MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Πανεπιστήμιο Κύπρου - Τεχνητή Νοημοσύνη

MAI612 - ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Διάλεξη 17: Διαδικασίες Αποφάσεων Markov και Δυναμικός Προγραμματισμός

Βασίλης Βασιλειάδης, PhD

Χειμερινό Εξάμηνο 2022/23

CYENS CENTRE OF EXCELLENCE

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Διάλεξη 17: Διαδικασίες Αποφάσεων Markov και Δυναμικός Προγραμματισμός

Μαθησιακά αποτελέσματα

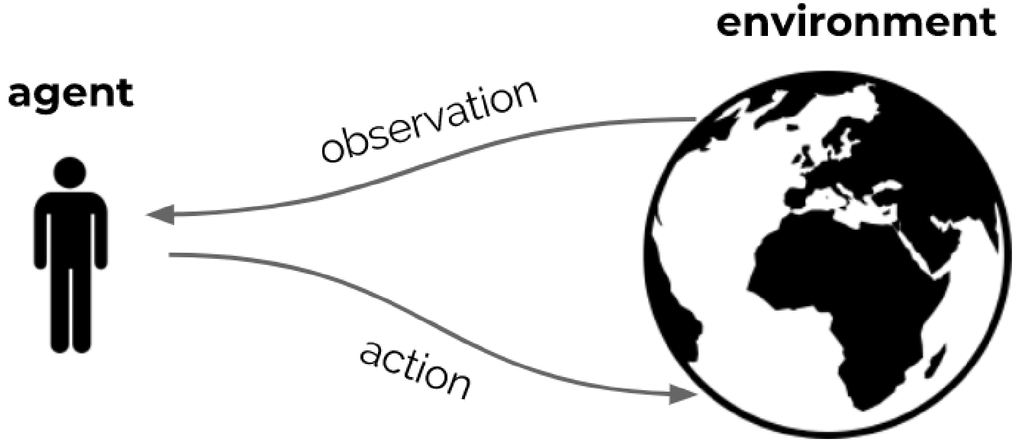
Θα καταλάβετε:

1. Πώς να επισημοποιηθεί το πρόβλημα RL χρησιμοποιώντας το πλαίσιο των διαδικασιών λήψης αποφάσεων Markov (ΔΑΜ).
2. Τι είναι οι συναρτήσεις αξίας και πώς σχετίζονται με τους στόχους σε ένα ΔΑΜ.
3. Δύο κατηγορίες προβλημάτων RL: πρόβλεψη και έλεγχος.
4. Πώς να χρησιμοποιήσετε τις εξισώσεις Bellman για την επίλυση προβλημάτων RL.
5. Δυναμικός προγραμματισμός και αλγόριθμοι επανάληψης αξίας και επανάληψης πολιτικής για την εξεύρεση λύσεων σε ΔΑΜ.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επανάληψη



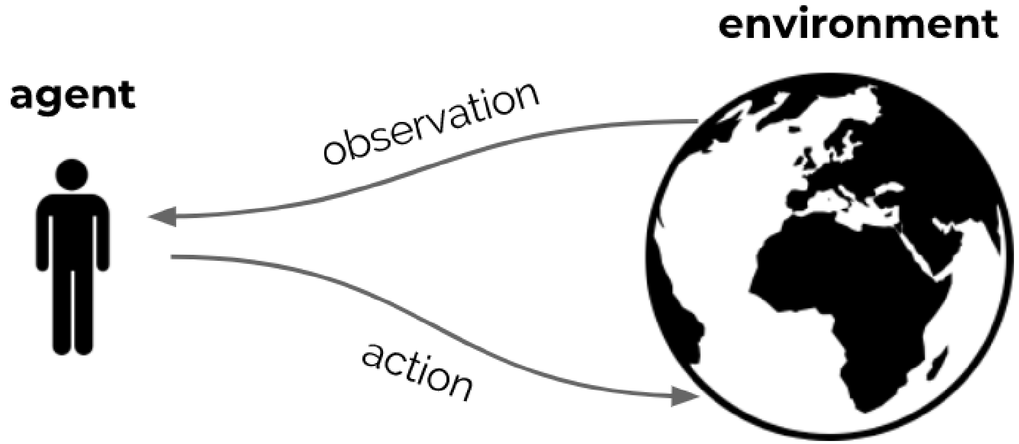
- Ενισχυτική μάθηση είναι η επιστήμη της μάθησης για τη λήψη αποφάσεων
- Οι πράκτορες μπορούν να μάθουν μια **πολιτική**, μια **συνάρτηση αξίας** ή/και ένα **μοντέλο**
- Το γενικό πρόβλημα περιλαμβάνει τη συνεκτίμηση του **χρόνου** και των **συνεπειών**
- Οι αποφάσεις επηρεάζουν την **ανταμοιβή**, το **κράτος πράκτορα** και την κατάσταση του **περιβάλλοντος**
- Η μάθηση είναι **ενεργή**: οι αποφάσεις έχουν αντίκτυπο στα δεδομένα

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επισημοποίηση της διεπαφής ΕΜ



- Θα συζητήσουμε μια μαθηματική διατύπωση της αλληλεπίδρασης παράγοντα-περιβάλλοντος
- Αυτό ονομάζεται **Διαδικασία Απόφασης Markov (ΔΑΜ)**
- Μας δίνει τη δυνατότητα να μιλήσουμε ξεκάθαρα για τον **στόχο** και **τον τρόπο επίτευξής του**

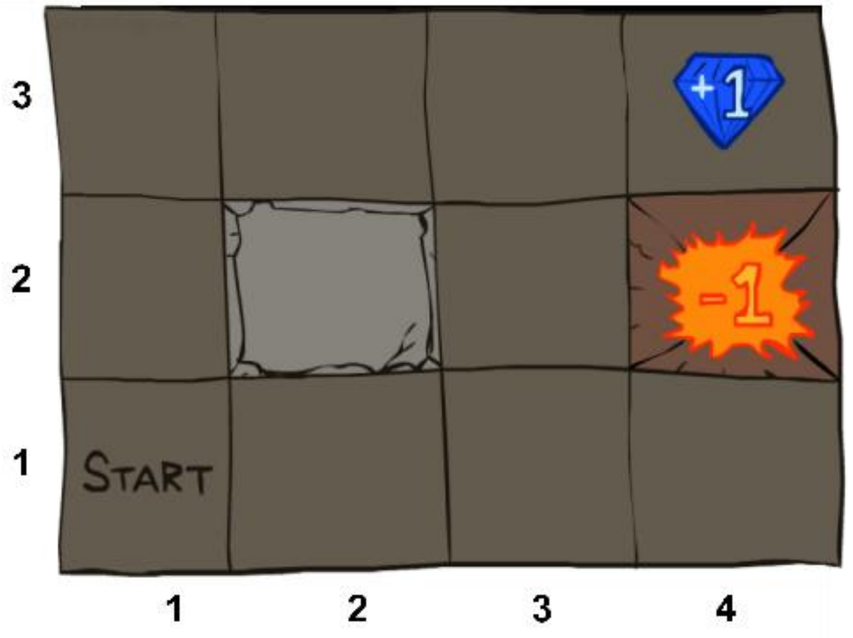
Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα: Ο Κόσμος του πλέγματος

- Ένα πρόβλημα που μοιάζει με λαβύρινθο
  - Ο πράκτορας ζει σε ένα πλέγμα
  - Τοίχοι μπλοκάρουν την πορεία του πράκτορα
- Θορυβώδης κίνηση: οι ενέργειες δεν πηγαίνουν πάντα όπως έχει προγραμματιστεί
  - 80 % του χρόνου, η δράση Βόρεια παίρνει τον πράκτορα Βόρεια (αν δεν υπάρχει τοίχος εκεί)
  - 10 % του χρόνου, αντί για βορεια ο πράκτορα πάει δυτικά και 10 % Ανατολικά
  - Αν υπάρχει τοίχος προς την κατεύθυνση που θα είχε πάρει ο πράκτορας, ο πράκτορας παραμένει στη θέση του.
- Ο πράκτορας λαμβάνει ανταμοιβές κάθε φορά που
  - Μικρή «ζωντανή» ανταμοιβή κάθε βήμα (μπορεί να είναι αρνητικό)
  - Μεγάλες ανταμοιβές έρχονται στο τέλος (καλό ή κακό)
- Ο στόχος μας: μεγιστοποιήστε το άθροισμα των ανταμοιβών



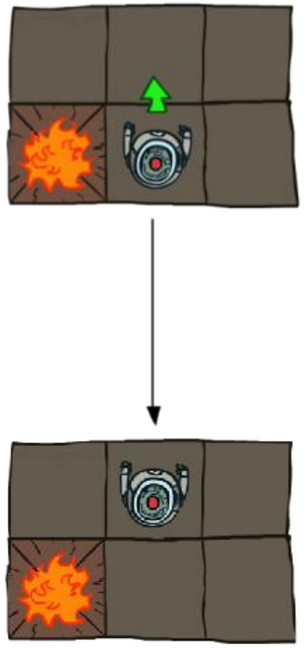
Διαφάνειες βασισμένες στο: *UC Berkeley CS188 – Intro to AI course*

Co-financed by the European Union
Connecting Europe Facility

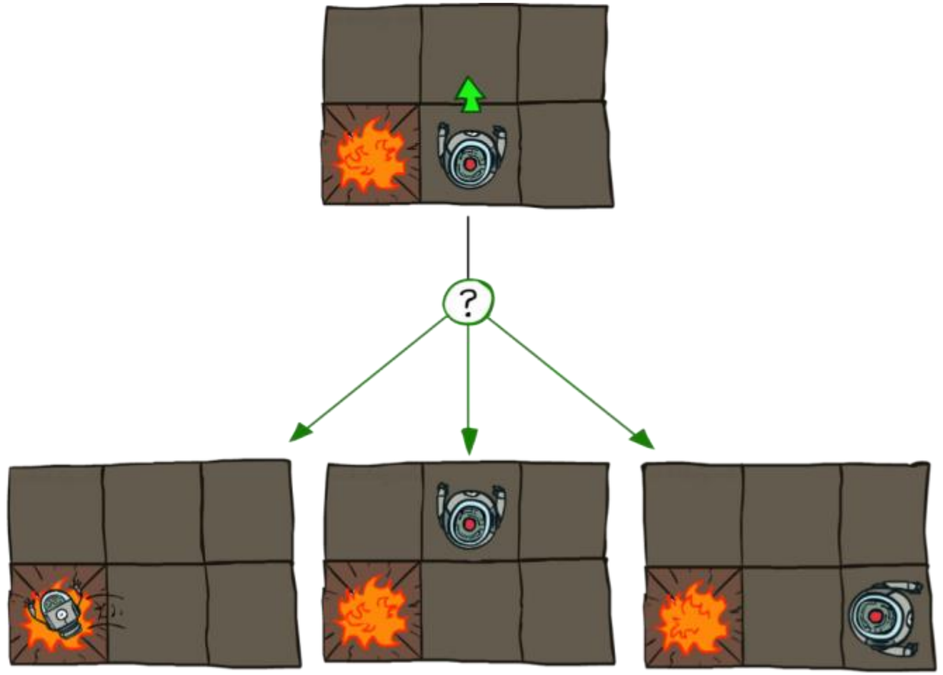
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα: Δράσεις στο κόσμος του πλέγματος

