Reinforcement Learning

Abe Vos

April 2022

Introductie

Korte terugblik

- Supervised learning
- Input/output paren (gelabelde data)
- Generaliseer naar nieuwe situaties

Reinforcement Learning

- Leren door interactie (trial & error)
- Maximaliseer beloning
- Exploration & exploitation

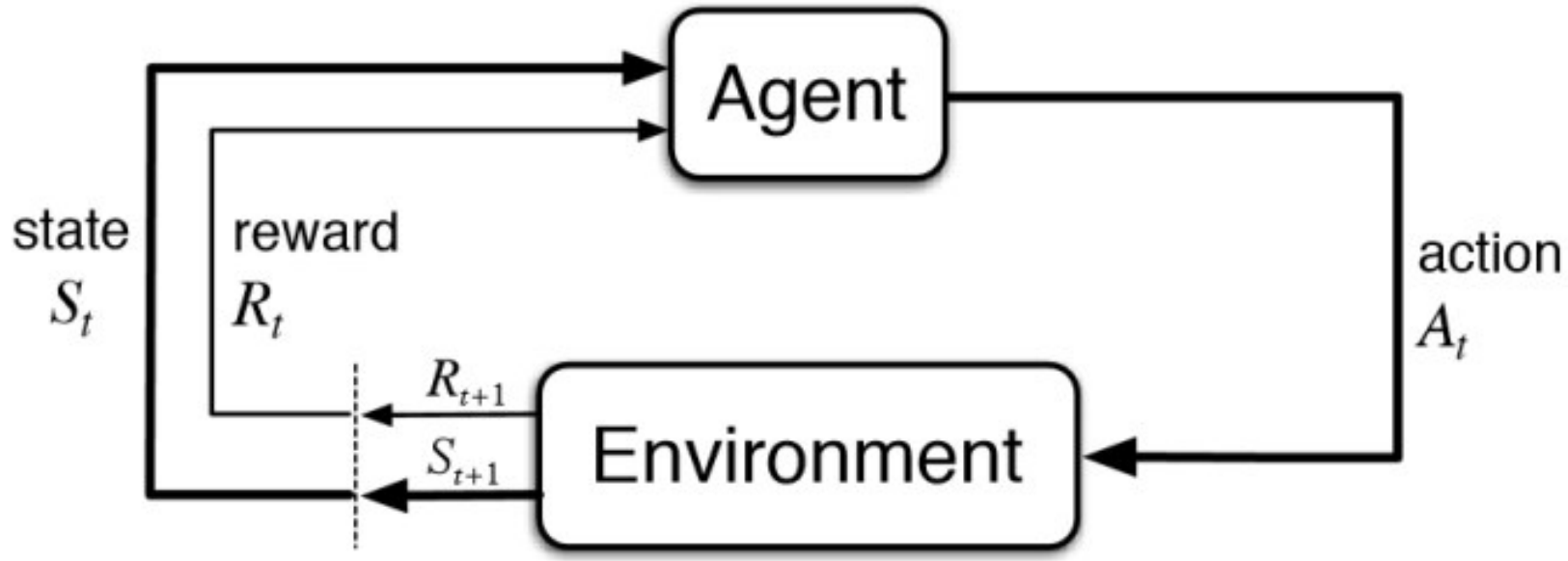
Reinforcement Learning

Eigenschappen

- *Agent* gebruikt *interactie*
- *Agent* bereikt een *doel*
- Houd rekening met *onzekerheid*

Agent en Interactie

- Agent voert acties uit
- Kies acties a.d.h.v. observaties
- Acties veranderen omgeving $\rightarrow$ nieuwe observatie


Agent
state
S_t
reward
R_t
action
A_t
R_{t+1}
S_{t+1}
Environment

Doel

- Gewenste uitkomst
- Niet altijd vanzelfsprekend

CoastRunners

Onzekerheid

- Effecten van acties zijn niet altijd voorspelbaar
- Observaties van de staat zijn niet altijd volledig

Figure 1: Acties in DND hebben een onzekere uitkomst

Figure 2: Strategiespellen hebben onzekere observaties

Intermezzo: kansrekening

Randomness

- Maakt een situatie onzeker
- Spelers gebruiken kansrekening om te redeneren over onzekerheid

