q-Funktion schätzt die erwarteten Kosten des Agenten ab, wenn dieser im Zustand *i* die Aktion *a* wählen würde und sich danach gemäß Strategie π verhalten würde.
- Es gilt einerseits:

$$V^{\pi}(i) = \sum_{j=0}^{n} p_{ij}(\pi(i)) \cdot (c(i, \pi(i)) + V^{\pi}(j))$$

Die Q-Funktion (2)

Erläuterung:

- Eine Q-Funktion schätzt die erwarteten Kosten des Agenten ab, wenn dieser im Zustand *i* die Aktion *a* wählen würde und sich danach gemäß Strategie π verhalten würde.
- Es gilt einerseits:

$$V^{\pi}(i) = \sum_{j=0}^{n} p_{ij}(\pi(i)) \cdot (c(i, \pi(i)) + V^{\pi}(j))$$

- Und andererseits:

$$Q^{\pi}(i, a) = \sum_{j=0}^{n} p_{ij}(a) \cdot (c(i, a) + Q^{\pi}(j, \pi(j)))$$

- Also:

Die Q-Funktion (2)

Erläuterung:

- Eine Q-Funktion schätzt die erwarteten Kosten des Agenten ab, wenn dieser im Zustand *i* die Aktion *a* wählen würde und sich danach gemäß Strategie π verhalten würde.
- Es gilt einerseits:

$$V^{\pi}(i) = \sum_{j=0}^{n} p_{ij}(\pi(i)) \cdot (c(i, \pi(i)) + V^{\pi}(j))$$

- Und andererseits:

$$Q^{\pi}(i, a) = \sum_{j=0}^{n} p_{ij}(a) \cdot (c(i, a) + Q^{\pi}(j, \pi(j)))$$

- Also:

$$V^{\pi}(i) = Q^{\pi}(i, \pi(i))$$

Die Q-Funktion (3)

Frage: Wie aber kann man eine Funktion Q^π erhalten?

- Die Q-Funktion ist gegenüber der V-Funktion nicht nur über S, sondern über $S \times A$ definiert.
- Sie enthält in diesem Sinne “mehr” Informationen als eine V-Funktion, weil sie auch Bewertungen für alle anderen Aktionen (also insbesondere solche, die nicht gemäß π gewählt würden) enthält.

Die Q-Funktion (3)

Frage: Wie aber kann man eine Funktion Q^π erhalten?

- Die Q-Funktion ist gegenüber der V-Funktion nicht nur über S, sondern über $S \times A$ definiert.
- Sie enthält in diesem Sinne “mehr” Informationen als eine V-Funktion, weil sie auch Bewertungen für alle anderen Aktionen (also insbesondere solche, die nicht gemäß π gewählt würden) enthält.

Antwort: Mit Hilfe der besprochenen Monte-Carlo-Methoden!

- Mittels “Roll-Outs” von Trajektorien, Ermittlung der entstehenden Kosten und Verrechnung gemäß stochatischer Approximation.
- Anstatt aber im Rahmen der MC-Strategiebewertung nur Aktionen gemäß π zu wählen, kann man auch gestatten, bei der ersten auszuführenden Aktion jede mögliche Aktion zu betrachten.

MC-Strategiebewertung für Q-Funktionen

Algorithmus: MC-Strategiebewertung für Q-Funktionen

Es sei $Q_0^\pi(i, a) = 0 \, \forall i \in S, a \in A$. Für jeden Zustand i und jede Aktion $a \in A$ werden durch den Agenten wiederholt Trajektorien in Interaktion mit der Umgebung abgelaufen, bei denen nach der Ausführung von a in i alle Aktionen gemäß der Strategie π gewählt werden.

MC-Strategiebewertung für Q-Funktionen

Algorithmus: MC-Strategiebewertung für Q-Funktionen

Es sei $Q_0^\pi(i,a) = 0 \, \forall i \in S, a \in A$. Für jeden Zustand i und jede Aktion $a \in A$ werden durch den Agenten wiederholt Trajektorien in Interaktion mit der Umgebung abgelaufen, bei denen nach der Ausführung von a in i alle Aktionen gemäß der Strategie π gewählt werden.

