Μέθοδοι Ελαχίστων Τετραγώνων για Έλεγχο στα Πλαίσια της Ενισχυτικής Μάθησης

Least-Squares Methods in Reinforcement Learning for Control

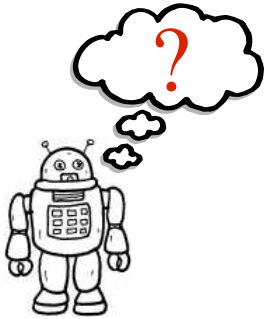
Μιχαήλ Γ. Λαγουδάκης
Ronald E. Parr
Michael L. Littman

Department of Computer Science
Duke University, Durham, NC 27708, U.S.A.

2ο Πανελλήνιο Συνέδριο Τεχνητής Νοημοσύνης
Θεσσαλονίκη, 11-12 Απριλίου 2002

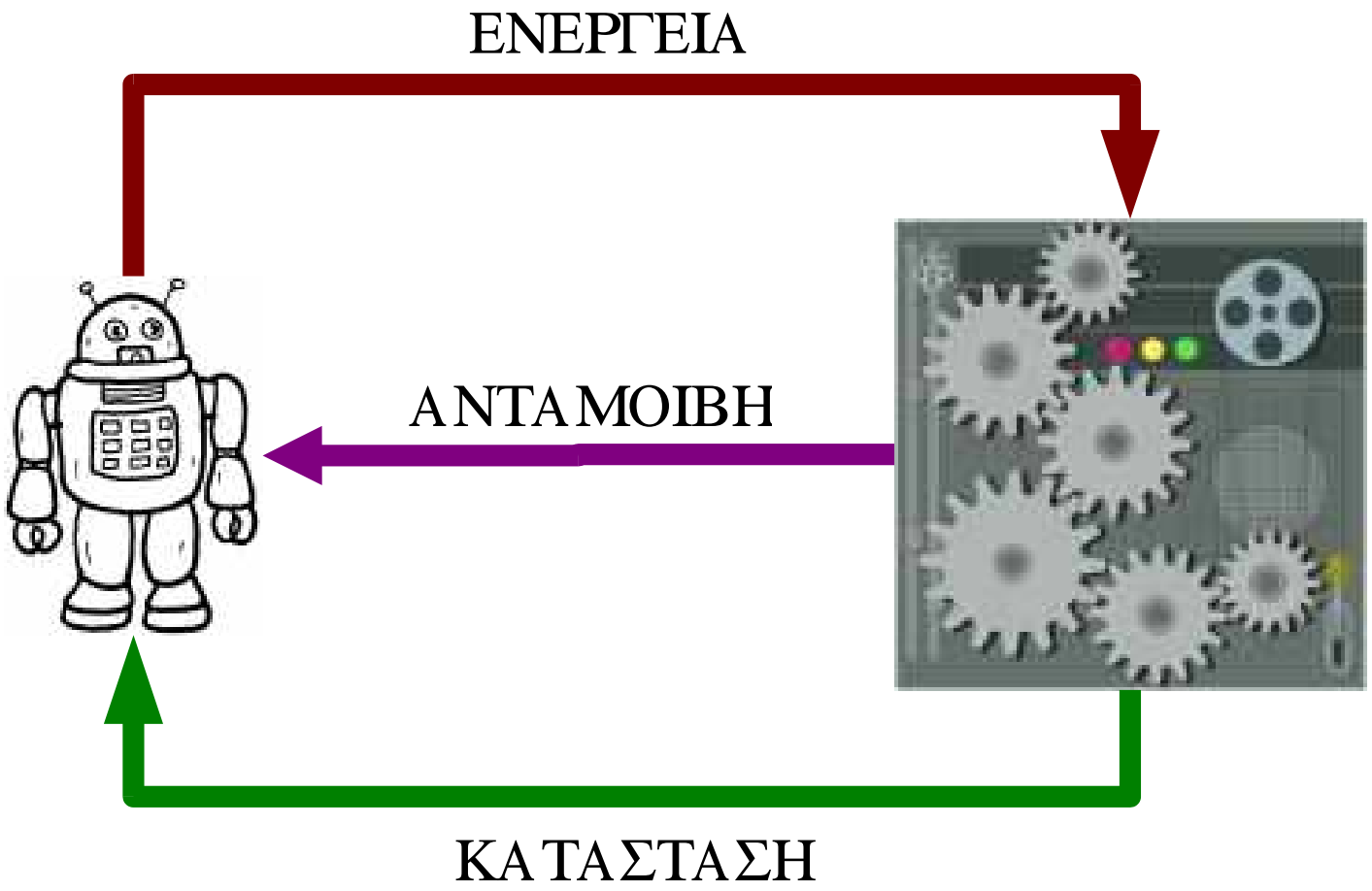
Λήψη Αποφάσεων υπό Αβεβαιότητα

Decision Making under Uncertainty



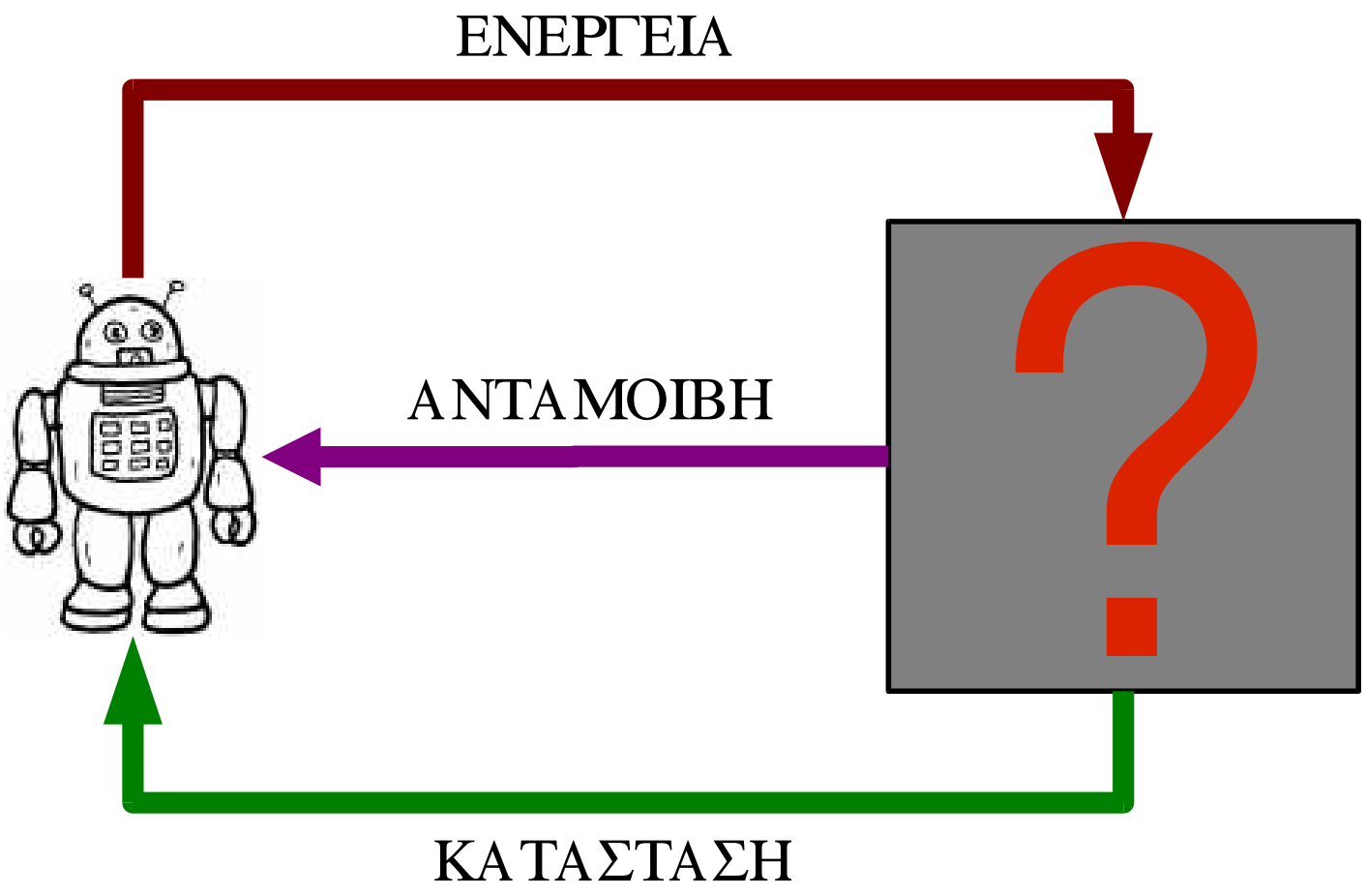
- Πώς μπορώ να βελτιστοποιήσω τη χρησιμοποίηση κάποιου πόρου;
- Πώς μπορώ να μεγιστοποιήσω την παραγωγή ενός εργοστασίου;
- Πώς μπορώ να ισορροπήσω ένα ποδήλατο;
- Πώς μπορώ να παίξω και να κερδίσω σε κάποιο παιχνίδι;
- Πώς μπορώ να κερδίσω στο χρηματιστήριο;