Stochastische variabele

- Variabele met onzekere waarde
- Functies van stochasten zijn ook stochastisch
- Notatie met hoofdletter (bijv.: X)
  - Uitkomst is kleine letter (bijv.: x)

Discreet en continu

- Discrete stochasten: telbaar
  - `random.randint()`
- Continue stochasten: ontelbaar
  - `random.random()`

Kanswaardes

- Elke uitkomst van een stochast heeft een kanswaarde: $p(X = x)$
- Kansverdeling: functie van uitkomst naar kanswaarde
- Twee voorwaarden:
  - $p(X = x) \geq 0, \forall x \in \mathcal{X}$
  - $\sum_{x \in \mathcal{X}} p(X = x) = 1$

Dobbelsteen

x	$p(X = x)$
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

Normaalverdeling

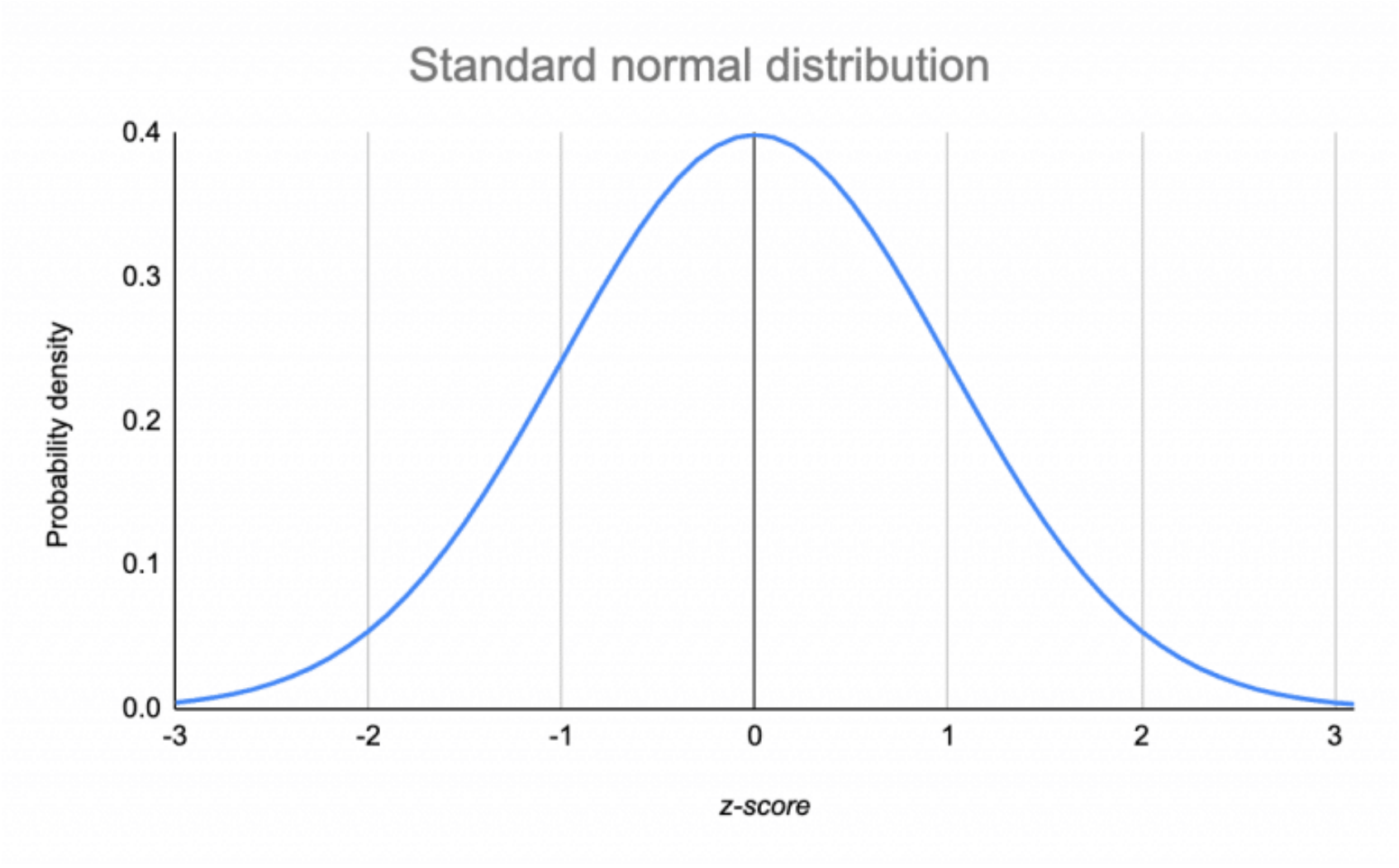


Figure 3: $p(X = x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Steekproeven/sampling

