MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Πανεπιστήμιο Κύπρου — Τεχνητή Νοημοσύνη

MAI612 — ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Διάλεξη 18: Ενισχυτική Μάθηση χωρίς μοντέλο

Βασίλης Βασιλειάδης, PhD

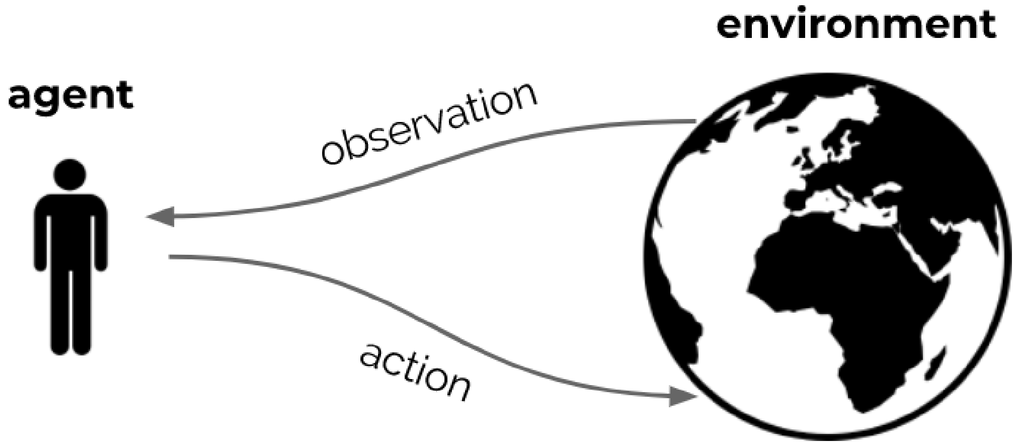
Χειμερινό Εξάμηνο 2022/23

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

Επανάληψη

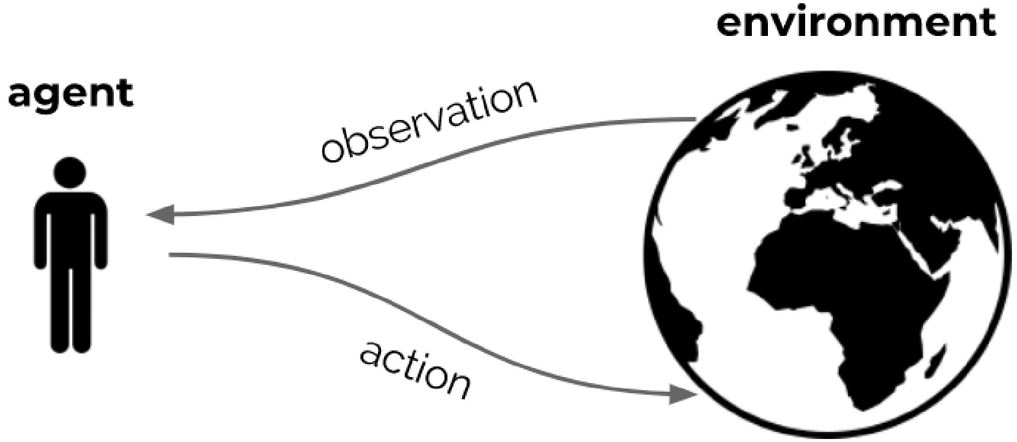


- Η Ενίσχυτική μάθηση είναι η επιστήμη της μάθησης για τη λήψη αποφάσεων
- Οι πράκτορες μπορούν να μάθουν μια **πολιτική**, μια **συνάρτηση αξίας** ή/και ένα **μοντέλο**
- Το γενικό πρόβλημα περιλαμβάνει τη συνεκτίμηση του **χρόνου** και των **συνεπειών**
- Οι αποφάσεις επηρεάζουν την **ανταμοιβή**, την **κατάσταση του 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



- Τελευταία διάλεξη:
  - Αν γνωρίζουμε ένα μοντέλο του MDP μπορούμε να χρησιμοποιήσουμε το σχεδιασμό με δυναμικό προγραμματισμό για να το λύσουμε.
- Αυτή η διάλεξη:
  - Αν δεν γνωρίζουμε το μοντέλο του MDP μπορούμε να χρησιμοποιήσουμε μεθόδους δειγματοληψίας

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

Διάλεξη 18: Ενισχυτική μάθηση χωρίς μοντέλο

Μαθησιακά αποτελέσματα

Θα μάθετε για:

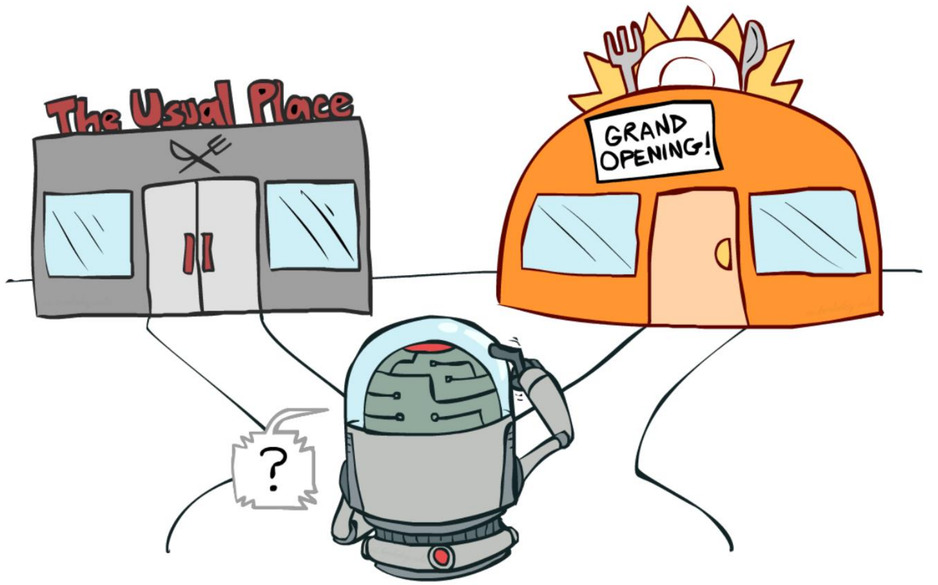
1. Το απλούστερο πλαίσιο των πολλαπλών κουλοχέρηδων και η ανταλλαγή εξερεύνησης-εκμετάλλευσης
2. Πρόβλεψη χωρίς μοντέλο για την εκτίμηση των τιμών σε άγνωστο ΔΑΜ: Monte Carlo, Temporal-Difference (TD) Learning και Multi-step TD learning
3. Έλεγχος χωρίς μοντέλο για τη βελτιστοποίηση των τιμών σε άγνωστο ΔΑΜ: Αλγόριθμοι Sarsa και Q-learning
4. Αισιόδοξη αρχικοποίηση της συνάρτησης αξίας για να βοηθήσει στην εξερεύνηση

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

Εξερεύνηση vs Εκμετάλλευση

- Οι εκπαιδευτικοί πρέπει να ανταλλάξουν δύο πράγματα
  - **Η εκμετάλλευσή** τους: Μεγιστοποίηση της απόδοσης με βάση τις τρέχουσες γνώσεις
  - **Η Εξερεύνηση**: Αύξηση της γνώσης
- Πρέπει να συλλέξουμε πληροφορίες για να πάρουμε τις καλύτερες συνολικές αποφάσεις
- Η καλύτερη μακροπρόθεσμη στρατηγική μπορεί να περιλαμβάνει βραχυπρόθεσμες θυσίες



Πηγή εικόνας: UC Berkeley AI slide, διάλεξη 11.