Dann wird die Wertfunktion wie folgt aktualisiert:

$$Q_t^\pi(i,a) = Q_{t-1}^\pi(i,a) + \alpha_t(g(i,a,t) - Q_{t-1}^\pi(i,a))$$

wobei die Gesamtkosten
$g(i,a,t) = c(s_0) + c(s_1) + \ldots + c(s_{N-1}), \quad s_0 = i$ sind und wobei
z.B. $\alpha_t = 1/t$ verwendet werden kann.

Gierige Strategieverbesserung ohne Modell

Frage: Was haben wir hierdurch gewonnen?

Gierige Strategieverbesserung ohne Modell

Frage: Was haben wir hierdurch gewonnen?

Antwort: Wir können π ohne Kenntnis des Modells verbessern!

- Erinnerung: Gierige Strategieverbesserung auf Basis von $V^{\pi}(i)$ erfordert Kenntnis des Modells:

$$\pi'(i) = \arg\min_{a \in A(i)} \sum_{j=1}^{n} p_{ij}(a)(c(i,a) + V^{\pi}(j))$$

Gierige Strategieverbesserung ohne Modell

Frage: Was haben wir hierdurch gewonnen?

Antwort: Wir können π ohne Kenntnis des Modells verbessern!

- Erinnerung: Gierige Strategieverbesserung auf Basis von $V^\pi(i)$ erfordert Kenntnis des Modells:

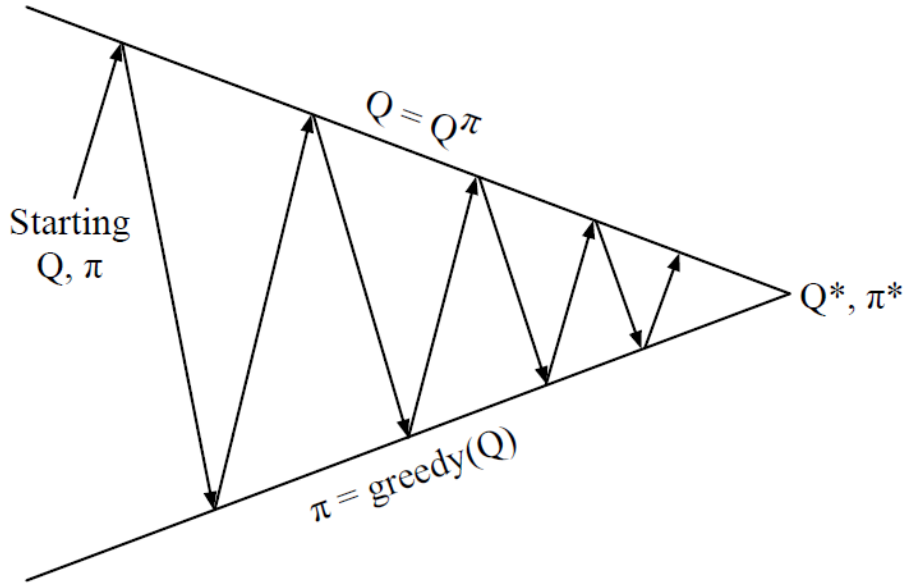
$$\pi'(i) = \arg\min_{a \in A(i)} \sum_{j=1}^{n} p_{ij}(a)(c(i,a) + V^\pi(j))$$

- Die gierige Strategieverbesserung auf Basis von $Q^\pi(i,a)$ kann hingegen modellfrei erfolgen:

$$\pi'(i) = \arg\min_{a \in A(i)} Q^\pi(i,a)$$

  - $\pi'(i)$ ist die gierige Aktion im Zustand i.