Ντετερμινιστικός κόσμος πλέγματος



Στοχαστικός κόσμος πλέγματος



Διαφάνειες βασισμένες στο: ***UC Berkeley CS188 – Intro to AI course***

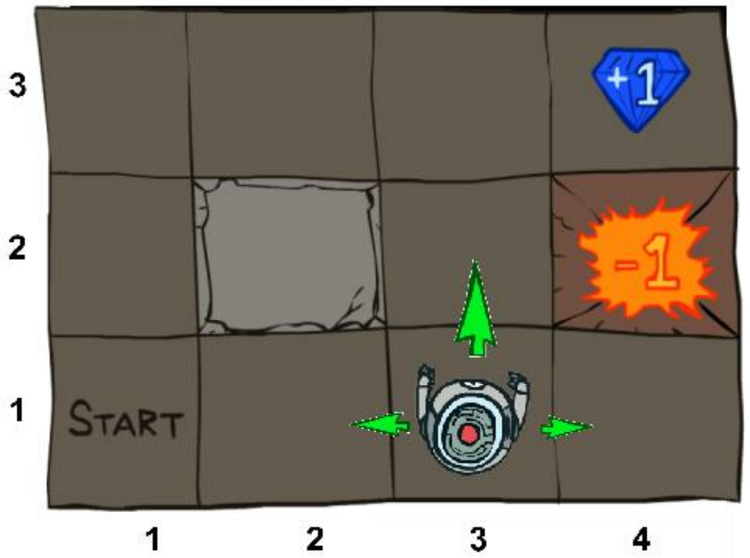
Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Διαδικασία Απόφασης Markov Ορισμός της διαδικασίας λήψης αποφάσεων

Η **διαδικασία λήψης αποφάσεων** Markov είναι μια πλειάδα (*S*, *A*, *T*, *r*) όπου:

- *S* είναι το σύνολο όλων των πιθανών καταστάσεων
- *A* είναι το σύνολο όλων των δυνατών ενεργειών (π.χ. μηχανοκίνητα χειριστήρια)
- *T(s,a,s)* είναι μια μεταβατική συνάρτηση που καθορίζει την πιθανότητα μετάβασης στην κατάσταση *s, δεδομένης μιας κατάστασης s και της δράσης a, δηλαδή, p(s'| s,a)*
  - καθορίζει τη **δυναμική** του προβλήματος
- *R(s,a)* είναι η συνάρτηση ανταμοιβής (μερικές φορές r(s), r(s) ή r(s,a,s))



Διαφάνειες βασισμένες στο: *UC Berkeley CS188 – Intro to AI course*

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Ιδιότητα Markov

- «Markov» σημαίνει ότι με δεδομένο το παρόν κράτος, το μέλλον είναι ανεξάρτητο από το παρελθόν.
- Για τις ΔΑΜ, «Markov» σημαίνει ότι τα αποτελέσματα της δράσης εξαρτώνται μόνο από την τρέχουσα κατάσταση.

$$P(S_{t+1} = s' | S_t = s_t, A_t = a_t, S_{t-1} = s_{t-1}, A_{t-1}, \dots S_0 = s_0)$$

$$=$$

$$P(S_{t+1} = s' | S_t = s_t, A_t = a_t)$$

Andrey Markov
(1856-1922)

Σε μία ΔΑΜ **όλες οι καταστάσεις** υποτίθεται ότι έχουν την ιδιοκτησία του Markov

- Η κατάσταση συλλαμβάνει όλες τις σχετικές πληροφορίες από το ιστορικό.
- Μόλις γίνει γνωστή η κατάσταση, το ιστορικό μπορεί να πεταχτεί.
- Η κατάσταση είναι ένα επαρκές στατιστικό στοιχείο του παρελθόντος.

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Για παράδειγμα: ρομπότ καθαρισμού

- Ένα ρομπότ που καθαρίζει τα κουτιά σόδας
- Δύο καταστάσεις: **υψηλή** φόρτιση μπαταρίας ή **χαμηλή** φόρτιση μπαταρίας
- Δραστηριότητες: {*περιμένετε, ψάξτε*} ψηλά, {*περιμένετε, αναζήτηση, επαναφόρτιση*} στο χαμηλό
- Η δυναμική μπορεί να είναι στοχαστική
  - P(*St+ 1* = υψηλό | *St* = υψηλό, *At* = αναζήτηση) = α
  - P(*St+ 1* = χαμηλό | *St* = υψηλό, *At* = αναζήτηση) = 1 − α
- Η ανταμοιβή θα μπορούσε να είναι αναμενόμενος αριθμός συλλεγόμενων κουτιών (deterministic) ή πραγματικός αριθμός συλλεγόμενων κουτιών (stochastic)

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα: ρομπότ καθαρισμού ΔΑΜ

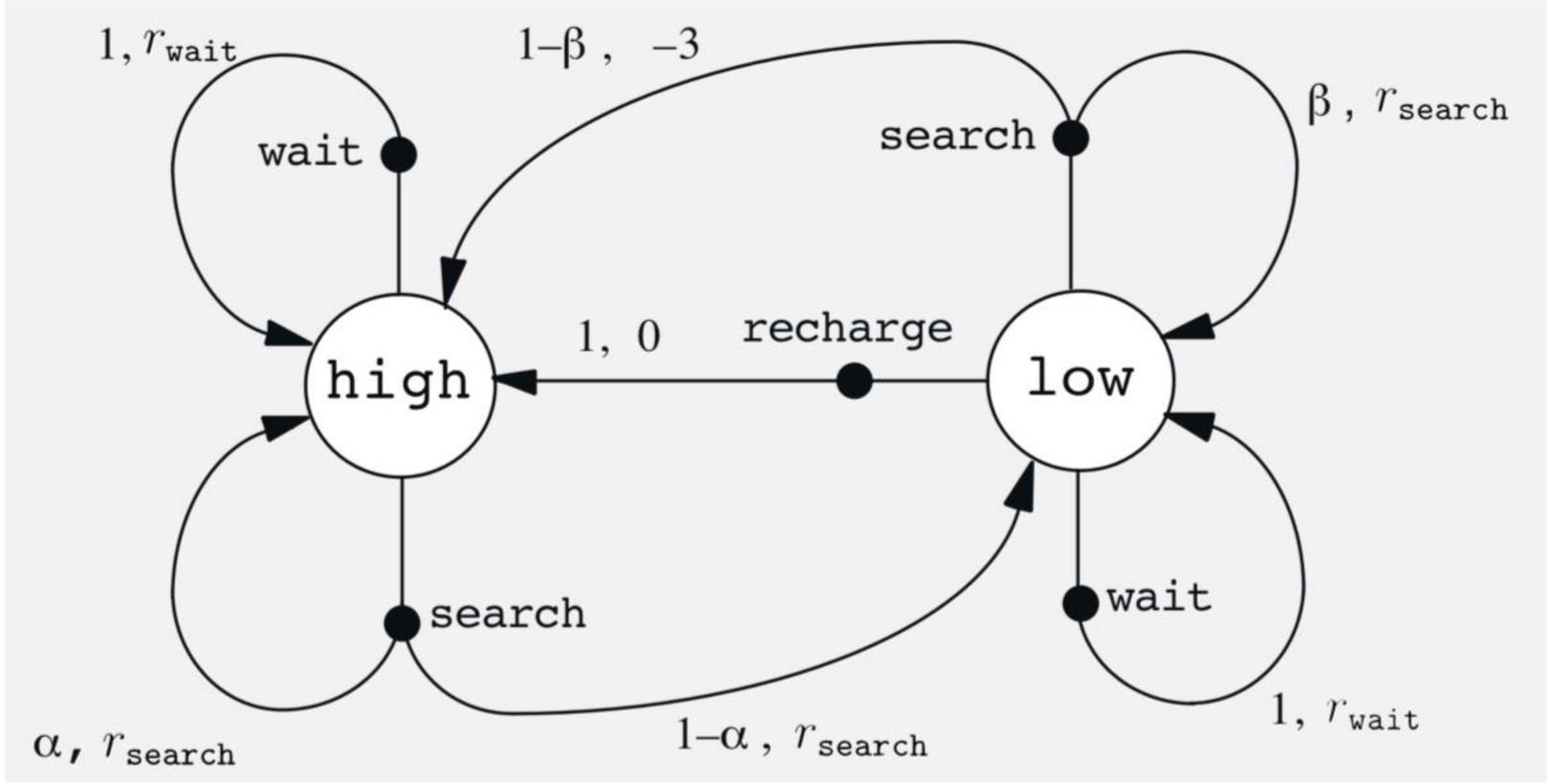
s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
high	search	high	α	$r_{\texttt{search}}$
high	search	low	$1-\alpha$	$r_{\texttt{search}}$
low	search	high	$1-\beta$	-3
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
high	wait	low	0	$r_{\texttt{wait}}$
low	wait	high	0	$r_{\texttt{wait}}$
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0
low	recharge	low	0	0