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

Πολλαπλοί κουλοχέρηδες

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

Ένα απλούστερο πρόβλημα: Πολλαπλοί κουλοχέρηδες

- Το περιβάλλον θεωρείται ότι έχει μόνο **μια κατάσταση**
- ⇒ οι ενέργειες δεν έχουν πλέον μακροπρόθεσμες συνέπειες στο περιβάλλον
- ⇒ οι ενέργειες εξακολουθούν να επηρεάζουν την **άμεση ανταμοιβή**
- ⇒ άλλες παρατηρήσεις μπορούν να αγνοηθούν
- Συζητάμε πώς να μάθετε μια πολιτική σε αυτή τη ρύθμιση

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

Πολλαπλός κουλοχέρης

- Ένας πολλαπλός κουλοχέρης είναι ένα σύνολο διανομών $\{\mathcal{R}_a | a \in \mathcal{A}\}$
- $\mathcal{A}$ είναι ένα (γνωστό) σύνολο ενεργειών (ή «όπλων»)
- $\mathcal{R}_a$ είναι μια διανομή σε ανταμοιβές, δεδομένης της δράσης
- Σε κάθε βήμα *o* πράκτορας επιλέγει μια ενέργεια $A_t \in \mathcal{A}$
- Το περιβάλλον δημιουργεί ανταμοιβή $R_t \sim \mathcal{R}_{A_t}$
- Ο στόχος είναι να μεγιστοποιηθεί η αθροιστική ανταμοιβή $\sum_{i=1}^{t} R_i$
- Το κάνουμε αυτό μαθαίνοντας μια **πολιτική**: μια διανομή στο $\mathcal{A}$

Διαφάνειες βασισμένες στις Διαλέξεις Ενίσχυσης του 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

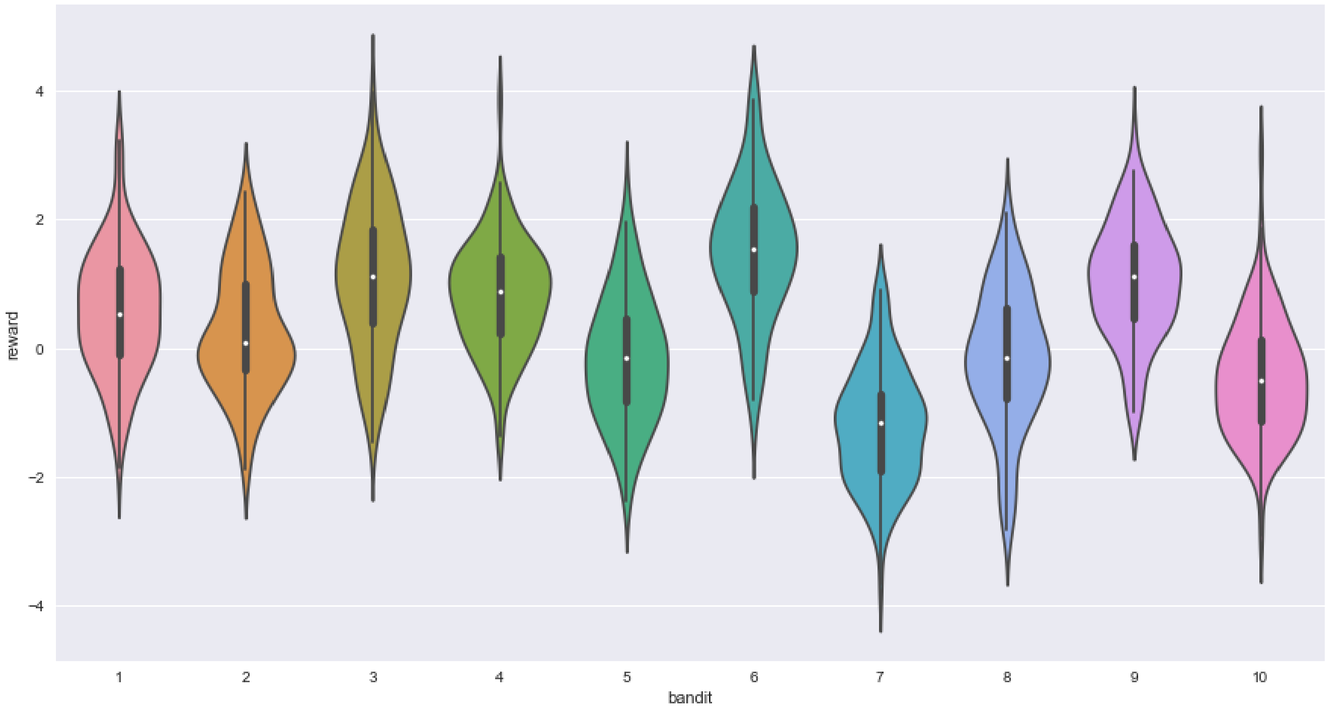
Ο πολλαπλός κουλοχέρης

Πηγή εικόνας

Ένας πολλαπλός κουλοχέρης
- Οι δράσεις είναι ανεξάρτητες

- k Κουλοχέρηδες
  - Διαφορετικές διανομές επιβράβευσης

- Στόχος είναι να βρούμε τον κουλοχέρη που αναμένεται να μας δώσει περισσότερη ανταμοιβή



Ποιο χέρι έχει την υψηλότερη μέση ανταμοιβή;

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

Ο πολλαπλός κουλοχέρης: βιομηχανικές εφαρμογές

- **Κλινικές δοκιμές:** ένα από τα κύρια πρακτικά προβλήματα που χρησιμοποιεί πολυόπλους ληστές.
  - Για την αξιολόγηση των πιθανών θεραπειών K για μια ασθένεια, η ομάδα των ασθενών N χωρίζεται τυχαία σε ομάδες K.
  - Η επιβράβευση: 1 εάν η θεραπεία είναι επιτυχής αλλιώς 0.
  - Μετά από λίγο η πλειοψηφία των ασθενών μπορεί να τεθεί στην καλύτερη θεραπεία.
- **Online ad-placement:** Αποφασίστε ποια διαφήμιση θα εμφανιστεί στην ιστοσελίδα.
- **Βελτιστοποίηση ιστοσελίδας:** διαδοχικά επιλέγοντας σχεδιαστικά στοιχεία (γραμματοσειρά, εικόνες, διάταξη) για την ιστοσελίδα για να μεγιστοποιήσετε το κλικ μέσω του ποσοστού
- **Δρομολόγηση πακέτων δικτύων:** ανταμοιβή είναι ο χρόνος που χρειάζεται για να παραδώσει ένα πακέτο από την πηγή στον προορισμό και υπάρχουν K διαφορετικές διαδρομές διαθέσιμες.

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

Τιμές

- Η **τιμή της δράσης** για τη δράση *α* είναι η αναμενόμενη ανταμοιβή

$$q(a) = \mathbb{E}\left[R_t | A_t = a\right]$$

- Η **βέλτιστη τιμή** είναι

$$v_* = \max_{a \in \mathcal{A}} q(a) = \max_a \mathbb{E}\left[R_t \mid A_t = a\right]$$

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

Αλγόριθμοι

- Υπάρχουν διάφοροι αλγόριθμοι για την εύρεση της πολιτικής:
  - Άπληστος
  - άπληστος
  - UCB
  - Δειγματοληψία Thomson
  - Βαθμίδες πολιτικής
- Θα μιλήσουμε για άπληστους και άπληστους:
  - Χρησιμοποιούν **εκτιμήσεις για την αξία της δράσης** $Q_t(a) \approx q(a)$

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

Εκτιμήσεις της αξίας της δράσης

- Η **τιμή της δράσης** για *τη δράση α* είναι η αναμενόμενη ανταμοιβή $q(a) = \mathbb{E}[R_t | A_t = a]$

- Μια απλή εκτίμηση είναι ο μέσος όρος των επιβράβευσης του δείγματος:

$\mathcal{I}(\cdot)$ είναι η λειτουργία **δείκτη**: $\mathcal{I}(\text{True}) = 1 \text{ and } \mathcal{I}(\text{False}) = 0$

$$Q_t(a) = \frac{\sum_{n=1}^{t} \mathcal{I}(A_n = a)\, R_n}{\sum_{n=1}^{t} \mathcal{I}(A_n = a)}$$