Σχεδιασμός (Planning)



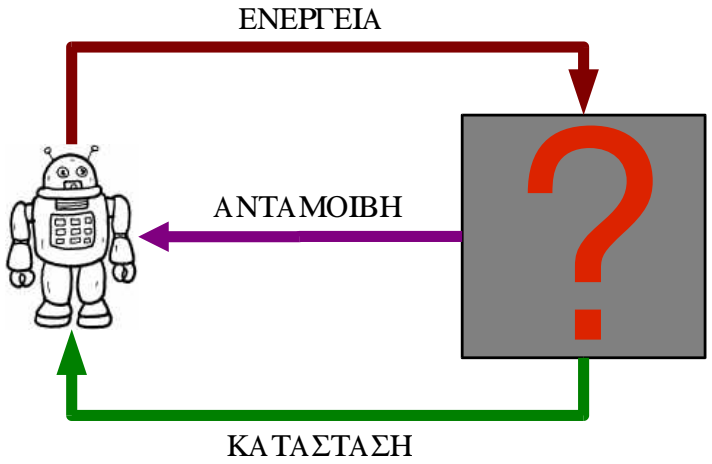
Η διεργασία είναι γνωστή

Μάθηση (Learning)



Η διεργασία είναι άγνωστη

Ενισχυτική Μάθηση (Reinforcement Learning)



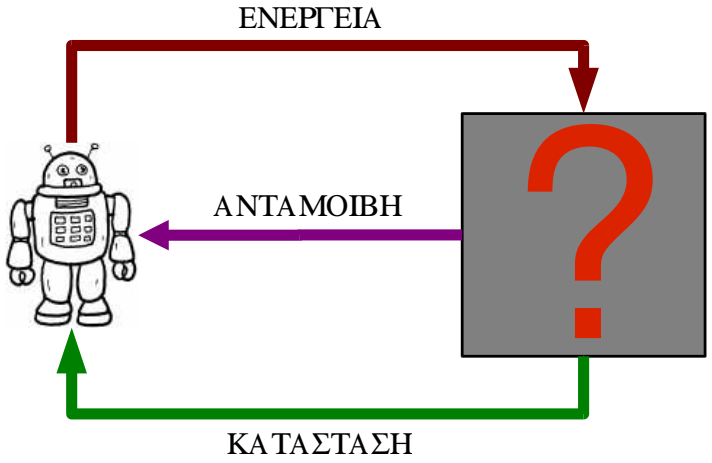
Πρόβλεψη (Prediction)

- Μάθε να προβλέπεις την προσδοκώμενη συνολική ανταμοιβή για μια δεδομένη αμετάβλητη πολιτική επιλογής ενεργειών.

Έλεγχος (Control)

- Μάθε να ελέγχεις τη διεργασία (κατάλληλη επιλογή ενεργειών) ώστε να μεγιστοποιήσεις την προσδοκώμενη συνολική ανταμοιβή.

Αλγόριθμοι Ενισχυτικής Μάθησης



Οι περισσότεροι αλγόριθμοι ενισχυτικής μάθησης ...

- απαιτούν πάρα πολλά δείγματα.
- κάνουν μικρές ρυθμίσεις για κάθε δείγμα και μετά το απορρίπτουν.

Ως αποτέλεσμα ...

- η μάθηση (σύγκλιση) είναι εξαιρετικά αργή!
- τα 'παθήματα του παρελθόντος' περνούν στη λήθη!

Το Κίνητρο

Least-Squares Temporal Difference (LSTD) Learning

[Bradtke and Barto, 1996]

Αποδοτικός αλγόριθμος για προβλήματα πρόβλεψης

- Χρησιμοποιεί τα δείγματα πολύ αποδοτικά
- Αντιμετωπίζει επιτυχώς προβλήματα μεγάλης κλίμακας
- Χρησιμοποιεί γραμμικές αρχιτεκτονικές προσέγγισης
- Χρησιμοποιεί μεθόδους ελαχίστων τετραγώνων

Ωστόσο, ο αλγόριθμος LSTD ...

- είναι κατάλληλος μόνο για προβλήματα πρόβλεψης
- είναι προβληματικός, αν χρησιμοποιηθεί για προβλήματα ελέγχου [Koller and Parr, 2000]

Ο Στόχος μας

Αποδοτικός αλγόριθμος για προβλήματα ελέγχου

Θέλουμε ...

- να μαθαίνει γρήγορα
- να χρησιμοποιεί τα δείγματα πολύ αποδοτικά
- να αντιμετωπίζει επιτυχώς προβλήματα μεγάλης κλίμακας

Χρησιμοποιήσαμε ...

- γραμμικές αρχιτεκτονικές προσέγγισης
- μεθόδους ελαχίστων τετραγώνων

Το αποτέλεσμα ...

- **LSQL**: *Least-Squares Q-Learning* [Lagoudakis and Littman, 2000]
- **LSPI**: *Least-Squares Policy Iteration* [Lagoudakis and Parr, 2001]

Διάγραμμα

- *Τεχνικό Υπόβαθρο*
- *Προσέγγιση*
- *Μάθηση*
- *Αποτελέσματα*
- *Συμπεράσματα*

Διάγραμμα

- **Τεχνικό Υπόβαθρο**
  - Μαρκωβιανή Διεργασία Απόφασης
  - Συνάρτηση Αξίας και Εξίσωση Bellman
  - Εύρεση Βέλτιστης Πολιτικής
- *Προσέγγιση*
- *Μάθηση*
- *Αποτελέσματα*
- *Συμπεράσματα*

Μαρκωβιανή Διεργασία Απόφασης (ΜΔΑ)

(Markov Decision Process)

- $\mathcal{S}$: Καταστάσεις (States)
- $\mathcal{A}$: Ενέργειες (Actions)
- P : Μοντέλο Μετάβασης (Transition Model)

$$P(s, a, s') = P(s'|s, a)$$

- R : Συνάρτηση Ανταμοιβής (Reward Function), $R(s, a, s')$
- γ : Συντελεστής Έκπτωσης (Discount Factor), $\gamma \in (0, 1]$

$$s_0 \xrightarrow[r_0]{a_0} s_1 \xrightarrow[r_1]{a_1} s_2 \xrightarrow[r_2]{a_2} s_3 \xrightarrow[r_3]{a_3} s_4 \quad \ldots$$

Μαρκωβιανή Ιδιότητα : Η επόμενη κατάσταση και η ανταμοιβή εξαρτώνται μόνο από την τρέχουσα κατάσταση και ενέργεια.

Πολιτικές

Μία πολιτική (policy) π απεικονίζει καταστάσεις σε ενέργειες:

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

Κάθε πολιτική χαρακτηρίζεται από την προσδοκώμενη συνολική εκπίπτουσα ανταμοιβή (ΠΣΕΑ) (expected total discounted reward) που αποκομίζει:

$$\Pi\Sigma\text{ΕΑ}(s_0) = E(r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \dots) = E\Big(\sum_t \gamma^t r_t\Big)$$

Για κάθε ΜΔΑ υπάρχει μία τουλάχιστον βέλτιστη (optimal) πολιτική π^* η οποία μεγιστοποιεί την ΠΣΕΑ σε όλες τις αρχικές καταστάσεις.