Verallgemeinerte Strategieiteration mit Q-Funktionen



Zweiphasiges Verfahren:

- Strategiebewertung mit Monte-Carlo, d.h. $Q = Q^\pi$
- Strategieverbesserung dank Q-Funktionen nun ohne Modell möglich
- Problem (weiterhin): Die Zwischenbewertungen werden eigentlich gar nicht (zumindest nicht 100% exakt) benötigt.
  - Hier wird eventuell zu viel Zeit auf eine zu exakte Lösung verschwendet.
    $\Rightarrow$ In Kürze mehr dazu ...

Modellfreies Lernen

Überblick

1. Motivation
2. Zustands-Aktions-Wertfunktionen
3. Exploration
4. Optimistische Strategieiteration

Exploration vs. Exploitation

Problemstellung:

- Bei vielen Problemen (insbesondere bei realen Systemen) kann man das zu steuernde System nicht einfach in einen bestimmten Zustand bringen, um von dort eine Trajektorie zu starten
  - Beispiele: Roboter, Inverses Pendel
- Man kann nur bestimmte Startzustände einnehmen.
- Viele Zustände sind im funktionierenden Fall “uninteressant”, aber die “interessanten” Zustände müssen im Lernverlauf irgendwie erreicht werden.

Exploration vs. Exploitation

$\Rightarrow$ Erforderlich ist das “Besuchen” möglichst vieler interessanter Zustände durch zufälliges Erkunden

Definition (Exploration)

Unter Exploration verstehen wir das gezielte Abweichen des Agenten von seiner bis zum aktuellen Zeitpunkt bekannten, bestmöglichen (gierigen) Strategie.

Beispiel

Beispiel: Notwendigkeit der Exploration

- zwei Türen zur Auswahl
- linke Tür geöffnet, Kosten von 0 erhalten $V(left) = 0$
- rechte Tür geöffnet, Kosten von -1 erhalten $V(right) = -1$
- rechte Tür geöffnet, Kosten von -3 erhalten $V(right) = -2$
- rechte Tür geöffnet, Kosten von -2 erhalten $V(right) = -2$
- ...
- Sind Sie sicher, die beste Tür ermittelt zu haben?

"Behind one door is tenure - behind the other is flipping burgers at McDonald's."

Explorationsansätze

- Ziel: Finden der möglichst optimalen Strategie
- Grundsätzliches Dilemma: Exploration vs. Exploitation
  “Neugier” vs. “greedy”/optimal bezüglich aktueller Kostenschätzung
- Implementierungsideen:

Explorationsansätze

- Ziel: Finden der möglichst optimalen Strategie
- Grundsätzliches Dilemma: Exploration vs. Exploitation
  “Neugier” vs. “greedy”/optimal bezüglich aktueller Kostenschätzung
- Implementierungsideen:
  - Perioden der Exploration (rein zufälliges Verhalten) und Perioden der gierigen Auswertung abwechseln
  - In jedem Schritt mit einer bestimmten Wahrscheinlichkeit von der gierigen Aktion abweichen.
  - Einen Zufallswert zur Pfadkostenfunktion hinzuaddieren
  - Abhängig von “Eindeutigkeit der Entscheidung” und im Laufe des Lernens abnehmender Explorationsfreude (Temperatur)
    $\Rightarrow$ Boltzmann-Verteilung

ε-gierige Aktionsauswahl (1)

Kernidee:

- Alle m möglichen Aktionen ($|A| = m$) werden in jedem Schritt mit einer Wahrscheinlichkeit größer null ausgewählt.
- Mit Wahrscheinlichkeit $1 - \varepsilon$ wird die bestmögliche (gierige) Aktion gewählt, mit Wahrscheinlichkeit ε wird eine zufällige Aktion ausgeführt.
- Ist die einfachste Idee, kontinuierliche Exploration zu gewährleisten.
- Typisch: Explorationswahrscheinlichkeit wird im Laufe der Zeit sukzessive reduziert.