- Η **καταμέτρηση** για τη δράση *α* είναι $N_t(a) = \sum_{n=1}^{t} \mathcal{I}(A_n = a)$

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

Εκτιμήσεις της αξίας της δράσης

- Αυτό μπορεί επίσης να επικαιροποιηθεί σταδιακά:

$$Q_t(A_t) = Q_{t-1}(A_t) + \alpha_t \underbrace{(R_t - Q_{t-1}(A_t))}_{\text{error}},$$

$$\forall a \neq A_t \ : Q_t(a) = Q_{t-1}(a)$$

με $\alpha_t = \frac{1}{N_t(A_t)}$ $\quad N_t(A_t) = N_{t-1}(A_t) + 1 \quad N_0(a) = 0.$

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

Η άπληστη πολιτική

- Μία από τις απλούστερες πολιτικές είναι η **άπληστη**:
  - Επιλέξτε δράση με την υψηλότερη τιμή: $A_t = \underset{a}{\text{argmax}}\ Q_t(a)$
  - Ισοδύναμα: $\pi_t(a) = \mathcal{I}(A_t = \underset{a}{\text{argmax}}\ Q_t(a))$
  - **Καθαρή εκμετάλλευση**: μπορεί να κολλήσει σε μια υποβέλτιστη δράση για πάντα

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

Απληστος αλγόριθμος

- **Ο άπληστος** αλγόριθμος:
  - Με πιθανότητα 1-ε επιλέξτε άπληστη δράση: $a = \underset{a \in \mathcal{A}}{\text{argmax}}\ Q_t(a)$
  - Με πιθανότητα ε επιλέξτε μια τυχαία ενέργεια
  - Ισοδύναμα: $$\pi_t(a) = \begin{cases} (1-\epsilon) + \epsilon/|\mathcal{A}| & \text{if } Q_t(a) = \max_b Q_t(b) \\ \epsilon/|\mathcal{A}| & \text{otherwise} \end{cases}$$
  - ο άπληστος **συνεχίζει να εξερευνά**

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

Κουίζ

Στον **ε-greedy αλγόριθμο**, η ρύθμιση ε=0 έχει ως αποτέλεσμα τον άπληστο αλγόριθμο (καθαρή εκμετάλλευση) και η ρύθμιση ε=1 έχει ως αποτέλεσμα την καθαρή εξερεύνηση. Σωστό ή λάθος;

Σωστό

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

Πρόβλεψη χωρίς μοντέλο

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

Αλγόριθμος Μόντε Κάρλο

- Μπορούμε να χρησιμοποιήσουμε δείγματα εμπειρίας για να μάθουμε χωρίς ένα μοντέλο
- Το Μόντε Κάρλο μαθαίνει: άμεση δειγματοληψία επεισοδίων
- Το MC είναι χωρίς μοντέλο: δεν απαιτείται γνώση της ΔΑΜ, μόνο δείγματα

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

Αξιολόγηση πολιτικής του Monte Carlo

- Ο στόχος μας: μάθετε v_π από επεισόδια εμπειρίας στο πλαίσιο της πολιτικής π

$$S_1, A_1, R_2, \ldots, S_k \sim \pi$$

- Η επιστροφή είναι η συνολική μειωμένη ανταμοιβή (για ένα επεισόδιο που λήγει τη χρονική στιγμή T>t):

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{T-t-1} R_T$$

- Η συνάρτηση τιμής είναι η αναμενόμενη απόδοση:

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

- Μπορούμε ακριβώς να χρησιμοποιήσουμε τη μέση απόδοση δειγμάτων αντί της αναμενόμενης επιστροφής
- Καλούμε αυτή την αξιολόγηση της πολιτικής του Μόντε Κάρλο

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

Μειονεκτήματα του Monte Carlo Learning

- Οι αλγόριθμοι MC μπορούν να χρησιμοποιηθούν για να μάθουν προβλέψεις τιμών
- Αλλά όταν τα επεισόδια είναι μεγάλα, η μάθηση μπορεί να είναι αργή
  - πρέπει να περιμένουμε μέχρι να τελειώσει ένα επεισόδιο για να μάθουμε
  - ... η επιστροφή μπορεί να έχει υψηλή διακύμανση

- Υπάρχουν εναλλακτικές λύσεις;

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

Μάθηση χρονικών διαφορών

- Προηγούμενη διάλεξη: Εξισώσεις Bellman

$$v_\pi(s) = \mathbb{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t \sim \pi(S_t)\right]$$

- Προηγούμενη διάλεξη: Κατά προσέγγιση με επανάληψη

$$v_{k+1}(s) = \mathbb{E}\left[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s, A_t \sim \pi(S_t)\right]$$

- Μπορούμε να το δοκιμάσουμε!

$$v_{t+1}(S_t) = R_{t+1} + \gamma v_t(S_{t+1})$$

- Αυτό είναι πιθανώς αρκετά θορυβώδες — καλύτερα να κάνετε ένα μικρό βήμα (με την παράμετρο α):

$$v_{t+1}(S_t) = v_t(S_t) + \alpha_t\Big(\underbrace{R_{t+1} + \gamma v_t(S_{t+1})}_{\text{target}} - v_t(S_t)\Big)$$

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

Μάθηση χρονικών διαφορών

- Ρύθμιση πρόβλεψης: μάθετε v_π online από την εμπειρία στο πλαίσιο της πολιτικής π
- Μόντε Κάρλο
  - Τιμή επικαιροποίησης $v_n(S_t)$ προς την επιστροφή του δείγματος G_t

$$v_{n+1}(S_t) = v_n(S_t) + \alpha\,(G_t - v_n(S_t))$$

- Μάθηση χρονικής διαφοράς:
  - Τιμή επικαιροποίησης $v_n(S_t)$ προς την εκτιμώμενη απόδοση $R_{t+1} + \gamma\, v(S_{t+1})$

$$v_{t+1}(S_t) \leftarrow v_t(S_t) + \alpha \left(\overbrace{\underbrace{R_{t+1} + \gamma v_t(S_{t+1})}_{\text{target}} - v_t(S_t)}^{\text{TD error}} \right)$$

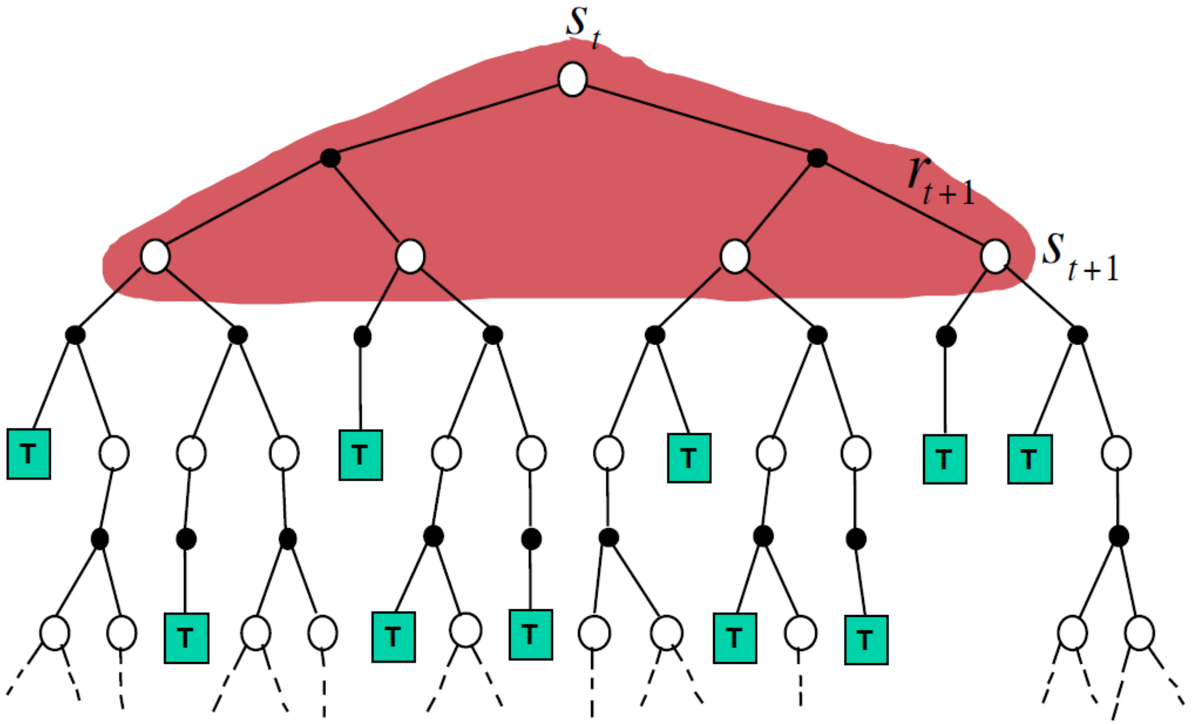
  - $\delta_t = R_{t+1} + \gamma\, v_t(S_{t+1}) - v_t(S_t)$ ονομάζεται σφάλμα TD