Στόχος : η εύρεση της βέλτιστης πολιτικής

Συνάρτηση Αξίας Q και Εξίσωση Bellman

Συνάρτηση Αξίας (Value Function) $Q^\pi(s,a)$

- $Q^\pi(s,a)$ είναι η *προσδοκώμενη συνολική εκπίπτουσα ανταμοιβή* ...
  - όταν ξεκινήσουμε στην κατάσταση s, ...
  - επιλέξουμε την ενέργεια a ...
  - και κατόπιν επιλέγουμε ενέργειες σύμφωνα με την π.

Εξίσωση Bellman (Bellman Equation)

$$Q^\pi(s,a) = \underbrace{\sum_{s'} P(s,a,s')R(s,a,s')}_{\text{Προσδ. ανταμ. από το πρώτο βήμα}} + \gamma \underbrace{\sum_{s'} P(s,a,s')Q^\pi\left(s',\pi(s')\right)}_{\text{Προσδ. ανταμ. από μετέπειτα βήματα}}$$

- $|\mathcal{S}| \times |\mathcal{A}|$ γραμμικές εξισώσεις με $|\mathcal{S}| \times |\mathcal{A}|$ αγνώστους
- Οι τιμές $Q^\pi(s,a)$ υπολογίζονται αναλυτικά ή επαναληπτικά

Χρήση Πινάκων

$$Q^\pi = \mathcal{R} + \gamma \mathbf{P}^\pi Q^\pi$$

- Q^π είναι ένα $(|\mathcal{S}||\mathcal{A}| \times 1)$ διάνυσμα
- $\mathbf{P}^\pi$ είναι ένας $(|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|)$ στοχαστικός πίνακας
- $\mathcal{R}$ είναι ένα $(|\mathcal{S}||\mathcal{A}| \times 1)$ διάνυσμα

$$\mathcal{R}(s,a) = \sum_{s'} P(s,a,s') R(s,a,s')$$

- Λύση: $Q^\pi = (\mathbf{I} - \gamma \mathbf{P}^\pi)^{-1} \mathcal{R}$

Επανάληψη στο Χώρο των Πολιτικών

Βρίσκει τη βέλτιστη πολιτική για κάποια ΜΔΑ.

Policy Iteration $(\mathcal{S}, \mathcal{A}, P, \mathcal{R}, \gamma, \pi_0)$

// $\mathcal{S}$: Καταστάσεις
// $\mathcal{A}$: Ενέργειες
// P : Μοντέλο μετάβασης
// $\mathcal{R}$: Συνάρτηση ανταμοιβής
// γ : Συντελεστής έκπτωσης
// π_0 : Αρχική πολιτική

$\pi' = \pi_0$
repeat
$\pi = \pi'$
$Q^\pi = (\mathbf{I} - \gamma \mathbf{P}^\pi)^{-1} \mathcal{R}$ // Εκτίμηση Πολιτικής
$\forall s \in \mathcal{S}, \pi'(s) = \arg\max_{a \in \mathcal{A}} Q^\pi(s, a)$ // Βελτίωση Πολιτικής
until $(\pi = \pi')$

return π

Πολυπλοκότητα: $O\left((|\mathcal{S}||\mathcal{A}|)^3 + |\mathcal{S}||\mathcal{A}|\right)$ χρόνος ανά επανάληψη, $O\left(|\mathcal{S}||\mathcal{A}|\right)$ χώρος

Πρακτικά Προβλήματα

Μεγάλος Χώρος Κατάστασης

- Ο ακριβής υπολογισμός των τιμών Q είναι ***πρακτικά ανέφικτος***.

$\Longrightarrow$ *Προσέγγιση Συνάρτησης Αξίας (Value Function Approximation)*

Άγνωστο Μοντέλο

- Ο ακριβής υπολογισμός των τιμών Q είναι ***αδύνατος***.

$\Longrightarrow$ *Ενισχυτική Μάθηση (Reinforcement Learning)*

Διάγραμμα

- *Τεχνικό Υπόβαθρο*
- **Προσέγγιση**
  - Προσέγγιση Συνάρτησης Αξίας
  - Γραμμικές Αρχιτεκτονικές
  - Προσέγγιση Ελαχίστων Τετραγώνων
- *Μάθηση*
- *Αποτελέσματα*
- *Συμπεράσματα*

Προσέγγιση Συνάρτησης Αξίας

Η εξαντλητική αναπαράσταση της συνάρτησης $Q^\pi(s,a)$
δεν είναι εφικτή για μεγάλα προβλήματα.

$$\Theta(|S||A|)$$

Λύση: ***Παραμετρική Προσέγγιση***

$$\widehat{Q}^\pi(s,a,w) \approx Q^\pi(s,a)$$

Γραμμικές Αρχιτεκτονικές Προσέγγισης

$$\widehat{Q}^{\pi}(s,a,w) = \sum_{i=1}^{k} \phi_i(s,a) w_i^{\pi} = \phi(s,a)^{\intercal} w^{\pi}$$

- $\phi_i(s,a)$: Συναρτήσεις βάσης (basis functions)
  - γενικά μη γραμμικές
  - γραμμικώς ανεξάρτητες
- w_i : Παράμετροι ή συντελεστές του γραμμικού συνδυασμού
- $k << |S||A|$
- Παραδείγματα: Πολυώνυμα, Ακτινικές συναρτήσεις βάσης (radial basis functions), Perceptrons, κλπ.

Προσέγγιση Ελαχίστων Τετραγώνων

$$\widehat{Q}^{\pi} = \mathbf{\Phi} w^{\pi} \approx Q^{\pi}$$

- Η συνάρτηση αξίας $\widehat{Q}^{\pi}$ πρέπει να ικανοποιεί την εξίσωση Bellman

$$\mathbf{\Phi} w^{\pi} = \underbrace{\mathbf{\Phi}\left(\mathbf{\Phi}^{\intercal}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\intercal}}_{\text{Ορθογώνια Προβολή}} \underbrace{\left(R + \gamma \mathbf{P}^{\pi}\mathbf{\Phi} w^{\pi}\right)}_{\text{Εξίσωση Bellman}}$$

$$\underbrace{\mathbf{\Phi}^{\intercal}\left(\mathbf{\Phi} - \gamma \mathbf{P}^{\pi}\mathbf{\Phi}\right)}_{A \quad (k \times k)} w^{\pi} = \underbrace{\mathbf{\Phi}^{\intercal} R}_{b}$$

- *Σταθερό σημείο* στο χώρο των συναρτήσεων αξίας

Διάγραμμα

- *Τεχνικό Υπόβαθρο*
- *Προσέγγιση*
- **Μάθηση**
  - Δειγματοληψία
  - LSQ: Μάθηση της Συνάρτησης Αξίας
  - LSPI: Least-Squares Policy Iteration
- *Αποτελέσματα*
- *Συμπεράσματα*

Μάθηση Προσέγγισης Ελαχίστων Τετραγώνων

Γνωστό Μοντέλο

$$\mathbf{A}w^{\pi} = b$$

$$\mathbf{A} = \mathbf{\Phi}^{\mathsf{T}}(\mathbf{\Phi} - \gamma\mathbf{P}^{\pi}\mathbf{\Phi}) \qquad \text{και} \qquad b = \mathbf{\Phi}^{\mathsf{T}}\mathcal{R}$$

Άγνωστο Μοντέλο

$$\widehat{\mathbf{A}}\widehat{w}^{\pi} = \widehat{b}$$

$$\widehat{\mathbf{A}} = \widehat{\mathbf{\Phi}}^{\mathsf{T}}(\widehat{\mathbf{\Phi}} - \gamma\widehat{\mathbf{P}^{\pi}\mathbf{\Phi}}) \qquad \text{και} \qquad \widehat{b} = \widehat{\mathbf{\Phi}}^{\mathsf{T}}\widehat{R}$$

$\widehat{\mathbf{\Phi}}$, $\widehat{\mathbf{P}^{\pi}\mathbf{\Phi}}$, και $\widehat{\mathcal{R}}$ δειγματοληπτικές προσεγγίσεις των $\mathbf{\Phi}$, $\mathbf{P}^{\pi}\mathbf{\Phi}$, και $\mathcal{R}$