ε-gierige Aktionsauswahl (2)

Bemerkung:

- Da die Strategie nun Wahrscheinlichkeiten liefert, ist sie nicht mehr als $\pi : S \to A$, sondern als Wahrscheinlichkeitsverteilung $\pi : S \times A \to [0, 1]$ definiert.

ε-gierige Aktionsauswahl (2)

Bemerkung:

- Da die Strategie nun Wahrscheinlichkeiten liefert, ist sie nicht mehr als $\pi : S \to A$, sondern als Wahrscheinlichkeitsverteilung $\pi : S \times A \to [0, 1]$ definiert.

Definition (ε-gierige Strategie)

Wenn wir im aktuellen Zustand i und bei gegebener Wertfunktion V bzw. Q die bestmögliche (gierige) Aktion als a^* bezeichnen, also

$$a^* = \arg\min_{a \in A} \sum_{j=0}^{n} p_{ij}(a)(c(i,a) + V(j)) \text{ bzw. } a^* = \arg\min_{a \in A} Q(i,a)$$

so ist die **ε-gierige Strategie** definiert als

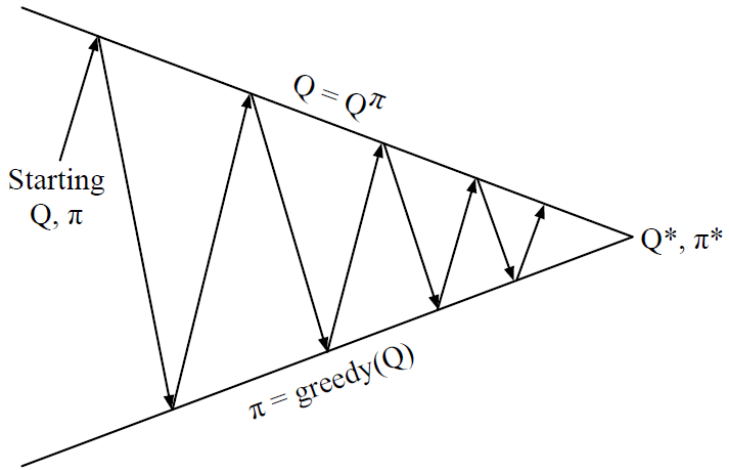
$$\pi(s,a) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{m} & \text{if } a = a^* \\ \frac{\varepsilon}{m} & \text{else} \end{cases}$$

ε-gierige Aktionsauswahl (3)

Theorem:

Für jede ε-gierige Strategie π, für die mittels Strategiebewertung die Zustands-Aktions-Funktion Q^π ermittelt worden ist, ist die ε-gierige Strategie, die aus gieriger Auswertung von Q^π hervorgeht, eine Verbesserung.
Formal: $V^{\pi'}(i) \leq V^\pi(s)$.

Also: Diese Abbildung gilt auch bei Strategien, die explorieren.



Konvergenz bei Exploration (1)

Bemerkungen:

- Auf der vorigen Folie war nur die Rede von einer Verbesserung.
- Wie sieht es aber mit Konvergenz aus: Konvergiert das generalisierte Strategieiterationsverfahren auf Basis von ε-gierigen Strategien auch gegen die optimale Strategie?

Konvergenz bei Exploration (1)

Bemerkungen:

- Auf der vorigen Folie war nur die Rede von einer Verbesserung.
- Wie sieht es aber mit Konvergenz aus: Konvergiert das generalisierte Strategieiterationsverfahren auf Basis von ε-gierigen Strategien auch gegen die optimale Strategie?