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

Διάγραμμα αντιγράφων: Δυναμικός προγραμματισμός

$$v(S_t) \leftarrow \mathbb{E}\left[R_{t+1} + \gamma v(S_{t+1}) \mid A_t \sim \pi(S_t)\right]$$

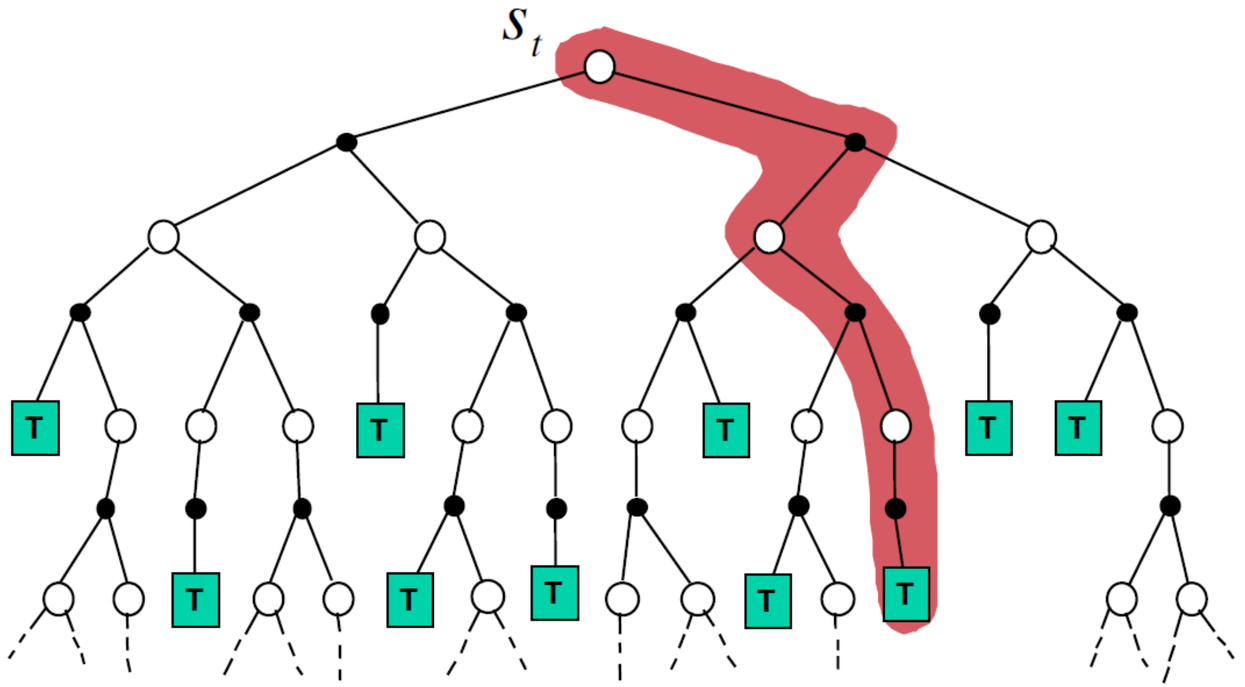


Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

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

Διάγραμμα αντιγράφων: Μόντε Κάρλο

$$v(S_t) \leftarrow v(S_t) + \alpha\,(G_t - v(S_t))$$

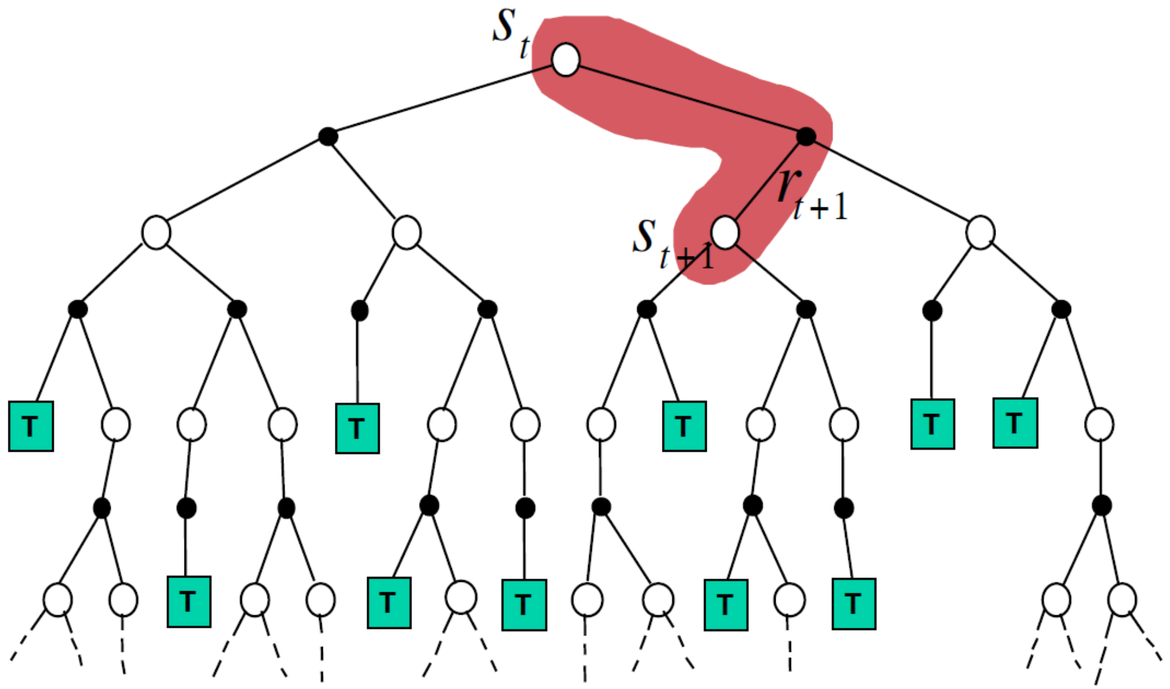


Διαφάνειες βασισμένες στις Διαλέξεις Ενισχυτικής μάθησης της DeepMind

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

Διάγραμμα αντιγράφων: Μάθηση χρονικής διαφοράς

$$v(S_t) \leftarrow v(S_t) + \alpha \left(R_{t+1} + \gamma v(S_{t+1}) - v(S_t)\right)$$



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

Bootstrapping και δειγματοληψία

- **Bootstrapping**: η ενημέρωση περιλαμβάνει μια εκτίμηση (δηλαδή, την αξία της επόμενης κατάστασης)
  - MC δεν κάνει bootstrap
  - DP κάνει bootstraps
  - TD κάνει bootstraps