Μια Κοντινότερη Ματιά

$$\mathbf{\Phi} = \begin{pmatrix} \phi(s_1, a_1)^\intercal \\ \dots \\ \phi(s, a)^\intercal \\ \dots \\ \phi(s_{|\mathcal{S}|}, a_{|\mathcal{A}|})^\intercal \end{pmatrix} \quad \mathbf{P}^\pi \mathbf{\Phi} = \begin{pmatrix} \sum_{s'} P(s_1, a_1, s') \phi(s', \pi(s'))^\intercal \\ \dots \\ \sum_{s'} P(s, a, s') \phi(s', \pi(s'))^\intercal \\ \dots \\ \sum_{s'} P(s_{|\mathcal{S}|}, a_{|\mathcal{A}|}, s') \phi(s', \pi(s'))^\intercal \end{pmatrix}$$

$$\mathcal{R} = \begin{pmatrix} \sum_{s'} P(s_1, a_1, s') R(s_1, a_1, s') \\ \dots \\ \sum_{s'} P(s, a, s') R(s, a, s') \\ \dots \\ \sum_{s'} P(s_{|\mathcal{S}|}, a_{|\mathcal{A}|}, s') R(s_{|\mathcal{S}|}, a_{|\mathcal{A}|}, s') \end{pmatrix}$$

Δειγματοληπτικές Προσεγγίσεις

$$D = \{(s_{d_i}, a_{d_i}, r_{d_i}, s'_{d_i}) \mid i = 1, 2, \ldots, L\}, \quad \pi$$

$$\widehat{\mathbf{\Phi}} = \begin{pmatrix} \phi(s_{d_1}, a_{d_1})^\intercal \\ \ldots \\ \phi(s_{d_i}, a_{d_i})^\intercal \\ \ldots \\ \phi(s_{d_L}, a_{d_L})^\intercal \end{pmatrix} \qquad \widehat{\mathbf{P}^\pi \mathbf{\Phi}} = \begin{pmatrix} \phi(s'_{d_1}, \pi\left(s'_{d_1}\right))^\intercal \\ \ldots \\ \phi(s'_{d_i}, \pi\left(s'_{d_i}\right))^\intercal \\ \ldots \\ \phi(s'_{d_L}, \pi\left(s'_{d_L}\right))^\intercal \end{pmatrix} \qquad \widehat{\mathcal{R}} = \begin{pmatrix} r_{d_1} \\ \ldots \\ r_{d_i} \\ \ldots \\ r_{d_L} \end{pmatrix}$$

Δειγματοληψία (Sampling)

Δείγμα (s, a, r, s')

$$\mathbf{\Phi} = \begin{pmatrix} \phi(s_1, a_1)^\intercal \\ \dots \\ \phi(s, a)^\intercal \\ \dots \\ \phi(s_{|\mathcal{S}|}, a_{|\mathcal{A}|})^\intercal \end{pmatrix} \quad \mathbf{P}^\pi \mathbf{\Phi} = \begin{pmatrix} \sum_{s'} P(s_1, a_1, s') \phi(s', \pi(s'))^\intercal \\ \dots \\ \sum_{s'} P(s, a, s') \phi(s', \pi(s'))^\intercal \\ \dots \\ \sum_{s'} P(s_{|\mathcal{S}|}, a_{|\mathcal{A}|}, s') \phi(s', \pi(s'))^\intercal \end{pmatrix}$$

$$\mathcal{R} = \begin{pmatrix} \sum_{s'} P(s_1, a_1, s') R(s_1, a_1, s') \\ \dots \\ \sum_{s'} P(s, a, s') R(s, a, s') \\ \dots \\ \sum_{s'} P(s_{|\mathcal{S}|}, a_{|\mathcal{A}|}, s') R(s_{|\mathcal{S}|}, a_{|\mathcal{A}|}, s') \end{pmatrix}$$

Σύγκλιση (Convergence)

$$\widehat{\mathbf{A}} = \widehat{\mathbf{\Phi}}^{\mathsf{T}}(\widehat{\mathbf{\Phi}} - \gamma\widehat{\mathbf{P}^{\pi}\mathbf{\Phi}}) \qquad \text{και} \qquad \widehat{b} = \widehat{\mathbf{\Phi}}^{\mathsf{T}}\widehat{R}$$

- Για ένα μεγάλο αριθμό ομοιόμορφα κατανεμημένων δειγμάτων

$$\widehat{\mathbf{A}} \approx m\mathbf{A} \qquad \text{και} \qquad \widehat{b} \approx mb$$

- το $\widehat{w}^{\pi}$ προσεγγίζει το w^{π}:

$$\widehat{w}^{\pi} = \widehat{\mathbf{A}}^{-1}\widehat{b} \quad \approx \quad (m\mathbf{A})^{-1}(mb) = \mathbf{A}^{-1}b = w^{\pi}$$

Προσαυξητικός Κάνονας Ενημέρωσης

Incremental Update Rule

- Για κάθε δείγμα (s, a, r, s')

$$\widehat{\mathbf{A}} \leftarrow \widehat{\mathbf{A}} + \phi(s,a)\Big(\phi(s,a) - \gamma\phi(s', \pi(s'))\Big)^{\mathsf{T}}$$

$$\widehat{b} \leftarrow \widehat{b} + \phi(s,a)r$$

- Παράμετροι:

$$\widehat{w}^{\pi} = \widehat{\mathbf{A}}^{-1}\widehat{b}$$

Ο Αλγόριθμος LSQ

LSQ $(D,\ k,\ \phi,\ \gamma,\ \pi)$ // Μαθαίνει τη συνάρτηση αξίας Q^π

// D : Εκπαιδευτικά Δείγματα
// k : Αριθμός συναρτήσεων βάσης
// ϕ : Συναρτήσεις βάσης
// γ : Συντελεστής έκπτωσης
// π : Πολιτική

$\widehat{\mathbf{A}} = 0$ // $(k \times k)$ πίνακας
$\widehat{b} = 0$ // $(k \times 1)$ διάνυσμα

for all $(s, a, r, s') \in D$
$\widehat{\mathbf{A}} \leftarrow \widehat{\mathbf{A}} + \phi(s,a)\Big(\phi(s,a) - \gamma\phi(s', \pi(s'))\Big)^\intercal$
$\widehat{b} \leftarrow \widehat{b} + \phi(s,a)r$

$\widehat{w}^\pi = \widehat{\mathbf{A}}^{-1}\widehat{b}$

return $\widehat{w}^\pi$

Πολυπλοκότητα: $O\left(|D|(k^2 + k|\mathcal{A}|) + k^3\right)$ χρόνος, $O\left(|D| + k^2\right)$ χώρος

Least-Squares Policy Iteration (LSPI)

LSPI $(k,\ \phi,\ \gamma,\ \epsilon,\ \pi_0,\ D)$ // Μαθαίνει μια καλή πολιτική από δείγματα

// k : Αριθμός συναρτήσεων βάσης
// ϕ : Συναρτήσεις βάσης
// γ : Συντελεστής έκπτωσης
// ϵ : Κριτήριο τερματισμού
// π_0 : Αρχική πολιτική, δοσμένη ως w_0, $\pi_0 = \pi(s, w_0)$ (συνήθως $w_0 = 0$)
// D : Εκπαιδευτικά δείγματα

$\pi' = \pi_0$ // Κατ' ουσίαν, $w' = w_0$

repeat
- $\pi = \pi'$ // $w = w'$
- $w' =$ **LSQ** $(D,\ k,\ \phi,\ \gamma,\ \pi)$ // Εκτίμηση Πολιτικής
- $\pi'(s) = \arg\max_a \phi(s,a)^\intercal w'$ // Βελτίωση Πολιτικής

until $(\pi \approx \pi')$ // τουτέστιν, $(||w - w'|| < \epsilon)$

return π // **return** w

Πολυπλοκότητα: $O\left(k^2|D||\mathcal{A}| + k^3\right)$ χρόνος ανά επανάληψη, $O\left(|D| + k^2\right)$ χώρος

Χαρακτηριστικά του Αλγορίθμου LSPI

Πετυχαίνει τους στόχους μας ...