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα: ρομπότ καθαρισμού ΔΑΜ



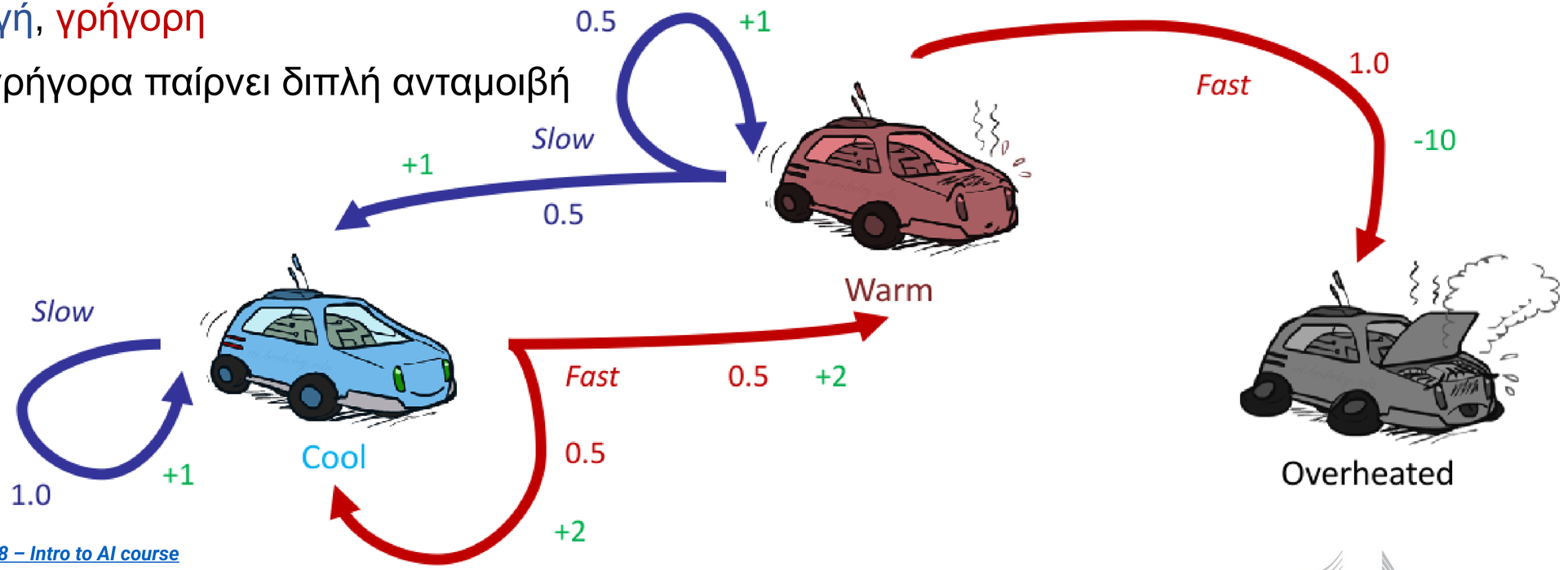
Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα: Αγώνας

- Ένα ρομποτικό αυτοκίνητο θέλει να ταξιδέψει μακριά, γρήγορα
- Τρεις καταστάσεις: Δροσερό, ζεστό, υπερθερμασμένο
- Δύο καταστάσεις: Αργή, γρήγορη
- Το να πηγαίνεις πιο γρήγορα παίρνει διπλή ανταμοιβή

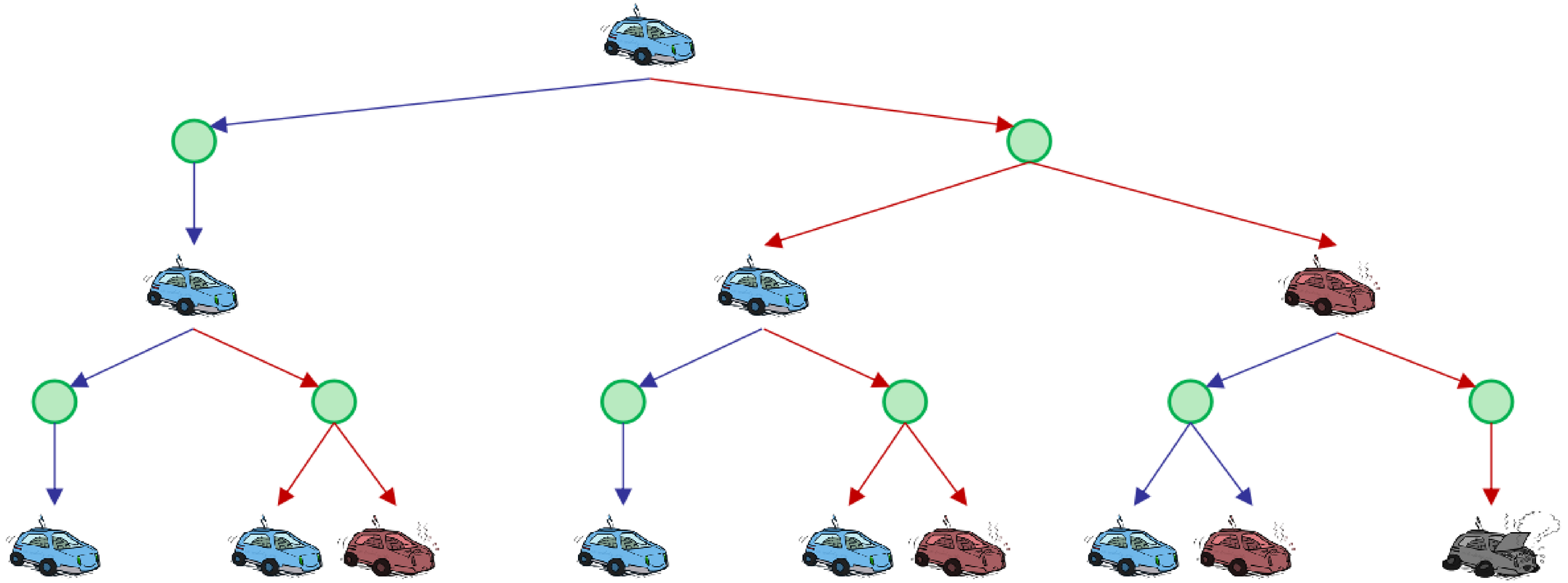


Διαφάνειες βασισμένες στο: *UC Berkeley CS188 – Intro to AI course*

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα: δέντρο αναζήτησης για τον αγώνα



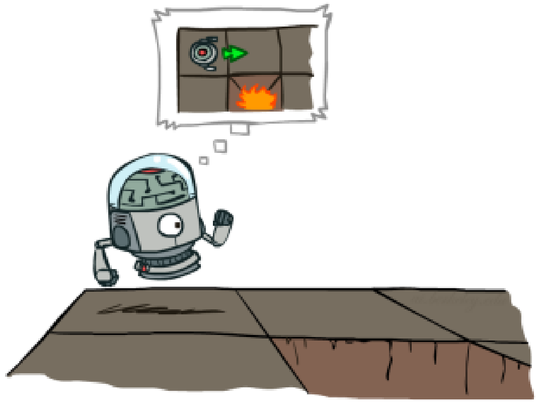
Διαφάνειες βασισμένες στο: *UC Berkeley CS188 – Intro to AI course*

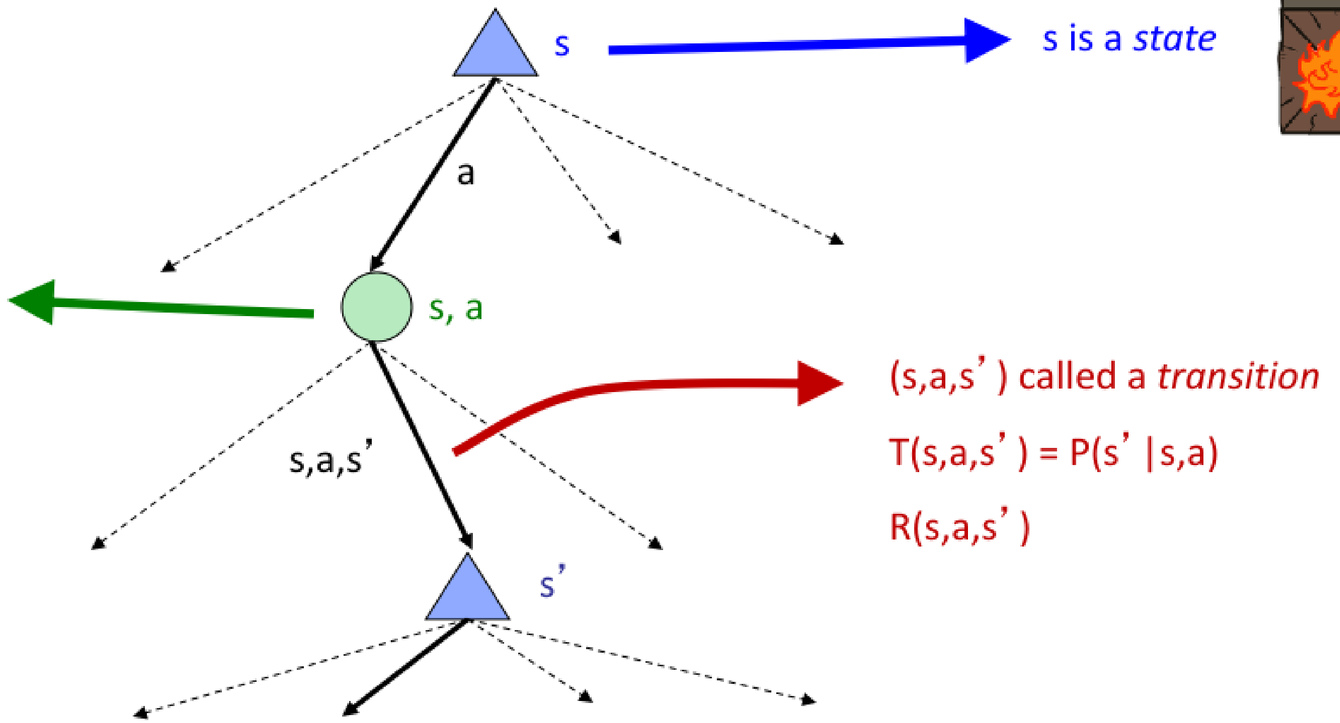
Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

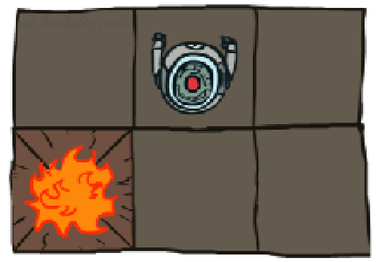
ΔΑΜ Δέντρα Αναζήτησης

- Κάθε κατάσταση ΔΑΜ προβλέπει ένα δέντρο αναζήτησης





s is a *state*



(s,a,s') called a *transition*

$T(s,a,s') = P(s' \mid s,a)$

$R(s,a,s')$

Διαφάνειες βασισμένες στο: ***UC Berkeley CS188 – Intro to AI course***

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Επισημοποίηση του στόχου

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Απόδοση

- Η δράση σε ένα ΔΑΜ έχει ως **αποτέλεσμα άμεσες ανταμοιβές** *Rt*, γεγονός που οδηγεί σε **απόδοση** *Gt*:
  - Απόδοση χωρίς έκπωση (πρόβλημα επισοδικών/πεπερασμένων οριζόντων)

$$G_t = R_{t+1} + R_{t+2} + ... + R_T = \sum_{k=0}^{T-t-1} R_{t+k+1}$$

  - Απόδοση με έκπωση(πρόβλημα πεπερασμένου ή άπειρου ορίζοντα)

$$G_t = R_{t+1} + \gamma R_{t+2} + ... + \gamma^{T-t} R_T = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}$$