- **Δειγματοληψία**: ενημέρωση δειγμάτων μιας προσδοκίας
  - MC προβαίνει σε δειγματοληψία
  - Η DP δεν προβαίνει σε δειγματοληψία (δηλαδή, υπολογίζει την επικαιροποίηση καθώς έχει γνώση του μοντέλου)
  - TD προβαίνει σε δειγματοληψία

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

Μάθηση χρονικών διαφορών

- Μπορούμε να εφαρμόσουμε την ίδια ιδέα στις αξίες δράσης
- Μάθηση TD για τις αξίες δράσης:
  - Τιμή επικαιροποίησης $q_t(S_t, A_t)$ προς την εκτιμώμενη απόδοση $R_{t+1} + \gamma\, q(S_{t+1}, A_{t+1})$

$$q_{t+1}(S_t, A_t) \leftarrow q_t(S_t, A_t) + \alpha \left(\overbrace{\underbrace{R_{t+1} + \gamma q_t(S_{t+1}, A_{t+1})}_{\text{target}} - q_t(S_t, A_t)}^{\text{TD error}} \right)$$

  - Αυτός ο αλγόριθμος είναι γνωστός **ως** SARSA, επειδή χρησιμοποιεί $(S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1})$

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

Μάθηση χρονικών διαφορών

- TD είναι μοντέλο-ελεύθερο (καμία γνώση του ΜΑΔ) και μαθαίνει απευθείας από την εμπειρία
- TD μπορεί να μάθει από ελλιπή επεισόδια με bootstrapping
- TD μπορεί να μάθει κατά τη διάρκεια κάθε επεισοδίου

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

Πλεονεκτήματα και μειονεκτήματα του MC vs TD

- Το TD μπορεί να μάθει πριν μάθει το τελικό αποτέλεσμα
  - Το TD μπορεί να μάθει online μετά από κάθε βήμα
  - Το MC πρέπει να περιμένει μέχρι το τέλος του επεισοδίου πριν γίνει γνωστή η επιστροφή
- Το TD μπορεί να μάθει χωρίς το τελικό αποτέλεσμα
  - TD μπορεί να μάθει από ελλιπείς ακολουθίες
  - MC μπορεί να μάθει μόνο από πλήρεις ακολουθίες
  - Το TD λειτουργεί σε συνεχή (μη τερματικά) περιβάλλοντα
  - Το MC λειτουργεί μόνο για επεισοδιακά (τερματικά) περιβάλλοντα
- Το TD είναι ανεξάρτητο από το χρονικό διάστημα της πρόβλεψης
  - Το TD μπορεί να μάθει από μεμονωμένες μεταβάσεις
  - Το MC πρέπει να αποθηκεύσει όλες τις προβλέψεις (ή καταστάσεις) για να ενημερώσει στο τέλος ενός επεισοδίου

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

MAI4CAREU

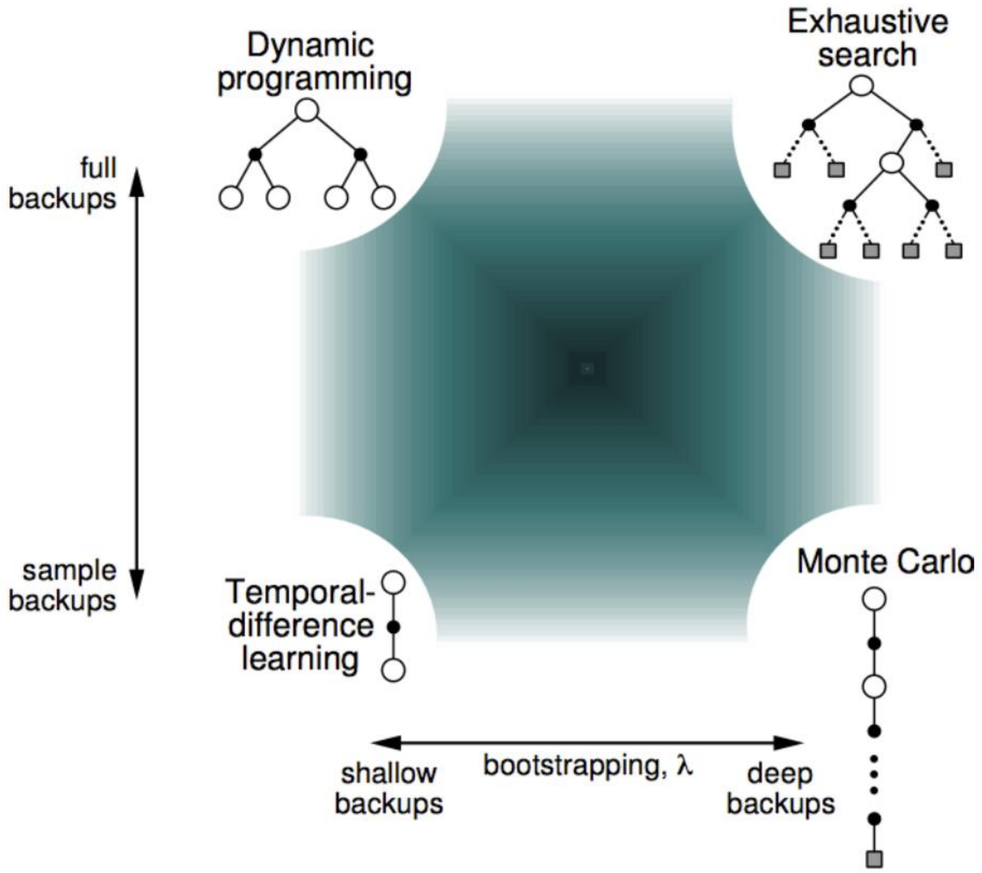
Master programmes in Artificial Intelligence 4 Careers in Europe

Μεταξύ MC και TD: Multi-Step TD

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

Ενοποιημένη άποψη της ΕΜ



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

Ενημερώσεις πολλών βημάτων

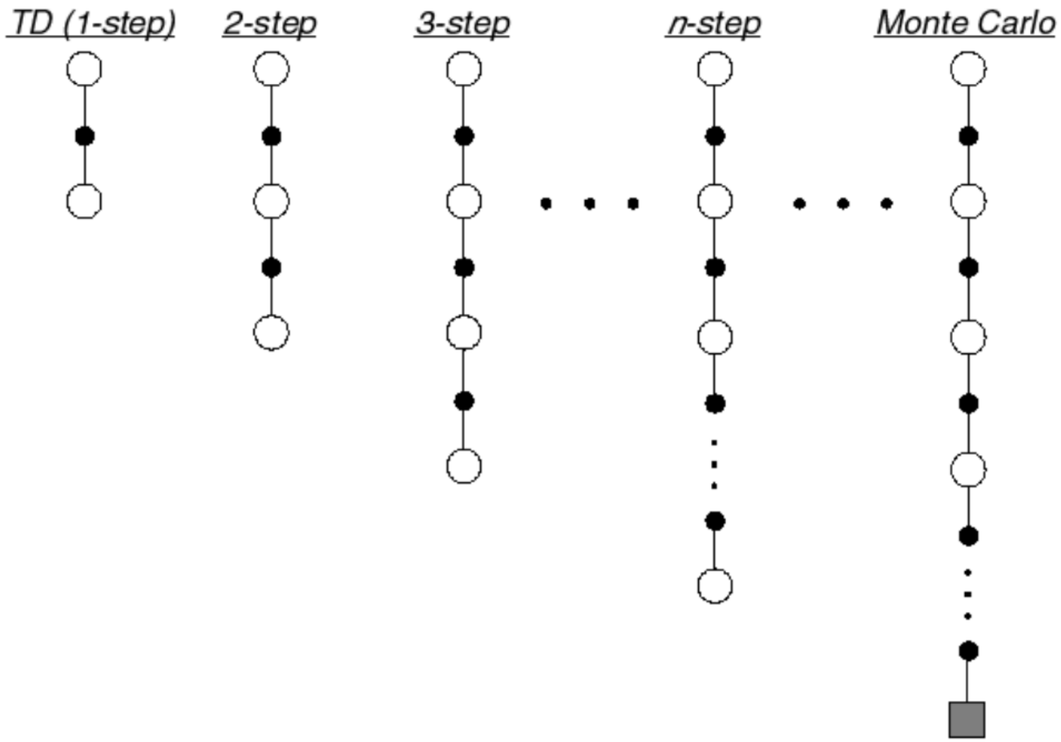
- Η TD χρησιμοποιεί εκτιμήσεις τιμών που μπορεί να είναι ανακριβείς
- Επιπλέον, οι πληροφορίες μπορούν να αναπαραχθούν αρκετά αργά
- Στην MC οι πληροφορίες διαδίδονται ταχύτερα, αλλά οι ενημερώσεις είναι πιο θορυβώδεις
- Μπορούμε να πάμε μεταξύ TD και MC