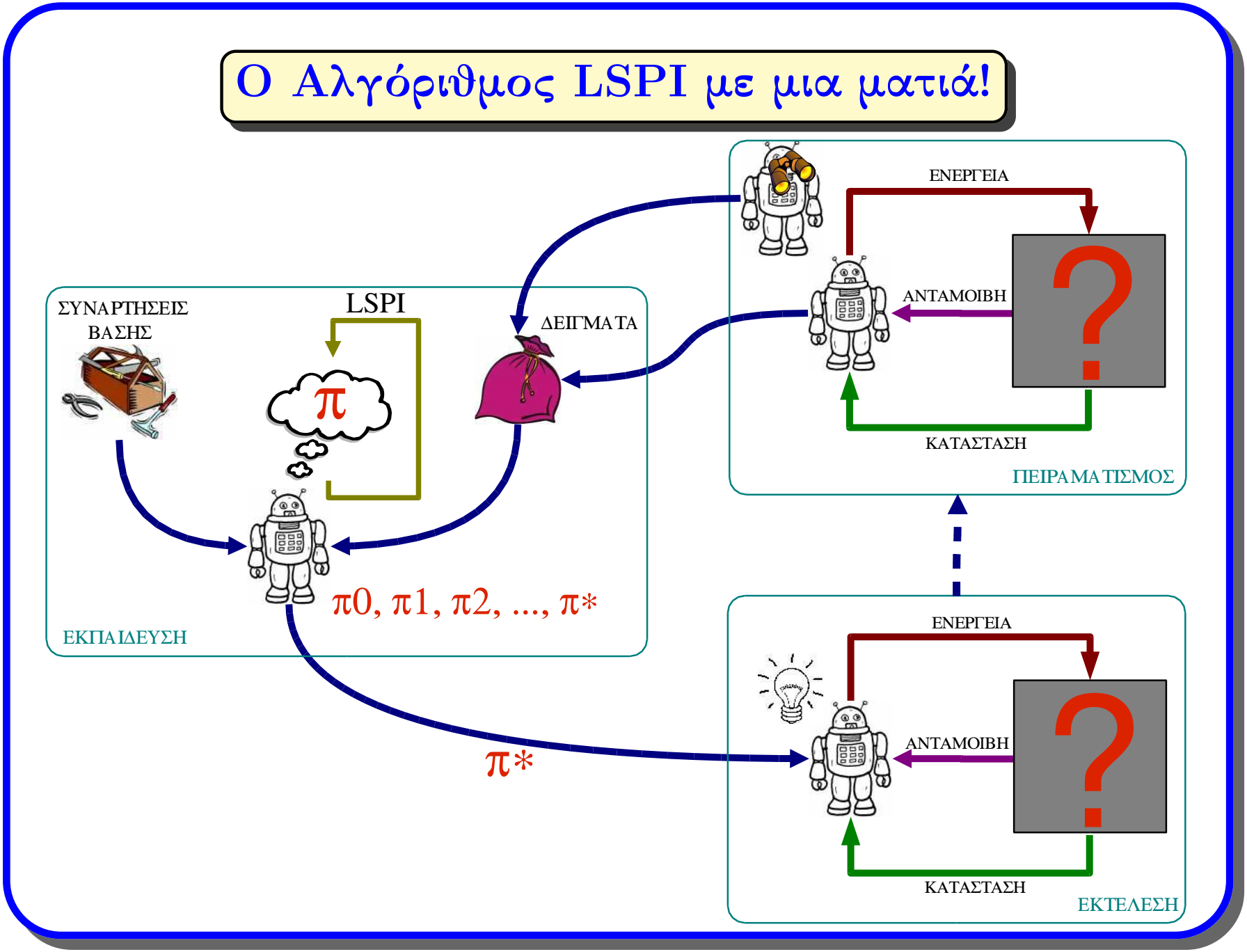
- Μαθαίνει γρήγορα √
- Χρησιμοποιεί τα δείγματα πολύ αποδοτικά √
- Αντιμετωπίζει επιτυχώς προβλήματα μεγάλης κλίμακας √

Επιπρόσθετα ...

- Επιλύει προβλήματα ελέγχου και πρόβλεψης.
- Δέχεται δείγματα από οποιαδήποτε πηγή.
- Επιτρέπει τη διερεύνηση εναλλακτικών αναπαραστάσεων.
- Είναι απλός και εύκολος στην υλοποίηση.

Μειονεκτήματα του Αλγορίθμου LSPI

- Ανομοιογενώς κατανεμημένα δείγματα
  - Μπορεί να ταλαντεύεται επ᾽ αόριστον
- Ανεπαρκείς συναρτήσεις βάσης
  - Μπορεί να συγκλίνει σε κάποια απαράδεκτη πολιτική
- Προσεγγιστικός αλγόριθμος
  - δεν μπορεί να εγγυηθεί σύγκλιση
  - δεν μπορεί να εγγυηθεί βέλτιστη λύση
- Καινούριο πρόβλημα : επιλογή συναρτήσεων βάσης


Ο Αλγόριθμος LSPI με μια ματιά!
ΕΝΕΡΓΕΙΑ
ΑΝΤΑΜΟΙΒΗ
ΚΑΤΑΣΤΑΣΗ
ΠΕΙΡΑΜΑΤΙΣΜΟΣ
ΣΥΝΑΡΤΗΣΕΙΣ ΒΑΣΗΣ
LSPI
ΔΕΙΓΜΑΤΑ
π
π0, π1, π2, ..., π*
ΕΚΠΑΙΔΕΥΣΗ
π*
ΕΝΕΡΓΕΙΑ
ΑΝΤΑΜΟΙΒΗ
ΚΑΤΑΣΤΑΣΗ
ΕΚΤΕΛΕΣΗ

Διάγραμμα

- *Τεχνικό Υπόβαθρο*
- *Προσέγγιση*
- *Μάθηση*
- **Αποτελέσματα**
  - Αντεστραμμένο Εκκρεμές
  - Επιλογή Αλγορίθμων (LSQL)
  - Ισορροπία και Οδήγηση Ποδηλάτου
  - Το Παιχνίδι Tetris
  - Διαχείριση Συστήματος
  - Ποδόσφαιρο
- *Συμπεράσματα*

Ισορροπία και Οδήγηση Ποδηλάτου

Ισορρόπησε και οδήγησε το ποδήλατο στο τέρμα 1 km μακρυά!

- $\mathcal{S} = \{(\theta, \dot{\theta}, \omega, \dot{\omega}, \ddot{\omega}, \psi)\}$

  θ: γωνία του τιμονιού, ω: κατακόρυφη γωνία, ψ: γωνία προς το τέρμα

- $\mathcal{A} = \{(\tau, \upsilon)\}$

  $\tau \in \{-2, 0, +2\}$: ροπή, $\upsilon \in \{-0.02, 0, +0.02\}$: μετατόπιση

- Θόρυβος n : $(\tau, \upsilon + n)$, $n \in [-0.02, +0.02]$

- Ανταμοιβή:
  - η μεταβολή στο ω^2, συν ...
  - 1% της μεταβολής στην απόσταση από το τέρμα