- Observatie van een stochast
- `random.random()` geeft een steekproef
  - Dobbelsteen: `random.randint(1,7)`
  - Normaalverdeling: `random.gauss(mu, sigma)`
- Sommige verdelingen zijn moeilijker

Simultane kansverdeling

- Kansverdeling voor twee of meer variabelen $p(X, Y)$
- "Wat is de kans om twee zessen te gooien met twee dobbelstenen?"
  - Niet hetzelfde als de kans om een 12 te gooien met twee dobbelstenen!
- Kansverdeling van twee onafhankelijke gebeurtenissen: $p(X, Y) = p(X)p(Y)$

Voorwaardelijke kansverdeling

- Kansverdeling voor een stochast nadat we een andere stochast hebben geobserveerd: $p(X|Y)$
- "Wat is de kans om een totaal van zeven te gooien met twee dobbelstenen, gegeven dat de eerste dobbelsteen een 4 rolde?"

Onafhankelijke gebeurtenissen

- Uitkomst van een stochast heeft geen invloed op andere stochasten
- $p(X|Y) = p(X)$
- Twee dobbelstenen
  - Eerste worp heeft geen invloed op tweede worp
  - Eerst worp heeft wel invloed op de som van uitkomsten

Eigenschappen van stochasten

- Verwachte waarde
- Variantie/standaard deviatie

Verwachte waarde

- $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} xp(X = x)$
- Gemiddelde uitkomst over een grote hoeveelheid uitkomsten
- Verwachting hoeft geen mogelijke uitkomst te zijn
  - Voorbeeld: verwachte uitkomst van een dobbelsteen is 3.5

Variantie

- $\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
  - $\text{Var}(X) = \sum_{x \in \mathcal{X}} p(X = x)(x - \mathbb{E}[X])^2$
- Gemiddeld verschil met gemiddelde

Scope

Elementen van de agent

- Beleid
- Beloning
- Waarde funcie

Beleid

- Gekozen actie a voor elke staat s: $\pi(a|s)$
- Kan een kansverdeling zijn
- Stochastische acties zijn nuttig voor verkenning

Beloning

- Doel is uitgedrukt als beloningssignaal
- Beloningen bij elke staat
- Totale beloning = som van alle beloningen

Waarde-functie

- Geeft “waarde” voor elke staat
- Waarde is de verwachte totale beloning in de toekomst
- Beloningen zijn direct; waarde is voor de lange termijn
- Waarde wordt geleerd en bepaalt het beleid

Finite Markov Decision Process

- Model voor beslissingsproblemen
- Verzameling van staten $s \in \mathcal{S}$ en acties in die staten $a \in \mathcal{A}(s), \forall s \in \mathcal{S}$
- Overgangsfunctie $p(S_{t+1} = s' | S_t = s, A_t = a)$
- Beloningsfunctie $R(s, a, s')$

Discount factor

- Waarde tussen 0 en 1
- Bepaalt invloed van beloningen in de toekomst op de waarde van het heden
- $\gamma = 1$; beloningen in de toekomst tellen even zwaar mee als directe beloningen
- $\gamma = 0$; alleen de directe beloning doet ertoe
- Hyperparameter in veel methodes

Markov eigenschap

- De toekomst is onafhankelijk van het verleden gegeven de toekomst
- Met een observatie in het heden weten we genoeg om een goede beslissing te maken
  - Kennis van het verleden maakt onze beslissingen in de toekomst niet beter
- $p(S_{t+1}|S_t, S_{t-1}) = p(S_{t+1}|S_t)$
- Dit maakt het probleem een stuk eenvoudiger

Optimalisatie objectief

- Functie die we willen maximaliseren
- Totale beloning: $G = \sum_{t=0}^{\infty} \gamma^t R(S_{t+1}, A_t, S_t)$
  - Elke stap krijgen we een actie van $\pi(A_t|S_t)$
  - En een nieuwe staat van $p(S_{t+1}|A_t, S_t)$
  - G is stochastisch, omdat S_{t+1} dat ook is
- Objectief: verwachte totale beloning $\max_\pi \mathbb{E}[G]$

Dynamic Programming

Wat leren we nou eigenlijk?