  - Μέση απόδοση (συνεχές, άπειρο πρόβλημα ορίζοντα)

$$G_t = \frac{1}{T-t-1}(R_{t+1} + R_{t+2} + ... + R_T) = \frac{1}{T-t-1} \sum_{k=0}^{T-t-1} R_{t+k+1}$$

Σημείωση: Πρόκειται για τυχαίες μεταβλητές που εξαρτώνται από το **ΔΑΜ** και την **πολιτική**

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Απόδοση με έκπτωση

- Είναι λογικό να μεγιστοποιηθεί το άθροισμα των ανταμοιβών
- Είναι επίσης λογικό να προτιμάτε τις ανταμοιβές τώρα από τις ανταμοιβές αργότερα
- Μια λύση: οι αξίες των ανταμοιβών διασπώνται εκθετικά
- Προεξοφλημένες **αποδόσεις** *Gt* για άπειρο ορίζοντα T → p: $G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$
- Η **έκπτωση** γ - [0, 1] είναι η παρούσα αξία των μελλοντικών ανταμοιβών
  - Η τιμή λήψης ανταμοιβής R μετά από k + 1 χρονικά βήματα είναι $\gamma^k R$

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

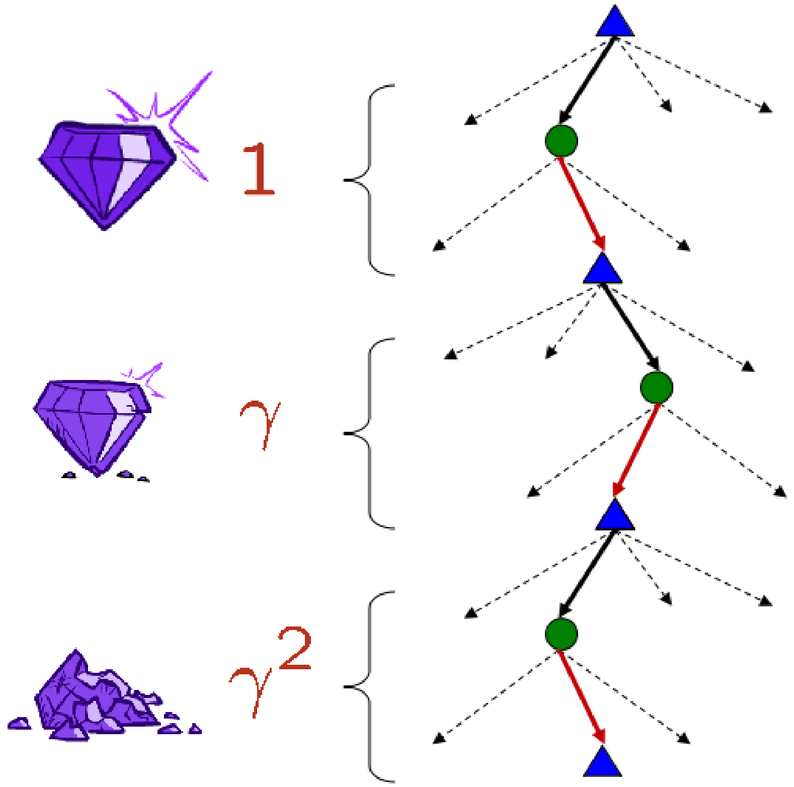
Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Έκπτωση

- Πώς να κάνετε έκπτωση;
  - Κάθε φορά που κατεβαίνουμε ένα επίπεδο, πολλαπλασιάζουμε στην έκπτωση μία φορά

- Γιατί έκπτωση;
  - Νωρίτερα οι ανταμοιβές πιθανόν να έχουν υψηλότερη αξία που αργότερα ανταμείβει
  - γ κοντά στο 0 οδηγεί σε «μυωπική» αξιολόγηση: οι άμεσες ανταμοιβές είναι πιο σημαντικές από τις καθυστερημένες ανταμοιβές
  - γ κοντά στο 1 οδηγεί σε «διορατική» αξιολόγηση
  - Μαθηματικά βολικό: αποφεύγει τις άπειρες αποδόσεις σε κυκλικές ΔΑΜ



Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

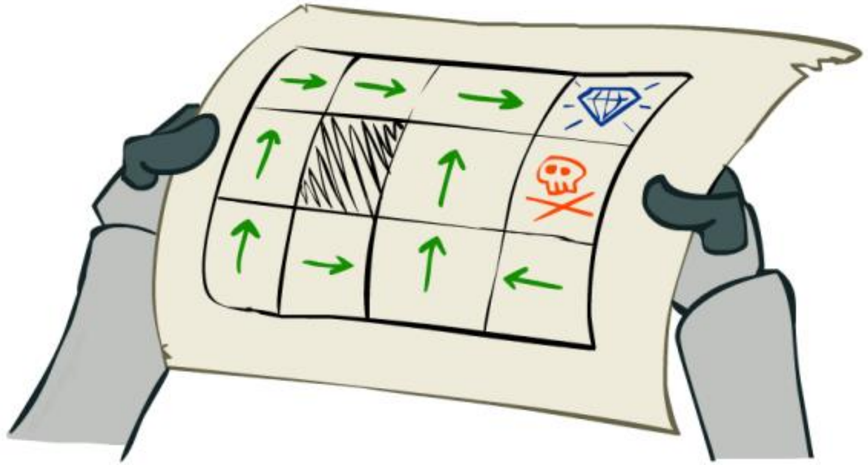
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Πολιτικές

Στόχος ενός πράκτορα της RL:

βρείτε μια πολιτική συμπεριφοράς που μεγιστοποιεί την αναμενόμενη απόδοση G_t

- Μια **πολιτική** είναι μια σχέση $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ η οποία, για κάθε κράτος *εκχωρεί* **για κάθε ενέργεια μια Α την πιθανότητα να αναλάβει αυτή τη δράση σε κατάσταση s**. Υποδεικνύεται από *π(a|s)*.
- Μια βέλτιστη πολιτική είναι αυτή που μεγιστοποιεί την αναμενόμενη απόδοση εάν ακολουθηθεί.
- Για τις ντετερμινιστικές πολιτικές, χρησιμοποιούμε μερικές φορές τον συμβολισμό $_t = \pi(s_t)$ *για* να υποδηλώσουμε τη δράση που έχει αναλάβει η πολιτική.



Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Συναρτήσεις αξίας

- Η **συνάρτηση** *αξίας vs δίνει* τη μακροπρόθεσμη τιμή (αναμενόμενη απόδοση) της κατάστασης *s*

$$v_\pi(s) = \mathbb{E}\left[G_t \mid S_t = s, \pi\right]$$

- Μπορούμε να ορίσουμε τις **αξίες (κατάστασης-)δράσης**:

$$q_\pi(s, a) = \mathbb{E}\left[G_t \mid S_t = s, A_t = a, \pi\right]$$

- Σύνδεση μεταξύ τους:

$$v_\pi(s) = \sum_a \pi(a \mid s) q_\pi(s, a) = \mathbb{E}\left[q_\pi(S_t, A_t) \mid S_t = s, \pi\right] \quad , \ \forall s$$

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Βέλτιστες συναρτήσεις αξίας

- Η βέλτιστη συνάρτηση κατάστασης-τιμής v(s) είναι η συνάρτηση μέγιστης τιμής σε όλες τις πολιτικές

$$v^*(s) = \max_\pi v_\pi(s)$$

- Η βέλτιστη συνάρτηση τιμής δράσης q(s, a) είναι η συνάρτηση μέγιστης τιμής δράσης σε όλες τις πολιτικές

$$q^*(s,a) = \max_\pi q_\pi(s,a)$$

- Η συνάρτηση βέλτιστης τιμής καθορίζει την καλύτερη δυνατή απόδοση στη ΔΑΜ
- Μία ΔΑΜ «επιλύεται» όταν γνωρίζουμε τη βέλτιστη συνάρτηση τιμής

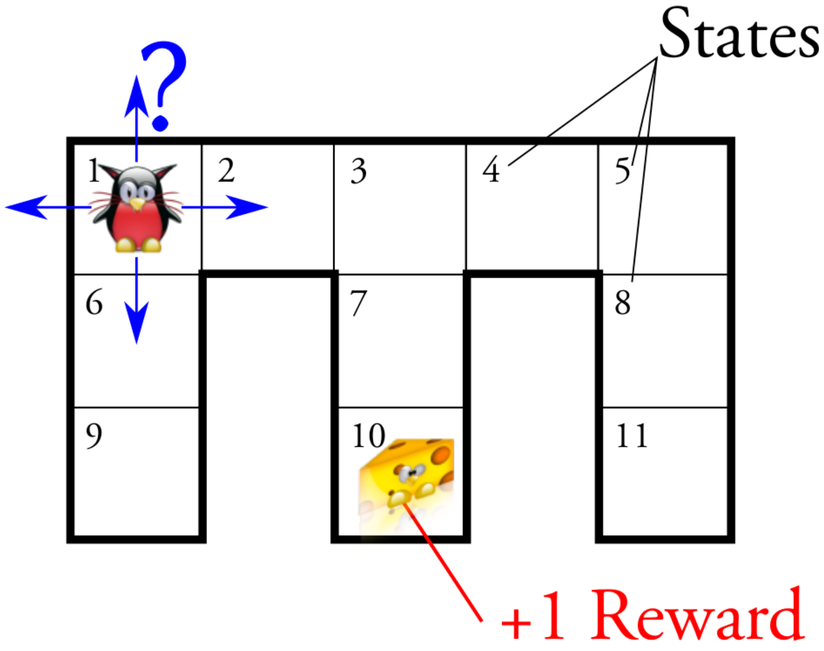
Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα συναρτήσεων αξίας

- **Στόχος**: αναζήτηση πολιτικής π: $S \rightarrow A$ για να μεγιστοποιήσει $\mathbb{E}_{\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$