- $\gamma = 0.8$

Ποδήλατο: Παράμετροι Μάθησης

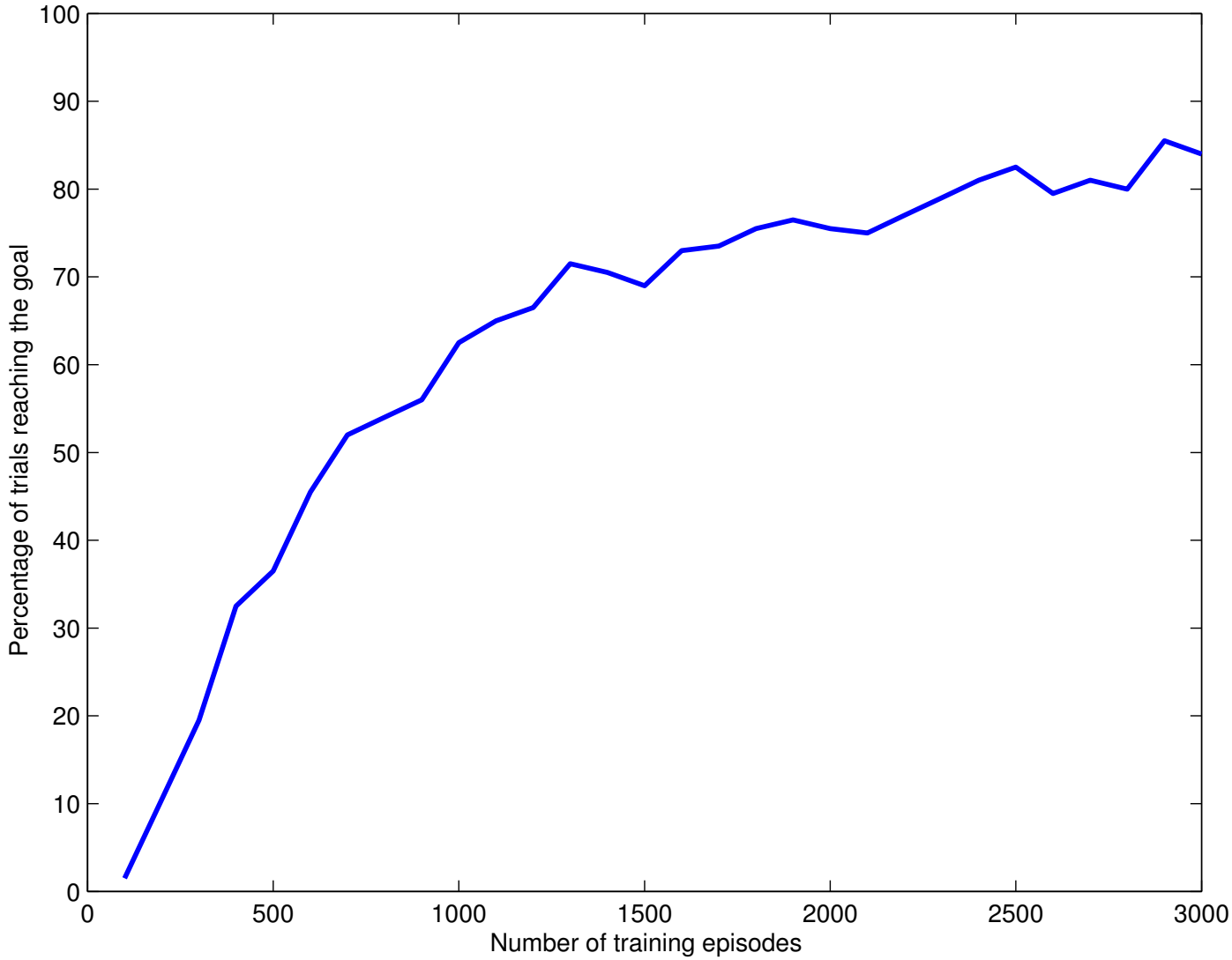
- Προσέγγιση: $k = 100$ (20 συναρτήσεις βάσης για κάθε ενέργεια)

  $(1, \omega, \dot{\omega}, \omega^2, \dot{\omega}^2, \omega\dot{\omega}, \theta, \dot{\theta}, \theta^2, \dot{\theta}^2, \theta\dot{\theta}, \omega\theta, \omega\theta^2, \omega^2\theta, \psi, \psi^2, \psi\theta, \bar{\psi}, \bar{\psi}^2, \bar{\psi}\theta)$

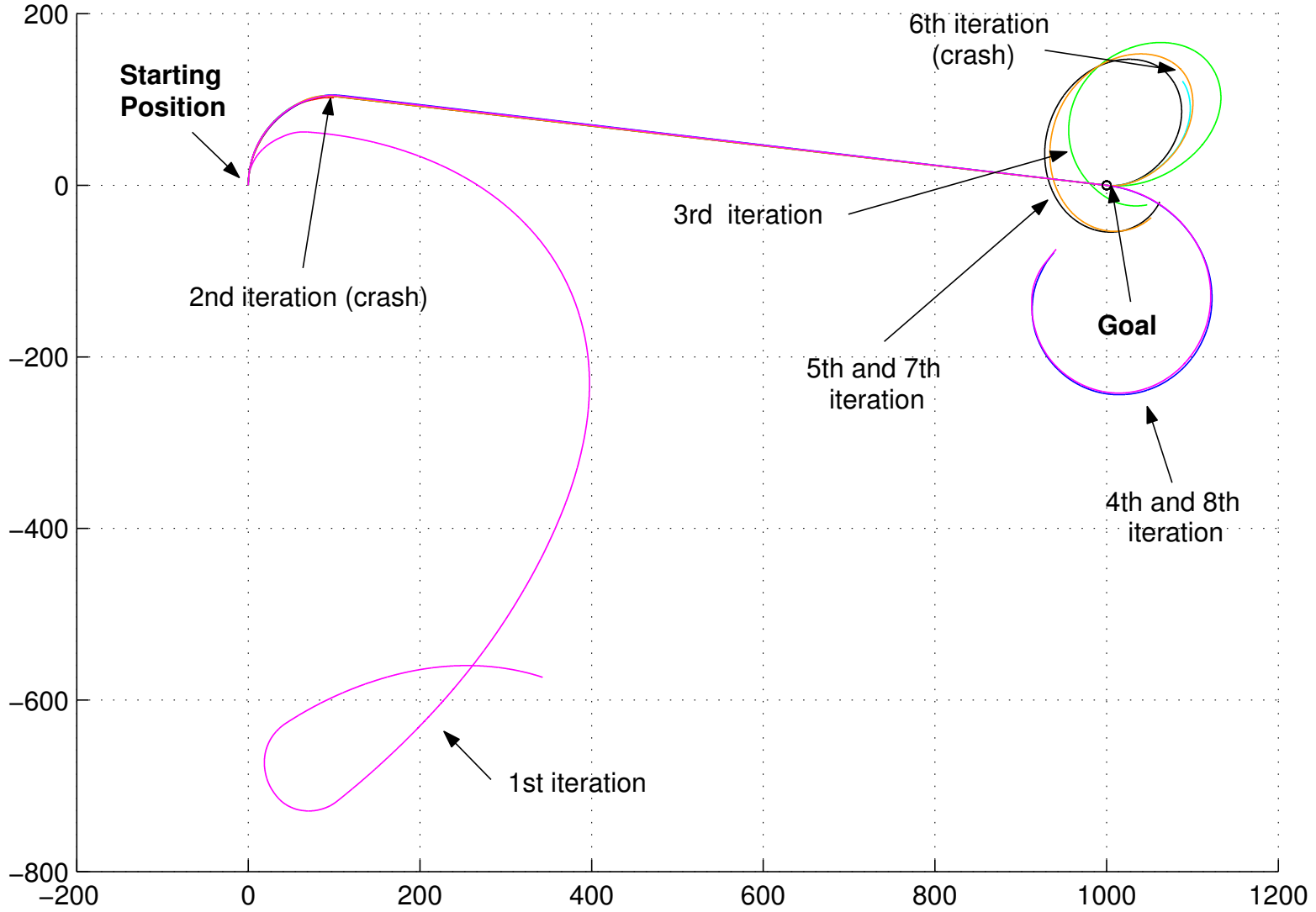
  $$\bar{\psi} = \pi - \psi \quad \text{εάν} \quad \psi > 0 \qquad \text{και} \qquad \bar{\psi} = -\pi - \psi \quad \text{εάν} \quad \psi < 0$$

- Δείγματα: Συλλογή απο 'τυχαία' επεισόδια τα οποία ...
  - ξεκινούν από μια τυχαία κατάσταση ...
  - κοντά στην αρχική κατάσταση και ...
  - ακολουθούν μια τελείως τυχαία πολιτική.

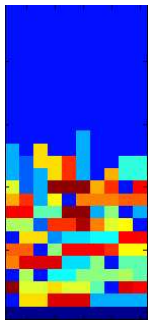
Ποδήλατο: Ποσοστό Επιτυχών Πολιτικών

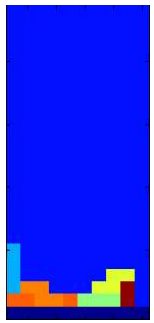


Ποδήλατο: Πολιτικές από μία Εκτέλεση του LSPI



Το Παιχνίδι Tetris





Μάθε να παίζεις καλά Tetris!