- Maximaliseer verwachte totale beloning
- Verwachte totale beloning wordt bepaalt door beleid
- Beleid wordt bepaalt door waarde-functie

Aanpakken van RL

- Maak een schatting van waarde-functie
  - Beleid kiest acties die tot hoogste waarde leiden
- Zoek direct naar een beleid

Policy Iteration

- Bereken de waarde-functie a.d.h.v. een beleid
  - Update daarna het beleid met de berekende waarde-functie
- Policy Evaluation & Policy Improvement
- Deterministisch beleid $a = \pi(s)$

Policy Evaluation benodigdheden

- Arbitraire beleidsfunctie $/pi(s)$
- Arbitraire waarde-functie $V(s)$
  - Als er een eindstaat s_{terminal} bestaat, dan moet $V(s_{\text{terminal}}) = 0$
- Maximale verandering $\Delta = 0$
  - Houdt *convergentie* bij
- Kleine drempelwaarde $\theta > 0$

Policy Evaluation

- Ga elke staat $s \in \mathcal{S}$ langs
- $V_{\text{oud}} \leftarrow V(s)$
- $V(s) \leftarrow \sum_{s',r} p(s', r|s, \pi(s))[r + \gamma V(s')]$
- $V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$
- Update $\Delta \leftarrow \max(\Delta, |V_{\text{oud}} - V(s)|)$
- Herhaal totdat $\Delta < \theta$

Gridworld

- Raster
- Agent kan omhoog, omlaag en opzij bewegen
- Staat is (x, y) positie van agent
- Elke overgang geeft een beloning
  - Acties die de agent van het bord af leiden, krijgen negatieve beloning
  - Eindstaten kunnen met beloningen aangegeven worden