Definition (GLIE)

Eine Strategie heißt GLIE (Greedy in the Limit with Infinite Exploration), wenn

- Alle Zustands-Aktions-Paare unendlich oft besucht werden: $\lim_{t\to\infty} N_t(i,a) = \infty$
- Die Strategie zu einer deterministischen Strategie konvergiert:

$$\lim_{t\to\infty} \pi_t(i,a) = \begin{cases} 1 & \text{if } a = \arg\min_{b\in A(i)} Q_t(i,b) \\ 0 & \text{else} \end{cases}$$

Konvergenz bei Exploration (2)

Bemerkungen zu GLIE:

- Eine ε-Strategie ist GLIE, wenn ε gegen 0 geht gemäß $\varepsilon_t = \frac{1}{t}$.
- Die durch den GLIE-MC-Algorithmus ermittelte Zustands-Aktions-Wertfunktion konvergiert gegen die optimale Wertfunktion, $Q(i, a) \rightarrow Q^*(i, a)$.
  - sh. in einigen Folien

Boltzmann-Exploration

Auswahl der Aktion gemäß einer 'Boltzmann'-Verteilung

$$\frac{e^{-Q(i,a)/T}}{\sum_{v \in A(i)} e^{-Q(i,v)/T}}$$

mit T als "Temperatur"-Parameter:

- Wenn T klein ist, wird die Aktion mit dem kleinsten Q-Wert nahezu mit Wahrscheinlichkeit 1 aufgerufen.
- oberer Term: exponentielle Gewichtung
- unterer Term: Normierung

Boltzmann-Exploration

Auswahl der Aktion gemäß einer 'Boltzmann'-Verteilung

$$\frac{e^{-Q(i,a)/T}}{\sum_{v \in A(i)} e^{-Q(i,v)/T}}$$

mit T als "Temperatur"-Parameter:

- Wenn T klein ist, wird die Aktion mit dem kleinsten Q-Wert nahezu mit Wahrscheinlichkeit 1 aufgerufen.
- oberer Term: exponentielle Gewichtung
- unterer Term: Normierung

Beispiel: $Q('a') = 0.3$, $Q('b') = 0.5$, $T = 0.1$ bzw. 1

$$e^{-0.3/0.1} = 0.049; e^{-0.5/0.1} = 0.006; p('a') = .89; p('b') = .11$$

$$e^{-0.3/1} = 0.74; e^{-0.5/1} = 0.6; p('a') = .55; p('b') = .45$$

Modellfreies Lernen

Überblick

1. Motivation
2. Zustands-Aktions-Wertfunktionen
3. Exploration
4. Optimistische Strategieiteration

Optimistische Strategieiteration (1)

Erinnerung:

1. Wir müssen eine Möglichkeit finden, eine gierige Strategieverbesserung durchzuführen (auch ohne Modell).
2. Wir müssen uns die Frage stellen, wie häufig wir die Strategie anpassen, also durch gierige Strategieverbesserung modifizieren.
   - Soll wirklich gewartet werden, bis durch die Strategiebewertung die perfekte V^{π} vorliegt?

Optimistische Strategieiteration (1)

Erinnerung:

1. Wir müssen eine Möglichkeit finden, eine gierige Strategieverbesserung durchzuführen (auch ohne Modell).
2. Wir müssen uns die Frage stellen, wie häufig wir die Strategie anpassen, also durch gierige Strategieverbesserung modifizieren.
   - Soll wirklich gewartet werden, bis durch die Strategiebewertung die perfekte V^{π} vorliegt?

Zum Punkt 1.: Monte-Carlo-basierte Strategiebewertung
Zum Punkt 2.: Auf die perfekte Bewertung der aktuellen Strategie zu warten, würde bedeuten, unendlich viele Trajektorien laufen zu lassen, um so den Erwartungswert der Kosten, die von einem Zustand aus auftreten, exakt zu ermitteln.

Optimistische Strategieiteration (2)

Kernidee: Beendigung der Strategiebewertung bereits nach einer endlichen Anzahl k von Trajektorien (Rollouts).

- Klar: Die so gewonnene Bewertung Q^{π} der aktuellen Strategie ist nicht perfekt.

Frage: Ist dann trotzdem garantiert, dass die generalisierte Strategieiteration sukzessive verbesserte Strategien generiert?