- $\mathcal{S}$: $\approx 10^{61}$ καταστάσεις
- $\mathcal{A}$: ≈ 40 ενέργειες
- Θόρυβος: το επόμενο αντικείμενο επιλέγεται τυχαία
- Ανταμοιβή: $+1$ για κάθε συμπληρωμένη γραμμή, 0 διαφορετικά
- $\gamma = 1$

Tetris: Παράμετροι Μάθησης

- Προσέγγιση: $k = 10$ συναρτήσεις βάσης με ορίσματα (s, a)
  1. ο σταθερός όρος 1.0
  2. ο αριθμός συμπληρωμένων γραμμών εάν επιλεγεί a στην s
  3. το μέγιστο ύψος
  4. η μεταβολή στο #3
  5. ο συνολικός αριθμός 'κενών'
  6. η μεταβολή στο #5
  7. το μέσο ύψος
  8. η μεταβολή στο #7
  9. το άθροισμα των απολύτων διαφορών ύψους μεταξύ γειτονικών στηλών
  10. η μεταβολή στο #9

- Δείγματα: συλλογή ολοκληρωμένων παιχνιδιών από ...
  - έναν 'εμπειρικό παίκτη' ο οποίος ...
  - συγκεντρώνει περίπου 675 πόντους ανά παιχνίδι
  - Η 'τυχαία' πολιτική δεν παρέχει επαρκή κάλυψη

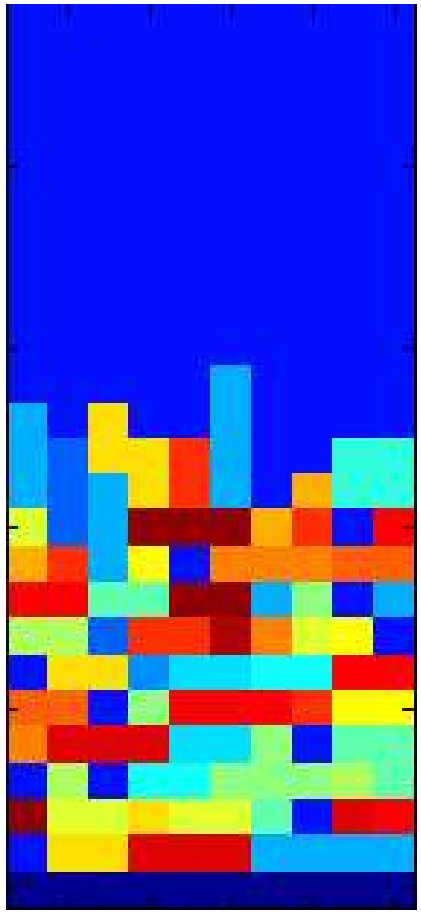
Tetris: Αποτελέσματα

Εμπειρικός Παίκτης

Μέσο σκόρ ≈ 675

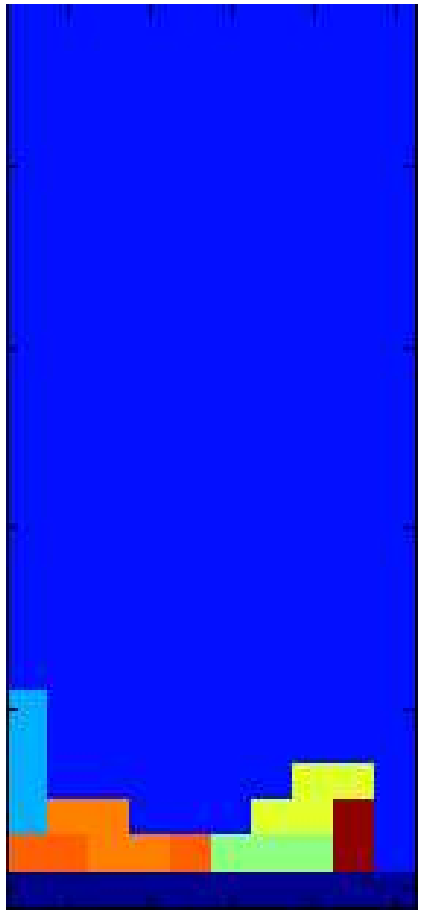
Πόντοι = 49

(150 βήματα)



Τελικός Παίκτης

Μέσο σκόρ ≈ 3000



Πόντοι = 58

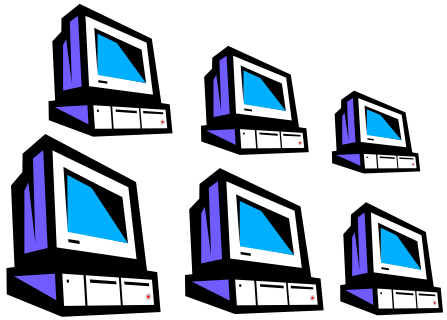
(150 βήματα)

Tetris: Σύγκριση με τη Μέθοδο NDP

- Νευρο-Δυναμικός Προγραμματισμός για Tetris
  [VanRoy, 1995], [Bertsekas and Tsitsiklis, 1996]
- Τα τελικά αποτελέσματα είναι συγκρίσιμα
- Οι μέθοδοι όμως είναι τελείως διαφορετικές

λ-Policy Iteration	**LSPI**
Χρήση της συνάρτησης αξίας $V(s)$	Χρήση της συνάρτησης αξίας $Q(s,a)$
22 συναρτήσεις βάσης	10 συναρτήσεις βάσης
Monte-Carlo εκτίμηση	Bellman εκτίμηση
Προσέγγιση ελαχιστοποίησης απόκλισης	Προσέγγιση σταθερού σημείου
Μοντέλο για επιλογή ενεργειών	Μοντέλο για βελτίωση προσέγγισης
Συλλογή δειγμάτων σε κάθε επανάληψη	Συλλογή δειγμάτων μόνο μία φορά
Η επανάληψη δε συγκλίνει	Η επανάληψη συγκλίνει

Διαχείριση Συστήματος Υπολογιστών



Μάθε να διαχειρίζεσαι το σύστημα ώστε να μεγιστοποιηθεί ο αριθμός των εργασιών που εκπληρώνονται επιτυχώς!

- $\mathcal{S}$: κατάσταση λειτουργίας και φόρτος
- $\mathcal{A}$: επανεκκίνηση (reboot) ή όχι
- Θόρυβος: άφιξη εργασιών, βλάβες και αλληλεπίδραση μηχανών
- Ανταμοιβή: $+1$ για κάθε εργασία που τελειώνει επιτυχώς
- $\gamma = 0.95$