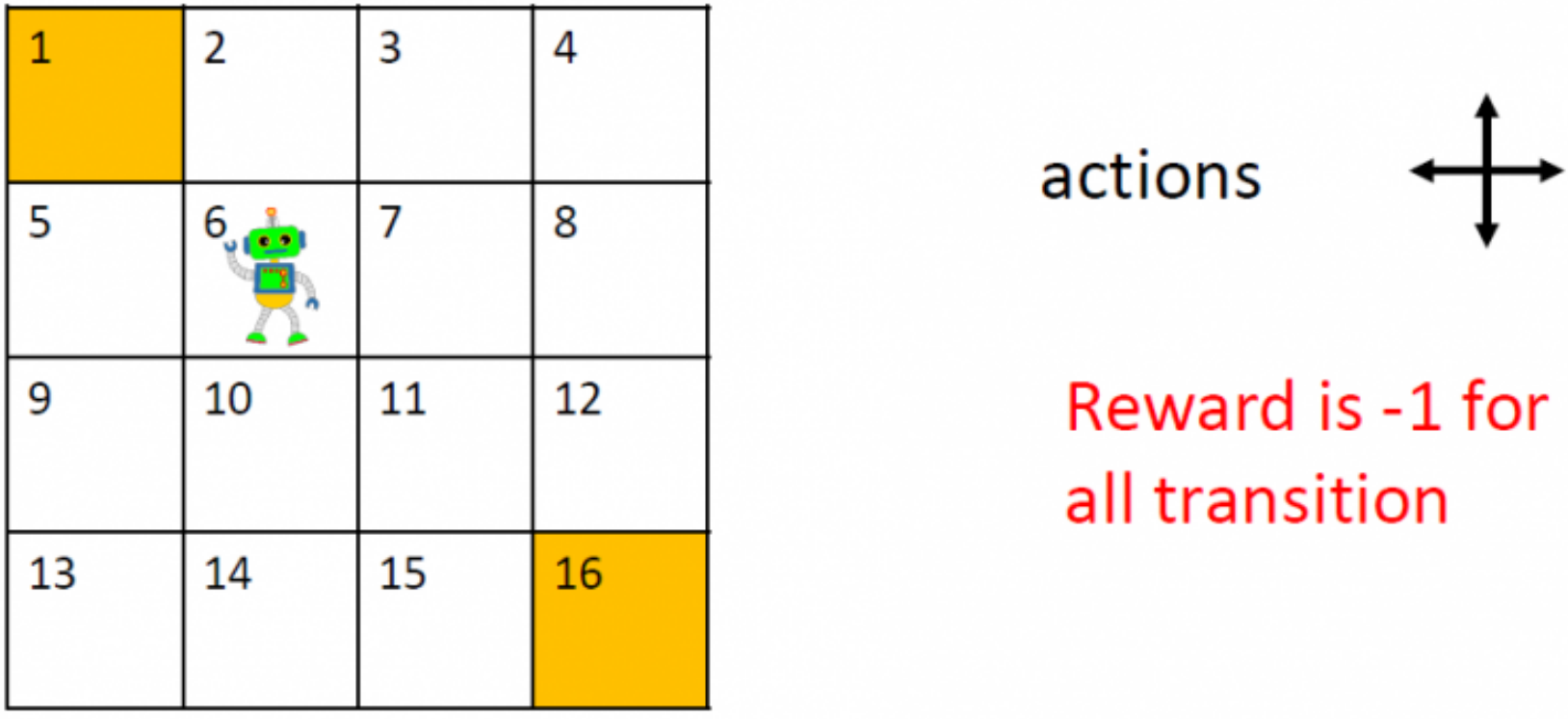


Figure 4: Gridworld voorbeeld

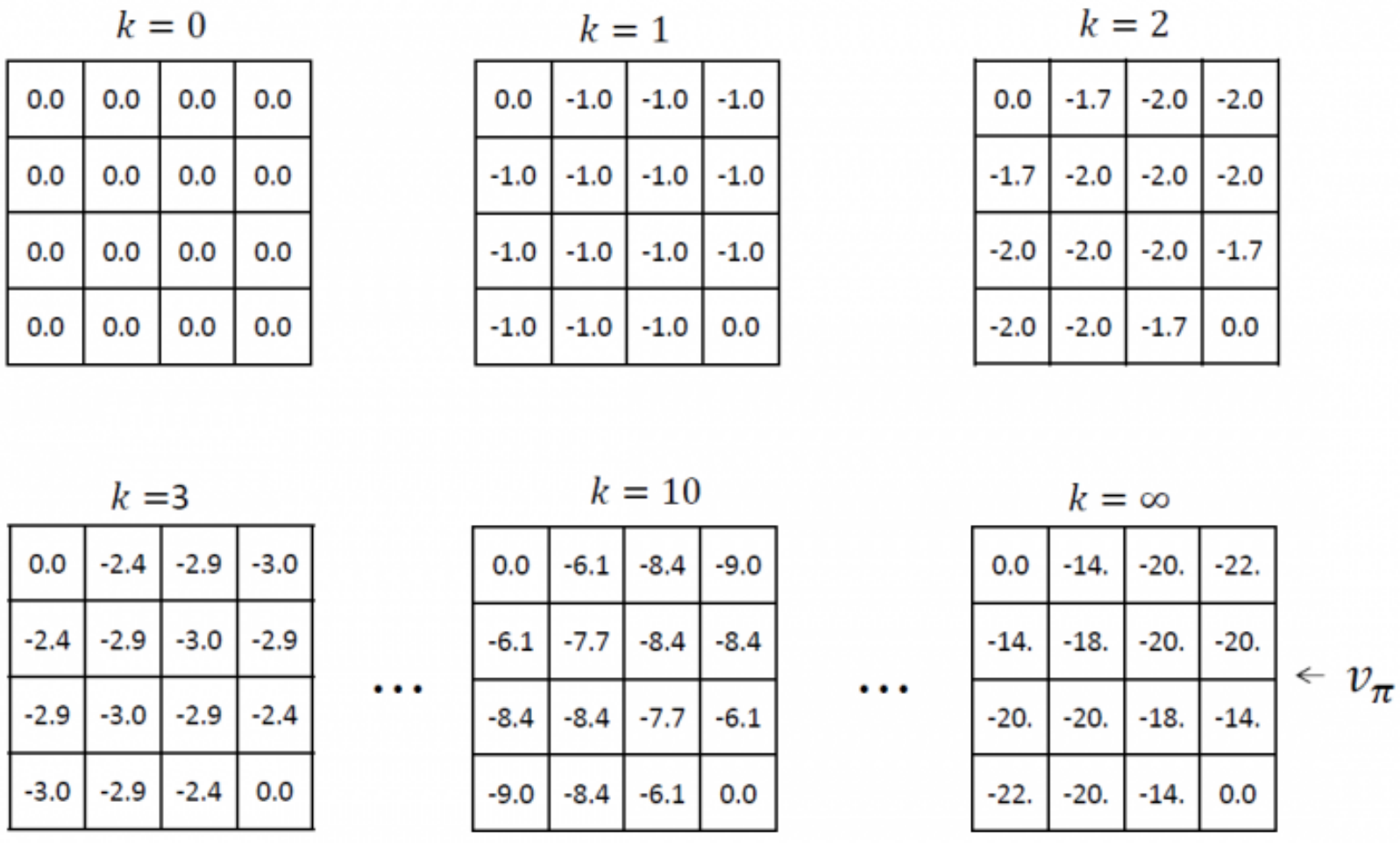


Figure 5: Policy Evaluation

Policy Improvement

- Gebruik $V(s)$ van Policy Evaluation om $\pi(s)$ te updaten
- Voor elke $s \in \mathcal{S}$
  - Sla de actie $\pi(s)$ tijdelijk op
  - Bereken $\sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$ voor elke actie
  - Stel $\pi(s)$ in op de actie met de hoogste waarde