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

Πρόβλεψη πολλαπλών βημάτων

- Αφήστε τον στόχο TD να κοιτάξει n βήματα στο μέλλον



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

Απόδοση πολλών βημάτων

- Εξετάστε τις παρακάτω n-επιστροφές βημάτων για $n = 1,2,\infty$:

$$\begin{aligned} n = 1 \quad & (TD) \quad G_t^{(1)} = R_{t+1} + \gamma v(S_{t+1}) \\ n = 2 \quad & \quad\;\; G_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 v(S_{t+2}) \\ \vdots \quad & \quad\;\; \vdots \\ n = \infty \quad & (MC) \quad G_t^{(\infty)} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{T-t-1} R_T \end{aligned}$$

- Σε γενικές γραμμές, η επιστροφή στο βήμα n ορίζεται από

$$G_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n v(S_{t+n})$$

- Multi-step TD learning

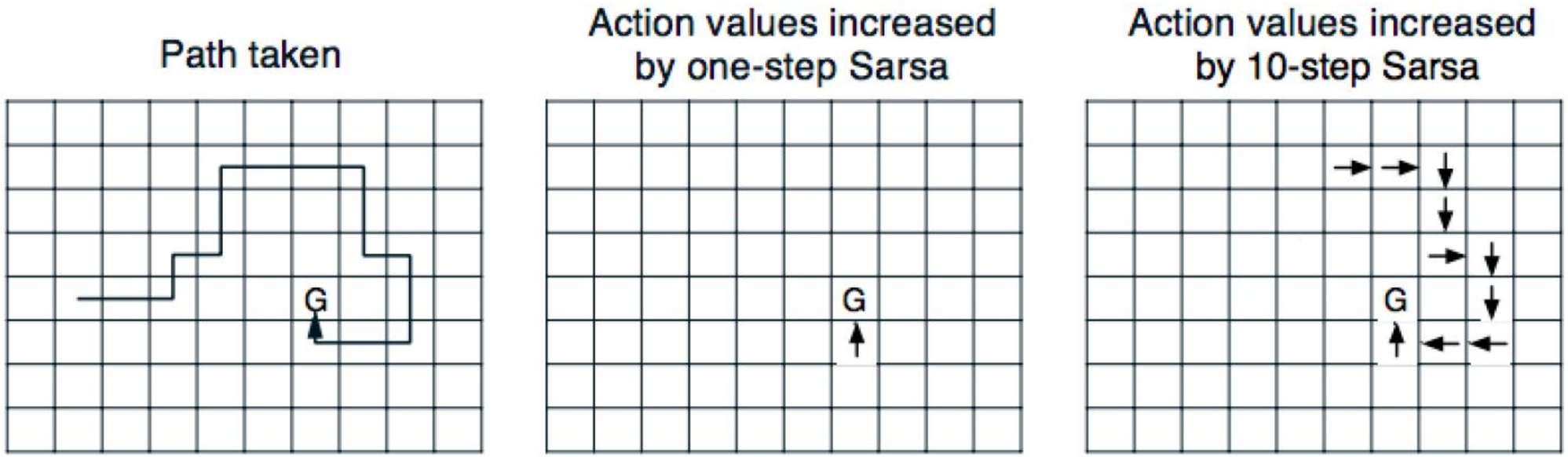
$$v(S_t) \leftarrow v(S_t) + \alpha \left(G_t^{(n)} - v(S_t) \right)$$

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 πλέγματος



(Υπενθύμιση: Το Sarsa είναι TD για τιμές δράσης$q(s,a)$)

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

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Μάθηση χρονικών διαφορών για έλεγχο

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

Μάθηση εντός και εκτός πολιτικής

- Μάθηση εντός πολιτικής
  - Μάθετε για την πολιτική συμπεριφοράς π από την εμπειρία του δείγματος από π
  - Αλγόριθμος Sarsa
- Μάθηση εκτός πολιτικής
  - Μάθετε σχετικά με την πολιτική -στόχο π από την εμπειρία του δείγματος από μ
  - Μάθετε «αντίστροφα» για άλλα πράγματα που θα μπορούσατε να κάνετε: «και αν...;»
  - Π.χ., «Τι γίνεται αν στρίψω αριστερά;» => νέες παρατηρήσεις, ανταμοιβές;
  - Π.χ., «Τι γίνεται αν θα έπαιζα πιο αμυντικά;» => διαφορετική πιθανότητα νίκης;
  - Π.χ., «Τι γίνεται αν συνέχιζα να προχωρώ;» => πόσο καιρό μέχρι να πέσω σε έναν τοίχο;

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

Αλγόριθμος Sarsa

Sarsa (on-policy TD control) for estimating $Q \approx q_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Choose A from S using policy derived from Q (e.g., ε-greedy)
  Loop for each step of episode:
    Take action A, observe R, S'
    Choose A' from S' using policy derived from Q (e.g., ε-greedy)
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma Q(S', A') - Q(S, A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until S is terminal

Πηγή: Sutton & Barto (2018). Reinforcement Learning – An Introduction (2nd edition)

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

Αλγόριθμος Sarsa

- Πίνακας SARSA συγκλίνει στη βέλτιστη συνάρτηση τιμής δράσης, εάν η πολιτική είναι Greedy στο όριο της άπειρης εξερεύνησης (GLIE)
  - Όλα τα ζεύγη κατάστασης-δράσης διερευνώνται απείρως πολλές φορές
  - Η πολιτική συγκλίνει σε μια άπληστη πολιτική
  - Για παράδειγμα, ε-άπληστος με την ε αποσύνθεση με την πάροδο του χρόνου

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

Μάθηση εκτός πολιτικής

- Αξιολόγηση της πολιτικής-στόχου $\pi(a|s)$ για τον υπολογισμό $v_\pi(s)$ ή $q_\pi(s,a)$
- Κατά τη χρήση της πολιτικής συμπεριφοράς $\mu(a|s)$ για τη δημιουργία δράσεων
- Γιατί είναι σημαντικό αυτό;
  - Μάθετε από την παρατήρηση ανθρώπων ή άλλων παραγόντων (π.χ. από καταγεγραμμένα δεδομένα)
  - Επαναχρησιμοποίηση της εμπειρίας από παλιές πολιτικές (π.χ. από τη δική σας εμπειρία στο παρελθόν)
  - Μάθετε για πολλές πολιτικές ενώ ακολουθείτε μία πολιτική
  - Μάθετε για την άπληστη πολιτική ακολουθώντας την διερευνητική πολιτική
- Q-learning εκτιμά την αξία της άπληστης πολιτικής

$$q_{t+1}(s,a) = q_t(S_t, A_t) + \alpha_t \left(R_{t+1} + \gamma \max_{a'} q_t(S_{t+1}, a') - q_t(S_t, A_t) \right)$$

  - Ενεργώντας άπληστος όλη την ώρα δεν θα διερευνήσει επαρκώς

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

Q-learning αλγόριθμος

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    Choose A from S using policy derived from Q (e.g., ε-greedy)
    Take action A, observe R, S'
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
    $S \leftarrow S'$
  until S is terminal