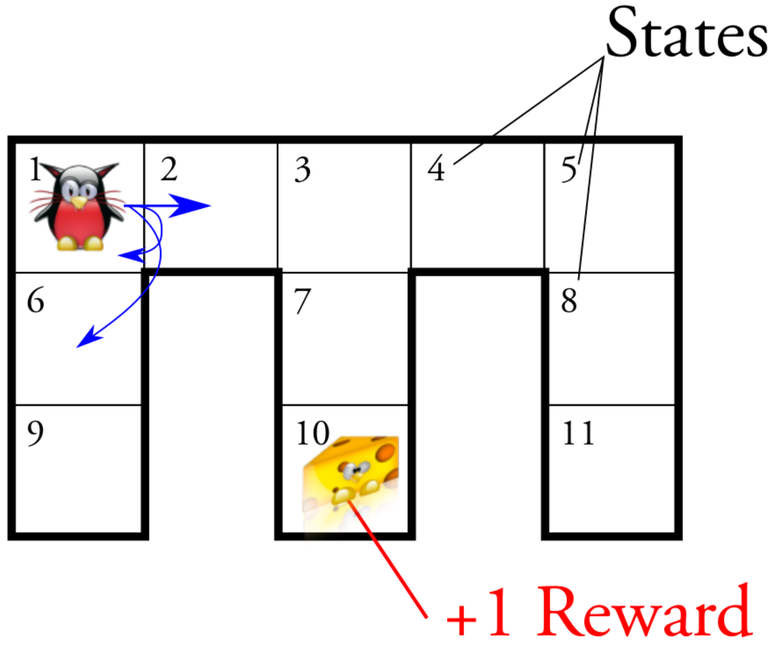


Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα συναρτήσεων αξίας

- **Στόχος**: αναζήτηση πολιτικής π: $S \rightarrow A$ για να μεγιστοποιήσει $\mathbb{E}_{\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$



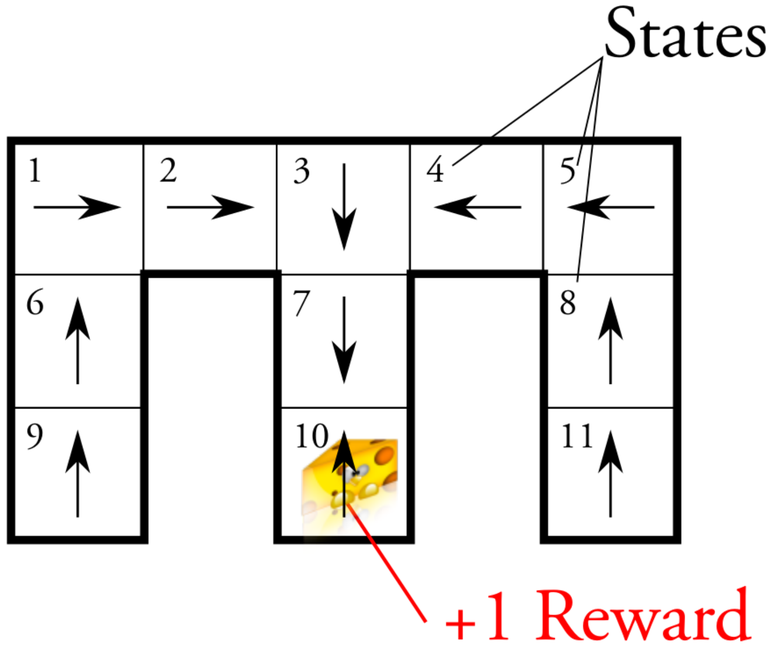
Οι ενέργειες του πράκτορα μπορεί να είναι στοχαστικές

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα συναρτήσεων αξίας

- **Στόχος**: αναζήτηση πολιτικής π: $S \rightarrow A$ για να μεγιστοποιήσει $\mathbb{E}_{\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$



Αυτή είναι η βέλτιστη πολιτική

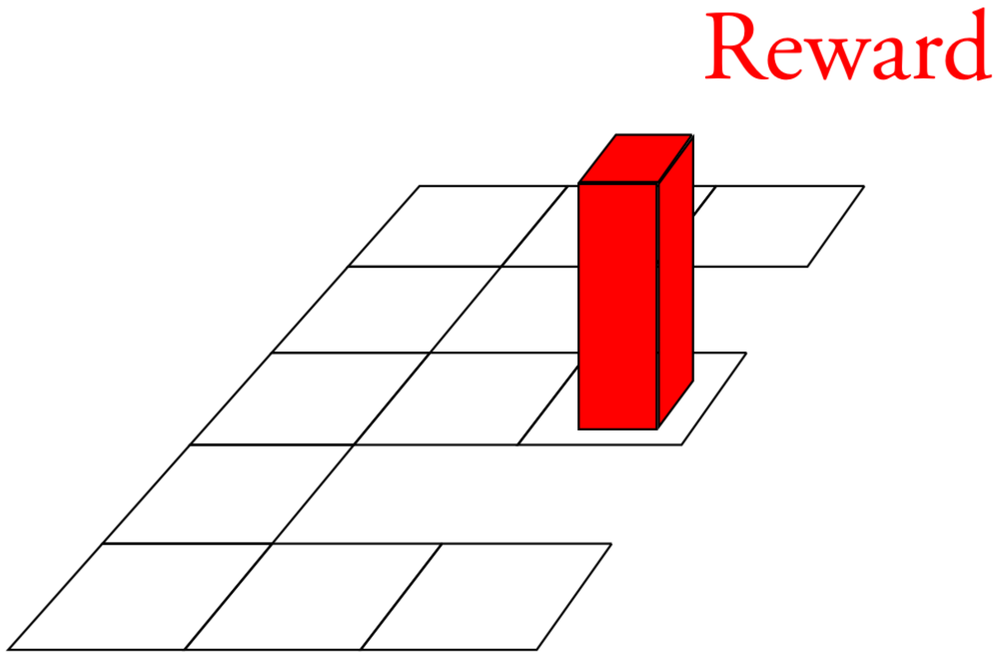
...αλλά πώς θα τη βρούμε;

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα συναρτήσεων αξίας

Η βέλτιστη συνάρτηση τιμής μεταμορφώνει αυτό το «αραιό τοπίο»

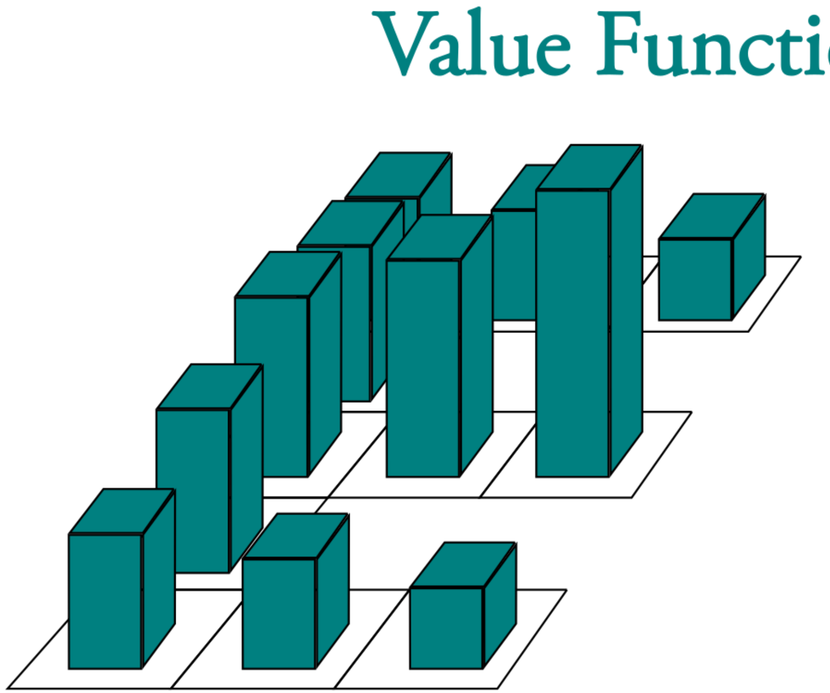


Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα συναρτήσεων αξίας

... σε αυτό το «πυκνό» ένα, το οποίο μπορεί να χρησιμοποιηθεί για να υπολογίσει τη βέλτιστη πολιτική



Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Εύρεση βέλτιστης πολιτικής

- Μια βέλτιστη πολιτική μπορεί να βρεθεί με τη μεγιστοποίηση πάνω από q(s, a)

$$\pi^*(s,a) = \begin{cases} 1 & \text{if } a = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, q^*(s,a) \\ 0 & otherwise \end{cases}$$

- ΠΑΡΑΤΗΡΗΣΕΙΣ:
  - Υπάρχει πάντα μια ντετερμινιστική βέλτιστη πολιτική για κάθε MDP
  - Εάν γνωρίζουμε q(s, a), έχουμε αμέσως τη βέλτιστη πολιτική
  - Μπορεί να υπάρχουν πολλαπλές βέλτιστες πολιτικές
  - Εάν οι πολλαπλές ενέργειες μεγιστοποιήσουν το q (s, ·), μπορούμε επίσης να επιλέξουμε οποιαδήποτε από αυτές (συμπεριλαμβανομένης της στοχαστικής)

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Εξισώσεις Bellman

Ρίτσαρντ Μπέλμαν

(1920-1984)

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Εξισώσεις Bellman

Συνάρτηση αξίας

- Η συνάρτηση τιμής v (s) δίνει τη μακροπρόθεσμη τιμή της κατά $v_\pi(s) = \mathbb{E}\left[G_t \mid S_t = s, \pi\right]$
- Μπορεί να οριστεί αναδρομικά:

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, \pi\right] \\
&= \mathbb{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t \sim \pi(S_t)\right] \\
&= \sum_a \pi(a \mid s) \sum_r \sum_{s'} p(r, s' \mid s, a)\left(r + \gamma v_\pi(s')\right)
\end{aligned}
$$

- Το τελευταίο βήμα γράφει ρητά την προσδοκία

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Εξισώσεις Bellman

Αξίες δράσης

- Μπορούμε να ορίσουμε τιμές κρατικής δράσης: $q_\pi(s,a) = \mathbb{E}\left[G_t \mid S_t = s, A_t = a, \pi\right]$
- Αυτό συνεπάγεται

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a\right] \\ &= \mathbb{E}\left[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a\right] \\ &= \sum_r \sum_{s'} p(r, s' \mid s, a) \left(r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s', a') \right) \end{aligned}$$

- Σημειώστε ότι: $v_\pi(s) = \sum_a \pi(a \mid s) q_\pi(s,a) = \mathbb{E}\left[q_\pi(S_t, A_t) \mid S_t = s, \pi\right] \;,\; \forall s$