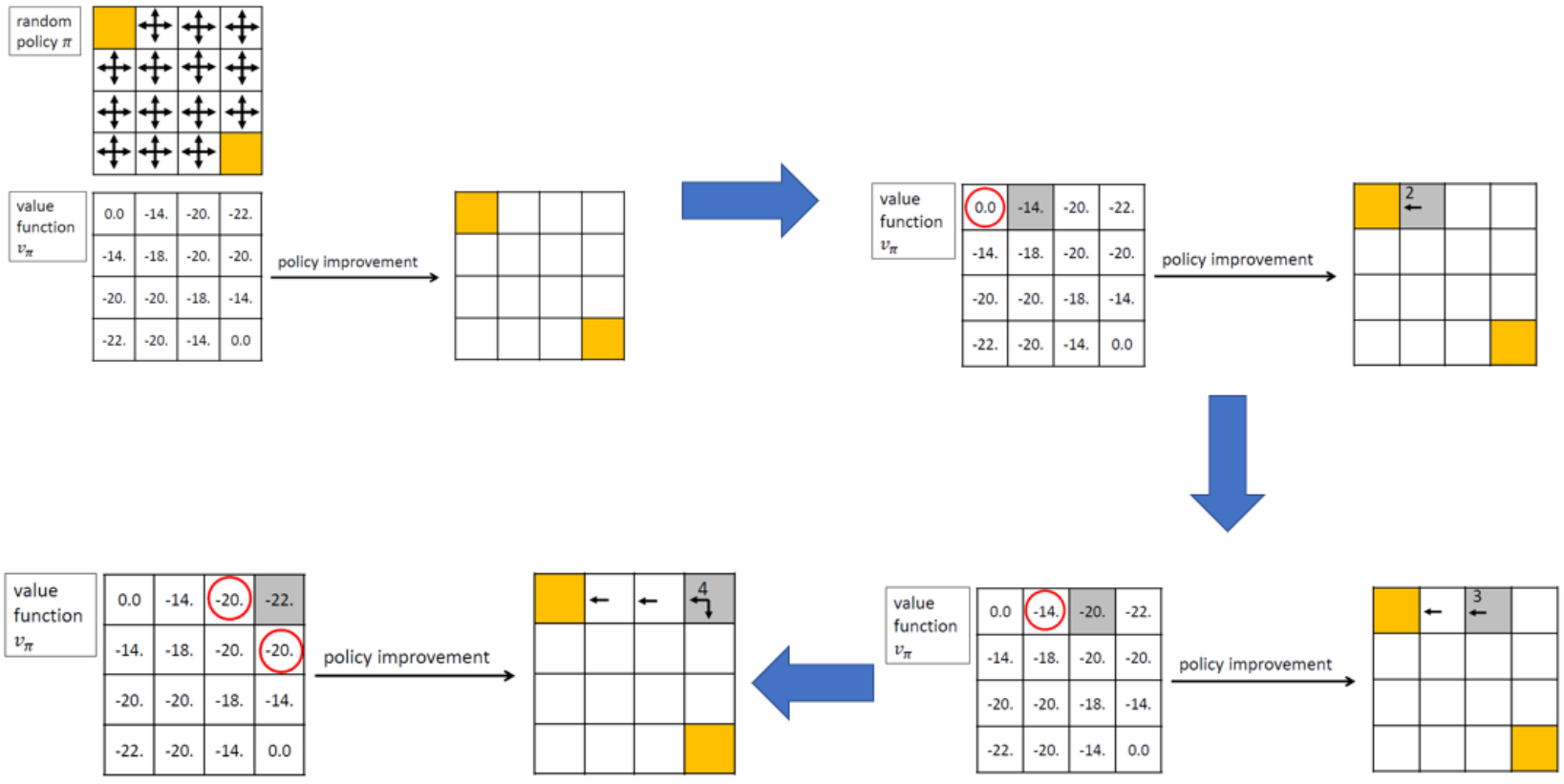


Figure 6: Policy Improvement

Policy Iteration

- Update $V(s)$ met Policy Evaluation
- Update $\pi(s)$ met Policy Improvement
- Herhaal totdat het beleid gelijk blijft voor elke staat
  - Dit gebeurt wanneer het beleid optimaal is

$$\pi_0 \xrightarrow{E} v_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} v_{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} \ldots \xrightarrow{I} \pi_* \xrightarrow{E} v_*$$

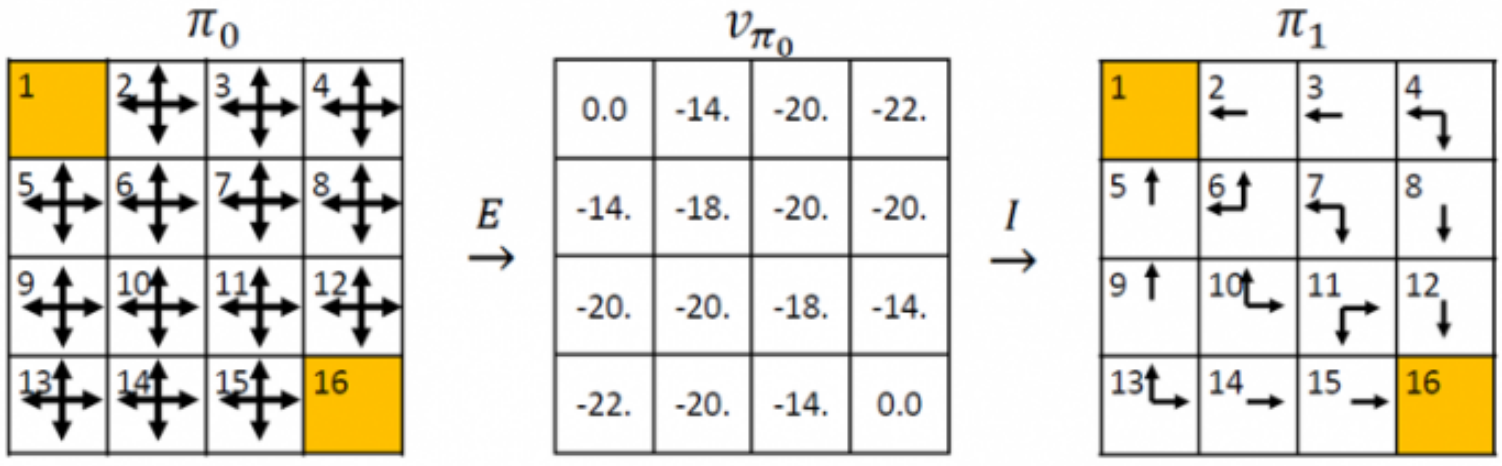


Figure 7: Policy Iteration

Value Iteration

- Policy Evaluation kan traag zijn
- Combineer Evaluation en Improvement stappen
- Vervang de update stap van policy evaluation
- $V(s) \leftarrow \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$
- 1 “Evaluation” en 1 “Improvement”

Tekortkomingen

- Zijn afhankelijk van de overgangsfunctie $p(s', r|s, a)$
- Moeten *alle* staten langs gaan
  - Vier-op-een-rij heeft 4531985219092 staten

Huiswerk

Github

- https://github.com/AbeVos/self-learning-ai-2022
- Huiswerk en slides
- Elk college geupdate

Eerste opdracht

- Programmeren en opdrachten
- Instructies voor inleveren in de opdracht
- Deadline 12 mei, 23:59u

Vragen/ziekmeldingen/etc.

- abe_vos@msn.com
- Discord?