Πηγή: Sutton & Barto (2018). Reinforcement Learning – An Introduction (2nd edition)

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

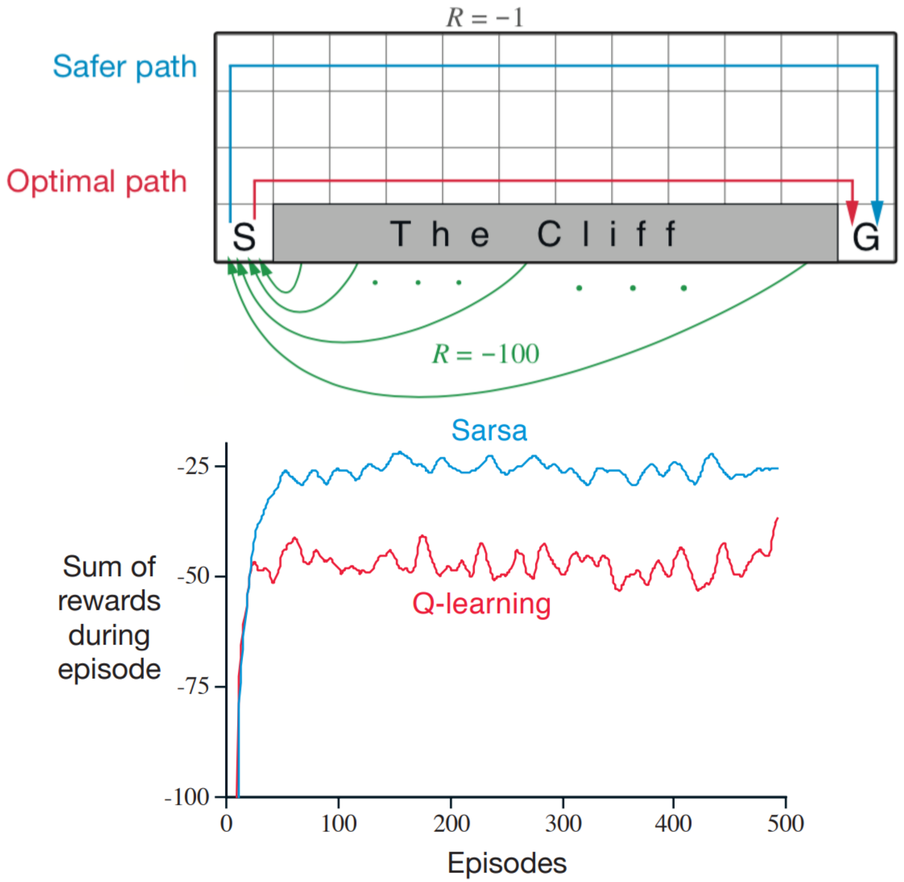
Q-learning αλγόριθμος

- Ο έλεγχος Q-learning συγκλίνει στη βέλτιστη συνάρτηση τιμής δράσης, αρκεί να αναλαμβάνουμε κάθε δράση σε κάθε κατάσταση απείρως συχνά.
- Δεν χρειάζεται άπληστη συμπεριφορά!
- Λειτουργεί για οποιαδήποτε πολιτική που τελικά επιλέγει όλες τις ενέργειες αρκετά συχνά
- Απαιτεί κατάλληλα αποσυντεθειμένα μεγέθη βημάτων

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

Παράδειγμα περπατήματος γκρεμού



Πηγή: Sutton & Barto (2018). Reinforcement Learning – An Introduction (2nd edition)

- Το Q-learning μαθαίνει αξίες για τη βέλτιστη πολιτική, αυτή που ταξιδεύει κατά μήκος της άκρης του γκρεμού.
- Αυτό έχει ως αποτέλεσμα να πέφτει περιστασιακά από το γκρεμό λόγω της άπληστης επιλογής δράσης
- Ο Sarsa λαμβάνει υπόψη την επιλογή δράσης και μαθαίνει την μακρύτερη αλλά ασφαλέστερη διαδρομή μέσω του άνω τμήματος του πλέγματος.
- Αν και η Q-learning μαθαίνει τις αξίες της βέλτιστης πολιτικής, η διαδικτυακή απόδοση της είναι χειρότερη από εκείνη της SARSA, η οποία μαθαίνει την πολιτική του κυκλικού κόμβου.
- Εάν το ε μειωνόταν σταδιακά, τότε και οι δύο μέθοδοι θα συγκλίνουν ασυμπτωματικά στη 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

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Αισιόδοξη αρχικοποίηση

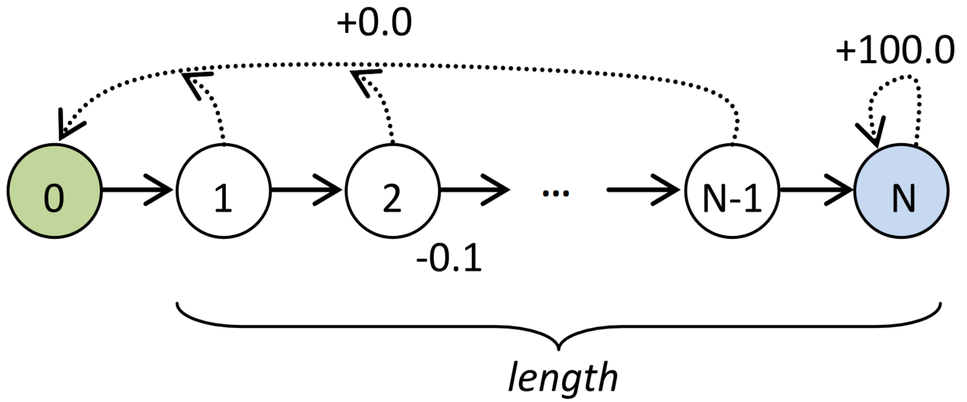
Co-financed by the European Union
Connecting Europe Facility

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

Αισιόδοξη αρχικοποίηση

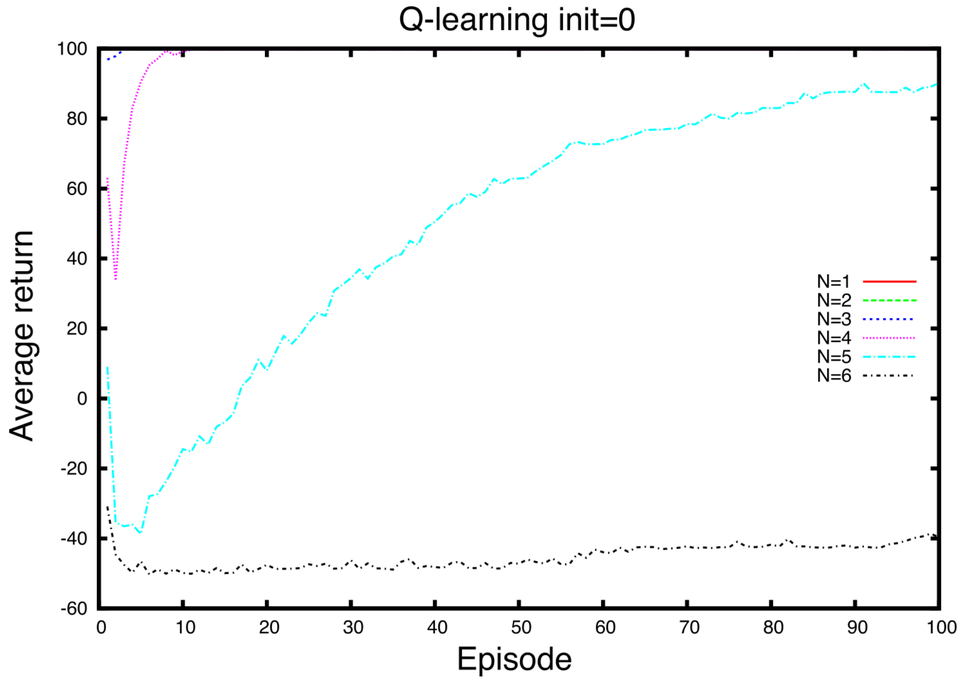
- Απλή προσέγγιση για τη διασφάλιση της εξερεύνησης του χώρου κατάστασης (δράσης)
- Αντί της αρχικοποίησης της συνάρτησης τιμών στο 0, αρχίζετε το r_{max} /(1-γ)
- Ενεργώντας άπληστα (αντί για ε-greedy) επιτρέπει στον αλγόριθμο να δοκιμάσει κάθε ενέργεια σε κάθε κατάσταση τουλάχιστον μία φορά
- Με αυτόν τον τρόπο, ο πράκτορας είναι σε θέση να ανακαλύψει μια καθυστερημένη ανταμοιβή