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Εξισώσεις Bellman

Εξισώσεις προσδοκιών του Μπέλμαν

Δεδομένου ενός MDP, για κάθε πολιτική π, οι συναρτήσεις τιμών υπακούουν στις ακόλουθες εξισώσεις προσδοκίας:

$$
\begin{aligned}
v_\pi(s) &= \sum_a \pi(s,a)\left[r(s,a) + \gamma \sum_{s'} p(s'|a,s) v_\pi(s')\right] \\
q_\pi(s,a) &= r(s,a) + \gamma \sum_{s'} p(s'|a,s) \sum_{a' \in \mathcal{A}} \pi(a'|s') q_\pi(s',a')
\end{aligned}
$$

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Bellman Εξισώσεις

Εξισώσεις βελτιστοποίησης Bellman

Δεδομένου ενός ΔΑΜ, οι συναρτήσεις βέλτιστης τιμής υπακούουν στις ακόλουθες εξισώσεις προσδοκίας:

$$v^*(s) = \max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|a,s) v^*(s') \right]$$

$$q^*(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|a,s) \max_{a' \in \mathcal{A}} q^*(s',a')$$

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Επίλυση προβλημάτων ΕΜ χρησιμοποιώντας τις εξισώσεις Bellman

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Προβλήματα στην ΕΜ

Δύο προβλήματα:

1. Η εκτίμηση v_π ή q_π ονομάζεται αξιολόγηση πολιτικής ή, απλά, πρόβλεψη
   - Λαμβάνοντας υπόψη μια πολιτική, ποια είναι η αναμενόμενη επιστροφή μου κάτω από αυτή τη συμπεριφορά;
   - Λαμβάνοντας υπόψη αυτό το πρωτόκολλο θεραπείας/τη στρατηγική συναλλαγών, ποια είναι η αναμενόμενη επιστροφή μου;

2. Η εκτίμηση v_* ή q_* ονομάζεται μερικές φορές έλεγχος, επειδή αυτά μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση της πολιτικής
   - Ποιος είναι ο βέλτιστος τρόπος συμπεριφοράς; Ποια είναι η βέλτιστη λειτουργία αξίας;
   - Ποια είναι η βέλτιστη θεραπεία; Ποια είναι η βέλτιστη πολιτική ελέγχου για την ελαχιστοποίηση του χρόνου, της κατανάλωσης καυσίμου κ.λπ.;

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Μια λύση

Εξίσωση Bellman σε μορφή πίνακα

- Η εξίσωση Bellman, για δεδομένη πολιτική π, μπορεί να εκφραστεί με τη χρήση πινάκων,

$$\mathbf{v} = \mathbf{r}^{\pi} + \gamma \mathbf{P}^{\pi} \mathbf{v}$$

που

$$v_i = v(s_i)$$
$$r_i^{\pi} = \mathbb{E}\left[R_{t+1} \mid S_t = s_i, A_t \sim \pi(S_t)\right]$$
$$P_{ij}^{\pi} = p(s_j \mid s_i) = \sum_a \pi(a \mid s_i) p(s_j \mid s_i, a)$$

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Μια λύση

Εξίσωση Bellman σε μορφή πίνακα

- Η εξίσωση Bellman, για δεδομένη πολιτική π, μπορεί να εκφραστεί με τη χρήση πινάκων,

$$\mathbf{v} = \mathbf{r}^\pi + \gamma \mathbf{P}^\pi \mathbf{v}$$

- Αυτή είναι μια γραμμική εξίσωση που μπορεί να λυθεί ά

$$\mathbf{v} = \mathbf{r}^\pi + \gamma \mathbf{P}^\pi \mathbf{v}$$
$$\left(\mathbf{I} - \gamma \mathbf{P}^\pi\right) \mathbf{v} = \mathbf{r}^\pi$$
$$\mathbf{v} = \left(\mathbf{I} - \gamma \mathbf{P}^\pi\right)^{-1} \mathbf{r}^\pi$$

- Η υπολογιστική πολυπλοκότητα είναι $O(|S|^3)$ - είναι δυνατή μόνο για μικρά προβλήματα
- Υπάρχουν **επαναληπτικές μέθοδοι** για μεγαλύτερα προβλήματα

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Μια λύση

Επίλυση της εξίσωσης της βελτιστοποίησης του Bellman

- Η εξίσωση βέλτιστης Bellman είναι μη γραμμική (λόγω του μέγιστου χειριστή)
- Δεν μπορεί να χρησιμοποιηθεί η ίδια άμεση λύση πίνακα όπως για την αξιολόγηση πολιτικής

- Πολλές επαναληπτικές μέθοδοι λύσης:
  - Χρήση μοντέλων/ **δυναμικού προγραμματισμού**
    - Επαναληπτική αξία
    - Επαναληπτική πολιτική
  - Χρήση δειγμάτων
    - Monte Carlo
    - Q-learning
    - SARSA

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Δυναμικός προγραμματισμός

Ο δυναμικός προγραμματισμός αναφέρεται σε μια συλλογή αλγορίθμων που μπορούν να χρησιμοποιηθούν για τον υπολογισμό βέλτιστων πολιτικών δεδομένου ενός τέλειου μοντέλου του περιβάλλοντος ως μια διαδικασία λήψης αποφάσεων Μάρκοβ (ΔΑΜ).

Sutton & Barto 2018

- Δυναμικός προγραμματισμός:
  - αποσυνθέτει ένα πρόβλημα σε μικρότερα υποπροβλήματα με αναδρομικό τρόπο
  - αποτελείται από δύο σημαντικά μέρη:

αξιολόγηση πολιτικής και **βελτίωση της πολιτικής**

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Αξιολόγηση πολιτικής

- Ξεκινάμε συζητώντας πώς να εκτιμήσουμε το $v_\pi(s) = \mathbb{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid s, \pi\right]$

- Ιδέα: μετατρέψτε αυτή την εξίσωση σε ενημέρωση

Αλγόριθμος

- Αρχικοποίηση v_0, π.χ., έως μηδέν
- Να επαναληφθεί: $$\forall s: \quad v_{k+1}(s) \leftarrow \mathbb{E}\left[R_{t+1} + \gamma v_k(S_{t+1}) \mid s, \pi\right]$$
- Σταματώντας: όποτε $v_{k+1}(s) = v_k(s)$, για όλα τα s, πρέπει να έχουμε βρει v_π

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Για παράδειγμα: Αξιολόγηση πολιτικής

actions

ΣΤΟΧΟΣ	1	2	3
4	5	6	7
8	9	10	11
12	13	14	ΣΤΟΧΟΣ

$$R_t = -1$$
on all transitions

π = τυχαία πολιτική

γ = 1,0

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα: Αξιολόγηση πολιτικής

$k = 0$

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0

$v_0 = 0$

$v_{k+1}(s_t) = E [R_{t+1} + \gamma v_k(s_{t+1})]$

$= E [-1 + v_k(s_{t+1})]$

$= -1 + E [v_k(s_{t+1})]$

$v_1(s_t) = -1 + E [v_0(s_{t+1})] = -1$

$k = 1$

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0

$v_2(s_t) = -1 + E [v_1(s_{t+1})]$

$= -1 + (-1-1-1+0)/4 = -1.75$

$k = 2$

0.0	-1.7	-2.0	-2.0
-1.7	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.7
-2.0	-2.0	-1.7	0.0

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα: Αξιολόγηση πολιτικής

$k = 3$

0.0	-2.4	-2.9	-3.0
-2.4	-2.9	-3.0	-2.9
-2.9	-3.0	-2.9	-2.4
-3.0	-2.9	-2.4	0.0

$k = 10$

0.0	-6.1	-8.4	-9.0
-6.1	-7.7	-8.4	-8.4
-8.4	-8.4	-7.7	-6.1
-9.0	-8.4	-6.1	0.0

Συνάρτηση τιμής της τυχαίας πολιτικής

$k = \infty$

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0

Η αξιολόγηση της πολιτικής μπορεί να απαιτήσει πολλά βήματα

$V_{k+1}(s_t) = -1 + E\,[v_k(s_{t+1})]$

$V_{k+1}(s_t) = v_k(s_{t+1})$

$V_k = v_\pi$

Όταν η συνάρτηση αξίας δεν αλλάζει, έχουμε ολοκληρώσει την αξιολόγηση πολιτικής

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Βελτίωση της πολιτικής

- Μπορούμε να χρησιμοποιήσουμε την αξιολόγηση για να βελτιώσουμε την πολιτική μας
- Το να είμαστε άπληστοι σε σχέση με τις αξίες της τυχαίας πολιτικής μπορεί να είναι αρκετό (δεν ισχύει γενικά)

Αλγόριθμος

- Επαναλάβετε χρησιμοποιώντας:

$$\begin{aligned}\forall s: \quad \pi_{\text{new}}(s) &= \operatorname*{argmax}_a q_\pi(s,a) \\ &= \operatorname*{argmax}_a \mathbb{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a\right]\end{aligned}$$

- Αξιολογήστε το νέο και επαναλάβετε

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

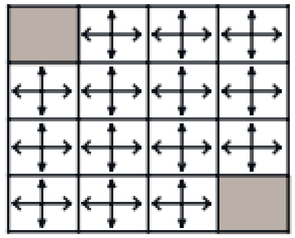
Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα: Αξιολόγιση Πολιτικής + Βελτίωση

$k = 0$

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0

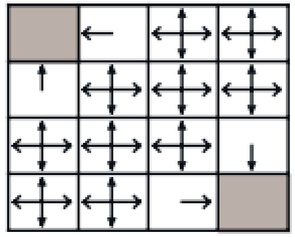


random policy

Μπορούμε να πάρουμε κάποια λειτουργία αξίας και να εξάγουμε κάποια πολιτική χρησιμοποιώντας άπληστη βελτίωση

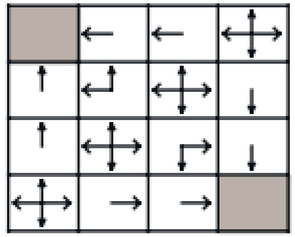
$k = 1$

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0



$k = 2$

0.0	-1.7	-2.0	-2.0
-1.7	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.7
-2.0	-2.0	-1.7	0.0



Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Παράδειγμα: Αξιολόγιση Πολιτικής + Βελτίωση

Μπορούμε να πάρουμε κάποια λειτουργία αξίας και να εξάγουμε κάποια πολιτική χρησιμοποιώντας άπληστη βελτίωση