Optimistische Strategieiteration (2)

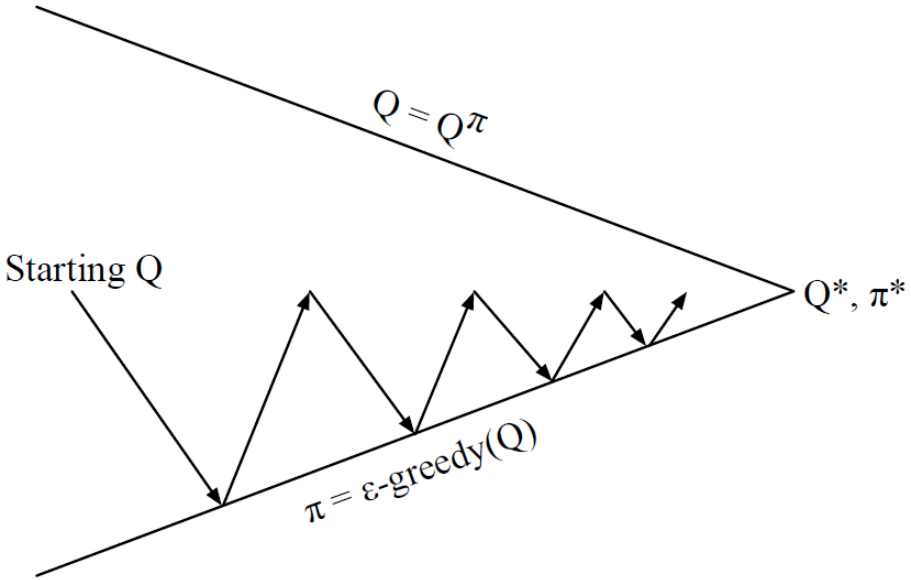
Kernidee: Beendigung der Strategiebewertung bereits nach einer endlichen Anzahl k von Trajektorien (Rollouts).

- Klar: Die so gewonnene Bewertung Q^{π} der aktuellen Strategie ist nicht perfekt.

Frage: Ist dann trotzdem garantiert, dass die generalisierte Strategieiteration sukzessive verbesserte Strategien generiert? (Erfreuliche) Antwort: Ja.

- sh. nächste Folie

Optimistische Strategieiteration (3)



Zweiphasiges Verfahren:

- Strategieverbesserung dank Q-Funktionen nun ohne Modell möglich
- Strategiebewertung mit Monte-Carlo, d.h. $Q \approx Q^\pi$
- Problem gelöst: Die Zwischenbewertungen werden gar nicht (zumindest nicht 100% exakt) benötigt.
  - Hier wird nun nicht mehr zu viel Zeit auf eine zu exakte Lösung verschwendet.

Konvergenz bei Exploration (3)

GLIE-MC-Algorithmus

- Erlebe Trajektorie t mit π: $\{s_0, a_0, c_0, \ldots, S_N\}$
- Für jeden Zustand s_k und Aktion a_k in der Trajektorie:

$$N(s_k, a_k) \leftarrow N(s_k, a_k) + 1$$

$$Q_t(s_k, a_k) \leftarrow Q_{t-1}(s_k, a_k) + \frac{1}{N(s_k, a_k)}(g(s_k, t) - Q_{t-1}(s_k, a_k))$$

- Führe Strategieverbesserungsschritt durch auf Basis der neuen Q-Funktion Q_t gemäß:

$$\varepsilon \leftarrow \frac{1}{t}$$

$$\pi \leftarrow \varepsilon\text{-greedy}(Q_t)$$

Zusammenfassung

Erreichte Ziele:

- Strategieverbesserung ohne Modell
- Strategiebewertung nur approximativ (nicht exakt, was “unendlich” lange dauern würde)

Ausblick:

- die modellfreie Strategieverbesserung funktioniert soweit nur in Kombination mit MC-basierter Strategiebewertung
- Ziel: zeitliche Differenzverfahren (TD) wieder in der Strategiebewertung integrieren und weiterhin eine modellfreie Strategieverbesserung gewährleisten