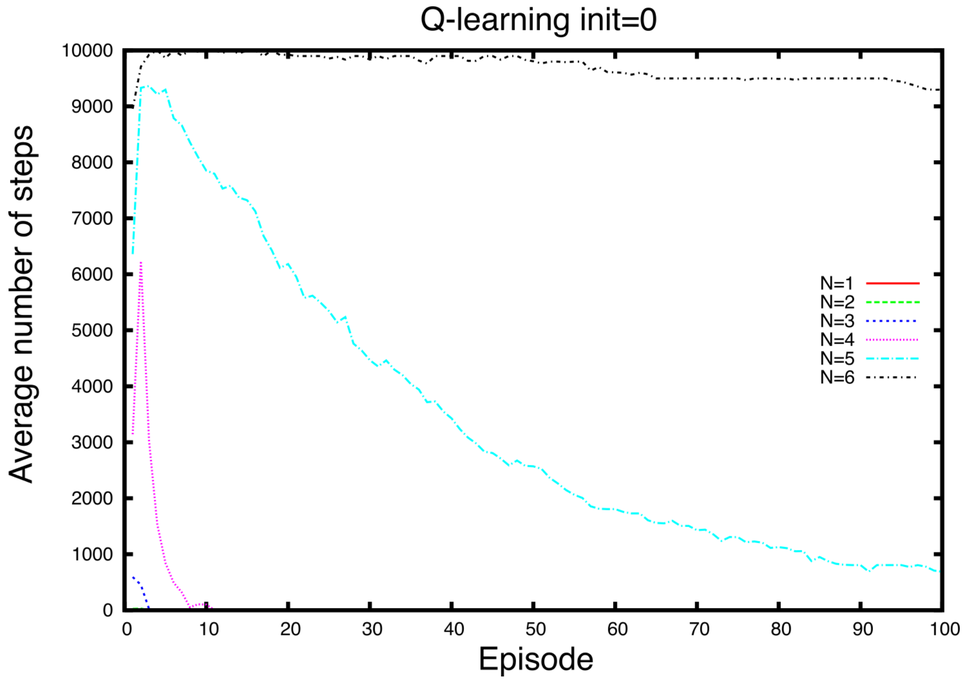
- Για παράδειγμα: Συνδυασμός-κλειδαριά MDP



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

Αποτελέσματα MDP συνδυασμού-κλειδώματος: $Q_{init} = 0$





Παράμετροι

ε=0.1

α=0.95

γ=0.95

επεισόδια = 100

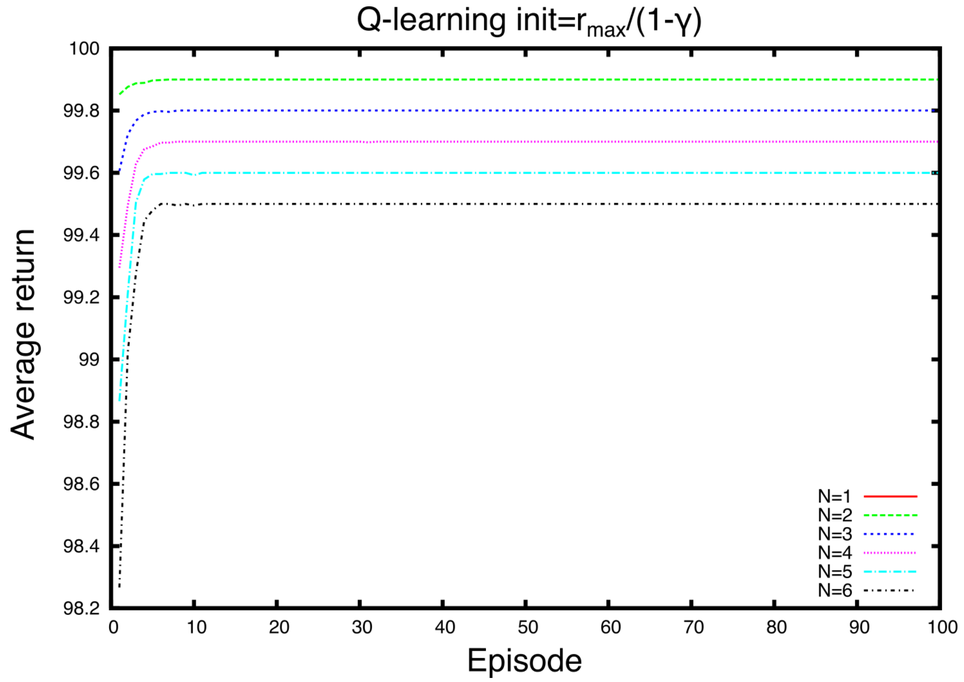
max_steps=10K

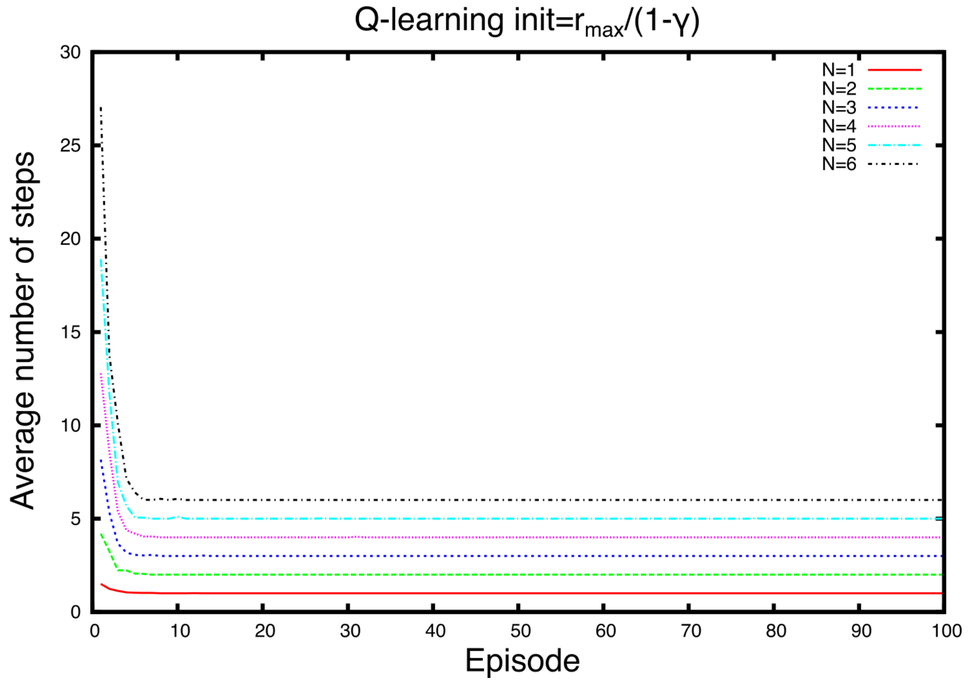
Run=100

Qinit=0

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

Αποτελέσματα MDP συνδυασμού-κλειδώματος: $Q_{init} = r_{max}/1 - \gamma$





Παράμετροι

ε=0

α=0.95

γ=0.95

επεισόδια = 100

max_steps=10K

Run=100

Qinit=2000

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