$k = 3$

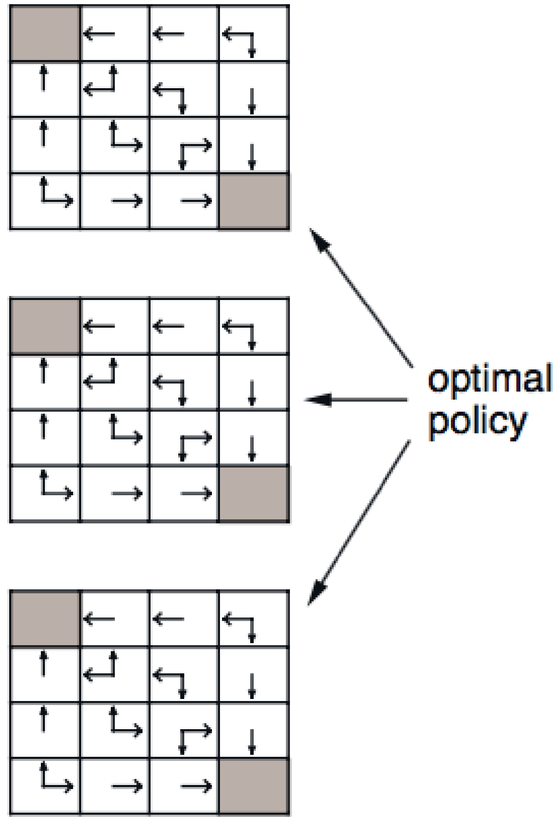
0.0	-2.4	-2.9	-3.0
-2.4	-2.9	-3.0	-2.9
-2.9	-3.0	-2.9	-2.4
-3.0	-2.9	-2.4	0.0

$k = 10$

0.0	-6.1	-8.4	-9.0
-6.1	-7.7	-8.4	-8.4
-8.4	-8.4	-7.7	-6.1
-9.0	-8.4	-6.1	0.0

$k = \infty$

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0



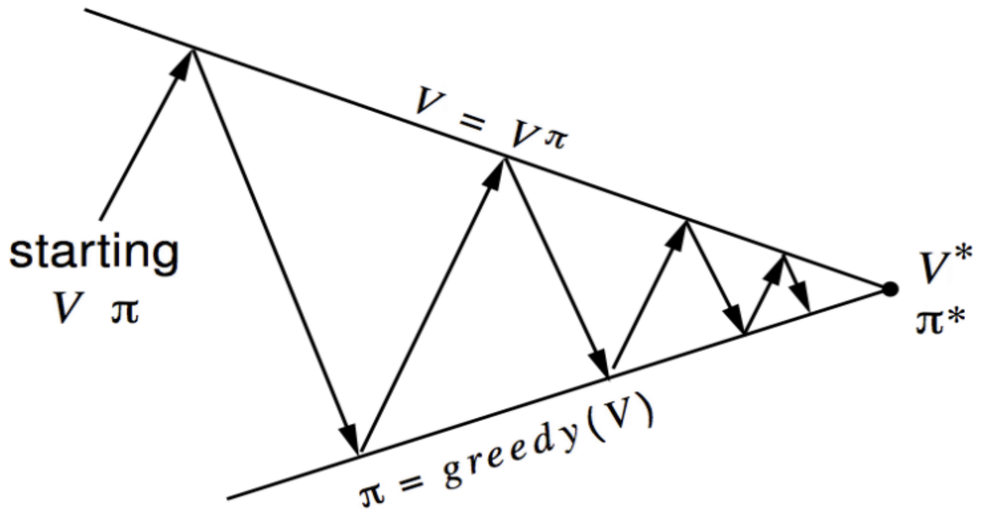
Παρατηρήστε ότι η συνάρτηση αξίας εξακολουθεί να αλλάζει, αλλά η βέλτιστη πολιτική έχει διδαχθεί από το k=3

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

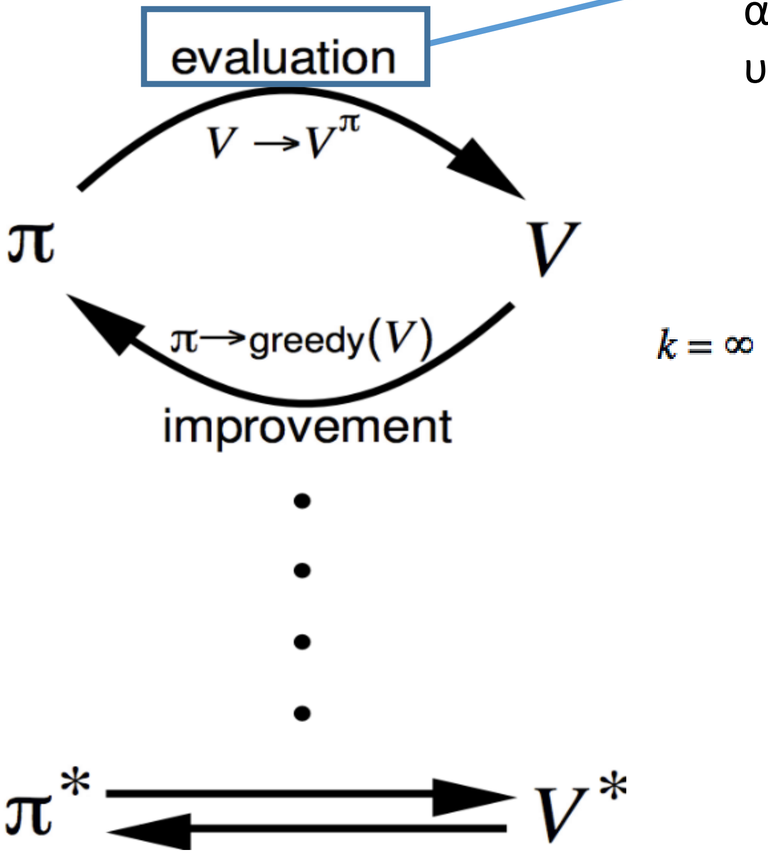
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επαναληπτική μέθοδος Πολιτικής



Policy evaluation Estimate v^{π}

Policy improvement Generate $\pi' \geq \pi$



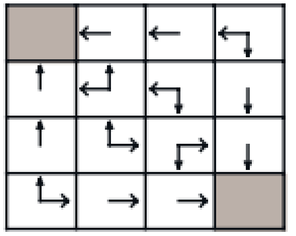
$k = \infty$

Επαναληπτικό, δηλαδή, μπορεί να απαιτήσει πολλούς υπολογισμούς

Αξιολόγηση πολιτικής

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0

Βελτίωση της πολιτικής



Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επαναληπτική μέθοδος Πολιτικής

- Χρειάζεται η αξιολόγηση πολιτικής να συγκλίνει προς το$_{\pi}$;
- Ή μήπως πρέπει να σταματήσουμε όταν είμαστε «κοντά»; (Π.χ., με κατώτατο όριο για τη μεταβολή των τιμών)
  - Ή απλά να σταματήσουμε μετά τις επαναλήψεις της επαναληπτικής αξιολόγησης πολιτικής;
  - Στο μικρό grid world k = 3 ήταν αρκετό για την επίτευξη βέλτιστης πολιτικής

- Ακραίο: Γιατί να μην ενημερώσετε την πολιτική κάθε επανάληψης - δηλαδή σταματήστε μετά το k = 1;
  - Αυτό είναι ισοδύναμο με την **επανάληψη αξίας**

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Επαναληπτική μέθοδος αξίας

- Θα μπορούσαμε να πάρουμε την εξίσωση βελτιστοποίησης Μπέλμαν, και να την μετατρέψουμε σε ενημέρωση.

$$\forall s: \quad v_{k+1}(s) \leftarrow \max_a \mathbb{E}\left[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s, A_t = s\right]$$

- Αυτό ισοδυναμεί με επανάληψη πολιτικής, με k = 1 βήμα αξιολόγησης πολιτικής μεταξύ κάθε δύο (άπληστων) σταδίων βελτίωσης της πολιτικής

Αλγόριθμος: Εξατομίκευση αξίας

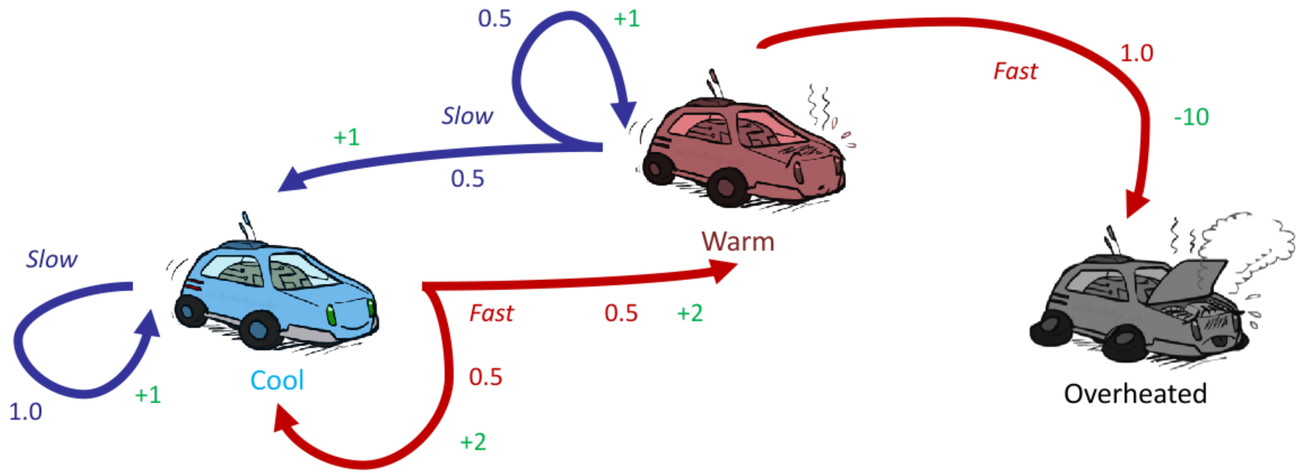
- Αρχικοποίησε v_0
- Επανέλαβε: $v_{k+1}(s) \leftarrow \max_a \mathbb{E}\left[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s, A_t = s\right]$
- Σταμάτησε: όποτε $v_{k+1}(s) = v_k(s)$, για όλα τα s, πρέπει να έχουμε βρει v^*

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επαναληπτική μέθοδος αξίας στο παράδειγμα του αγώνα



Αλγόριθμος

1. $\forall s \in S$, initialize $V_0(s) = 0$

2. Επαναλάβετε μέχρι τη σύγκλιση:

$$\forall s \in S,\ V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_k(s')]$$

Η Σύγκλιση:

$$\forall s, V_{k+1}(s) = V_k(s)$$

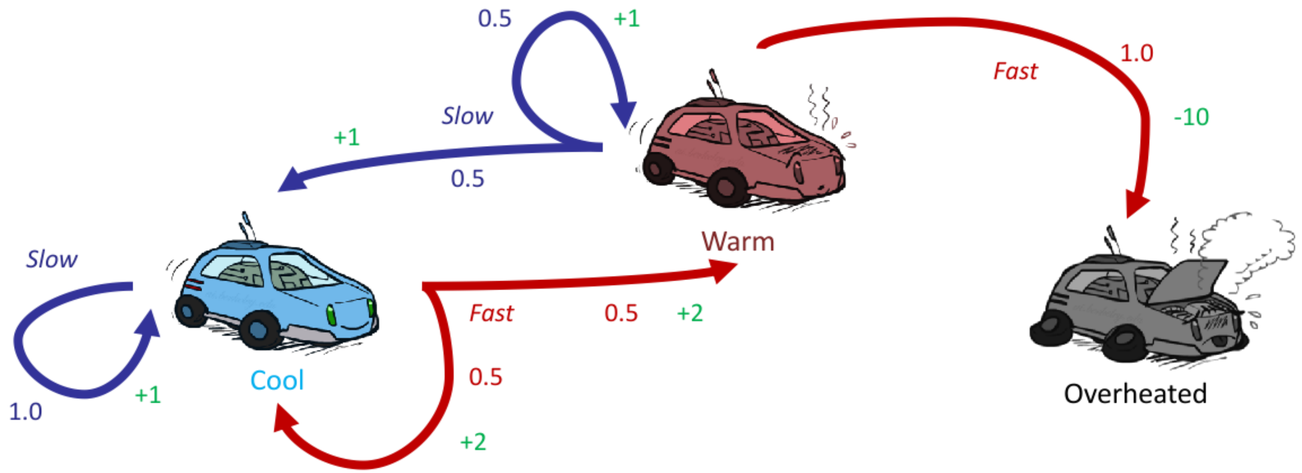
Διαφάνειες βασισμένες στο: ***UC Berkeley CS188 – Intro to AI course***

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επαναληπτική μέθοδος αξίας στο παράδειγμα του αγώνα

$$\forall s \in S,\ V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_k(s')]$$



	cool	warm	overheated
V_0	0	0	0

$$\begin{aligned} V_1(cool) &= \max\{1\cdot[1+0.5\cdot 0],\ 0.5\cdot[2+0.5\cdot 0]+0.5\cdot[2+0.5\cdot 0]\} \\ &= \max\{1,2\} \\ &= \boxed{2} \end{aligned}$$

$$\begin{aligned} V_1(warm) &= \max\{0.5\cdot[1+0.5\cdot 0]+0.5\cdot[1+0.5\cdot 0],\ 1\cdot[-10+0.5\cdot 0]\} \\ &= \max\{1,-10\} \\ &= \boxed{1} \end{aligned}$$

$$\begin{aligned} V_1(overheated) &= \max\{\} \\ &= \boxed{0} \end{aligned}$$

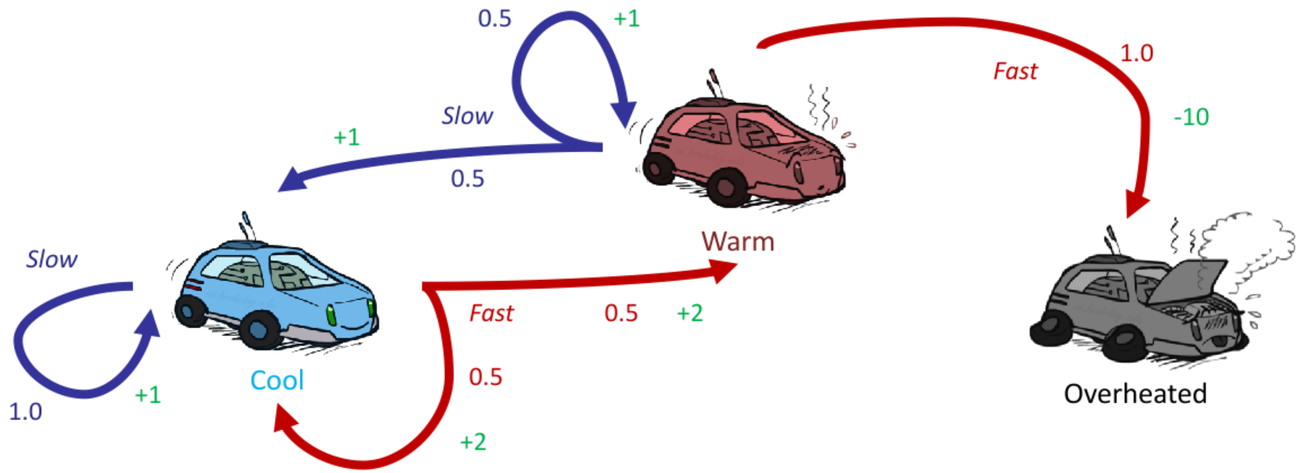
Διαφάνειες βασισμένες στο: *UC Berkeley CS188 – Intro to AI course*

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επαναληπτική μέθοδος αξίας στο παράδειγμα του αγώνα

$$\forall s \in S,\ V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_k(s')]$$



	cool	warm	overheated
V_0	0	0	0
V_1	2	1	0

$$\begin{aligned} V_2(cool) &= \max\{1\cdot[1+0.5\cdot 2],\ 0.5\cdot[2+0.5\cdot 2]+0.5\cdot[2+0.5\cdot 1]\} \\ &= \max\{2, 2.75\} \\ &= \boxed{2.75} \end{aligned}$$

$$\begin{aligned} V_2(warm) &= \max\{0.5\cdot[1+0.5\cdot 2]+0.5\cdot[1+0.5\cdot 1],\ 1\cdot[-10+0.5\cdot 0]\} \\ &= \max\{1.75, -10\} \\ &= \boxed{1.75} \end{aligned}$$

$$\begin{aligned} V_2(overheated) &= \max\{\} \\ &= \boxed{0} \end{aligned}$$

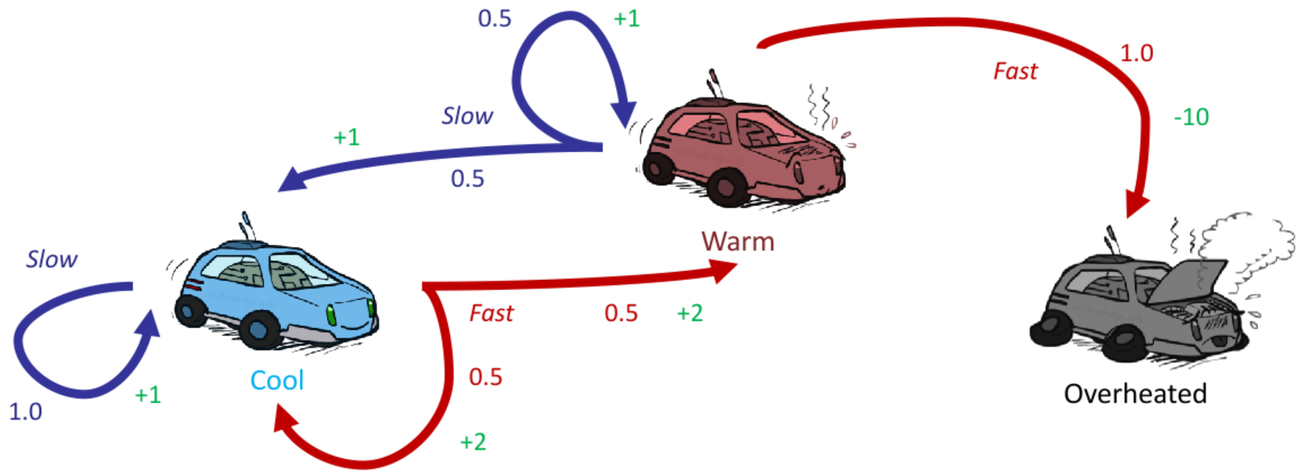
Διαφάνειες βασισμένες στο: *UC Berkeley CS188 – Intro to AI course*

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επαναληπτική μέθοδος αξίας στο παράδειγμα του αγώνα

$$\forall s \in S,\ V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_k(s')]$$



	cool	warm	overheated
V_0	0	0	0
V_1	2	1	0
V_2	2.75	1.75	0

Χρειάζονται περισσότερες επαναλήψεις για να συγκλίνουν στη βέλτιστη συνάρτηση τιμής

Διαφάνειες βασισμένες στο: *UC Berkeley CS188 – Intro to AI course*

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επαναληπτική μέθοδος πολιτικής Vs αξίας

- Επαναληπτική μέθοδος πολιτικής :
  - Ξεκινά με τυχαία πολιτική
  - Διενεργεί αξιολόγηση πολιτικής μέχρι τη σύγκλιση και, στη συνέχεια, ένα βήμα βελτίωσης της πολιτικής
  - Απαιτούνται λίγες επαναλήψεις για να συγκλίνουν, αλλά σε βάρος του περισσότερου χρόνου υπολογισμού λόγω της αξιολόγησης της πολιτικής

- Επαναληπτική μέθοδος αξίας:
  - Ξεκινά με μια τυχαία (ή μηδενική) συνάρτηση τιμής
  - Ακολουθεί ένα βήμα αξιολόγησης πολιτικής, ακολουθούμενο από βελτίωση της πολιτικής
  - Απαιτεί περισσότερες επαναλήψεις για να συγκλίνει στη βέλτιστη συνάρτηση τιμής

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Τι έχουμε καλύψει

- Διαδικασίες λήψης αποφάσεων Markov
- Στόχοι σε ένα MDP: διαφορετική έννοια της επιστροφής
- Συναρτήσεις τιμής — αναμενόμενες αποδόσεις, συνθήκη για την κατάσταση (και δράση)
- Αρχές βελτιστοποίησης σε MDPs: βέλτιστες λειτουργίες αξίας και βέλτιστες πολιτικές
- Bellman Εξισώσεις
- Δύο κατηγορίες προβλημάτων στην RL: αξιολόγηση και έλεγχος
- Πώς να υπολογίσετε v_π (άλλως λύστε ένα πρόβλημα αξιολόγησης/προβλεψιμότητας)
- Πώς να υπολογίσετε τη βέλτιστη συνάρτηση τιμής μέσω δυναμικού προγραμματισμού
  - Επαναληπτική μέθοδος πολιτικής
  - Επαναληπτική μέθοδος αξίας

Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επόμενες Διαλέξεις

- Ενίσχυμένη Μάθηση χωρίς μοντέλο
- Επανάληψη

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Σας ευχαριστούμε

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423