Σύστημα Υπολογιστών: Μάθηση

LSPI για Πολυπρακτορικά Συστήματα

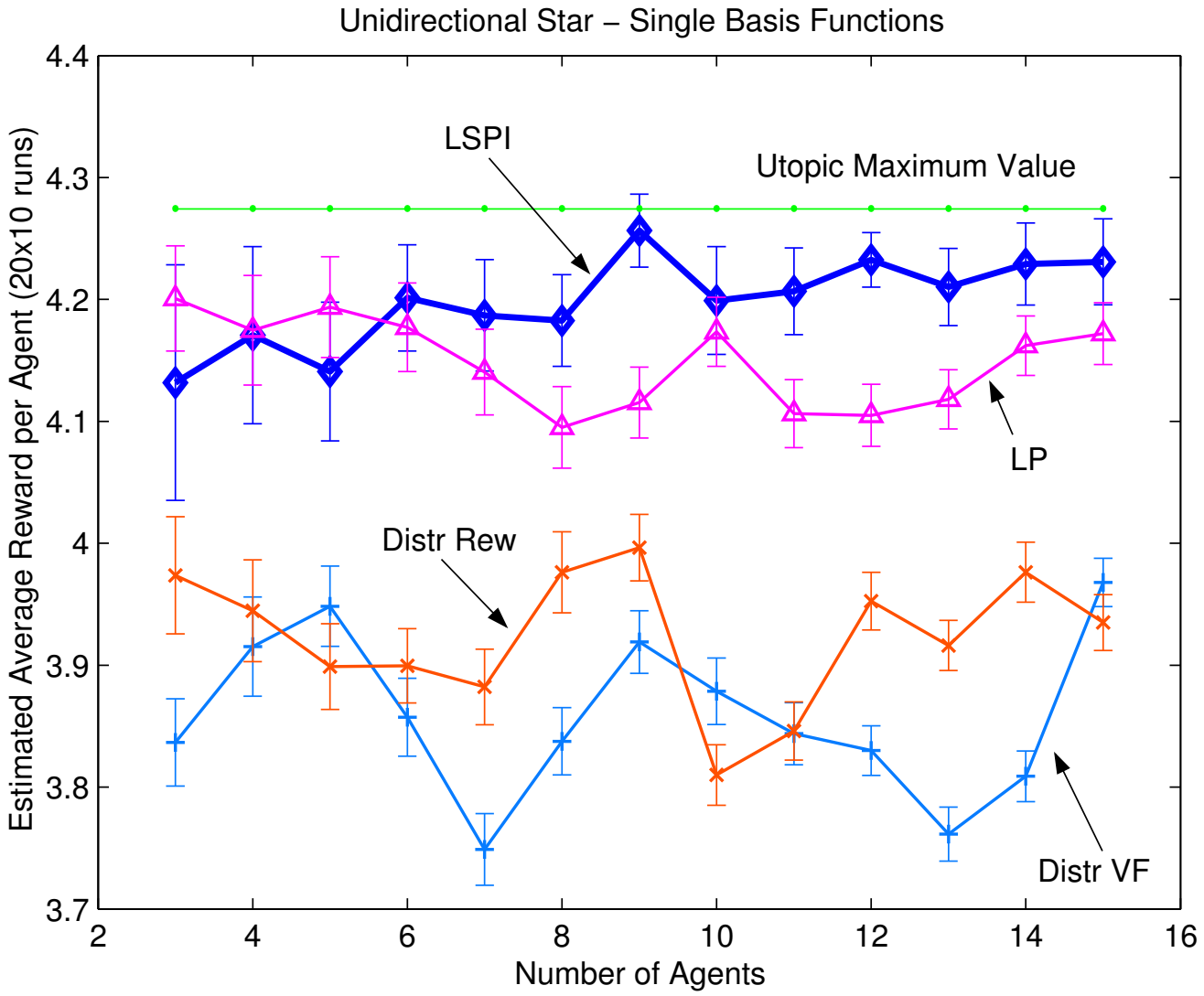
[Guestrin, Lagoudakis, and Parr, 2002]

- n μηχανές: $|S|^n$ καταστάσεις και $|A|^n$ ενέργειες
- Συνεργατική επιλογή ενεργειών (Collaborative Action Selection) [Guestrin, Koller, Parr, 2001].
- Αξιοποίηση της δομής στο σύστημα

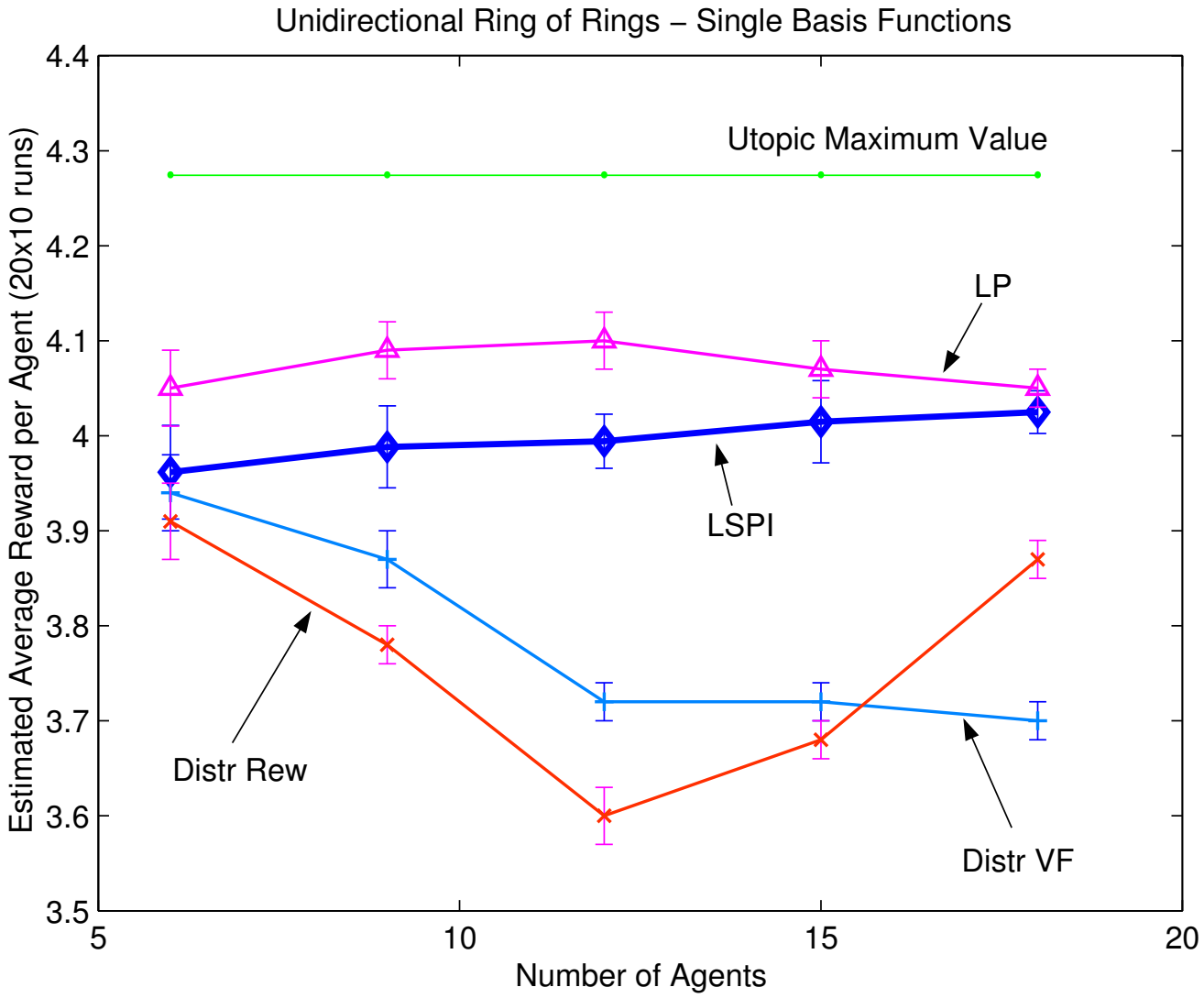
Παράμετροι Μάθησης

- Δείγματα: συλλογή απο 'τυχαία' επεισόδια
- Προσέγγιση: συναρτήσεις βάσης από [Guestrin, Koller, Parr, 2001]

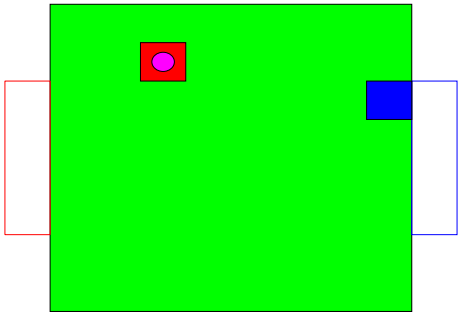
Σύστημα Υπολογιστών: Τοπολογία Αστέρα



Συστ. Υπολ.: Τοπολογία Δακτυλίου Δακτυλίων



Ποδόσφαιρο (Πολύ Απλουστευμένο)



Μάθε να παίζεις ποδόσφαιρο μ' έναν άγνωστο αντίπαλο!

- $\mathcal{S}$: $\{(x_1, y_1, x_2, y_2, b)\}$, θέσεις των παικτών και της μπάλλας
- $\mathcal{A}$: $\{\Uparrow, \Downarrow, \Rightarrow, \Leftarrow, *\}$, 5 ενέργειες για κάθε παίκτη
- Θόρυβος: Η προτεραιότητα σε κάθε βήμα καθορίζεται τυχαία.
- Ανταμοιβή: $+1/-1$ όταν σκοράρει κάποιος παίκτης, 0 διαφορετικά
- $\gamma = 0.9$ (όσο νωρίτερα σκοράρεις, τόσο καλύτερα)

Ποδόσφαιρο: Μάθηση

LSPI για Μαρκωβιανά Παιχνίδια Μηδενικού Αθροίσματος

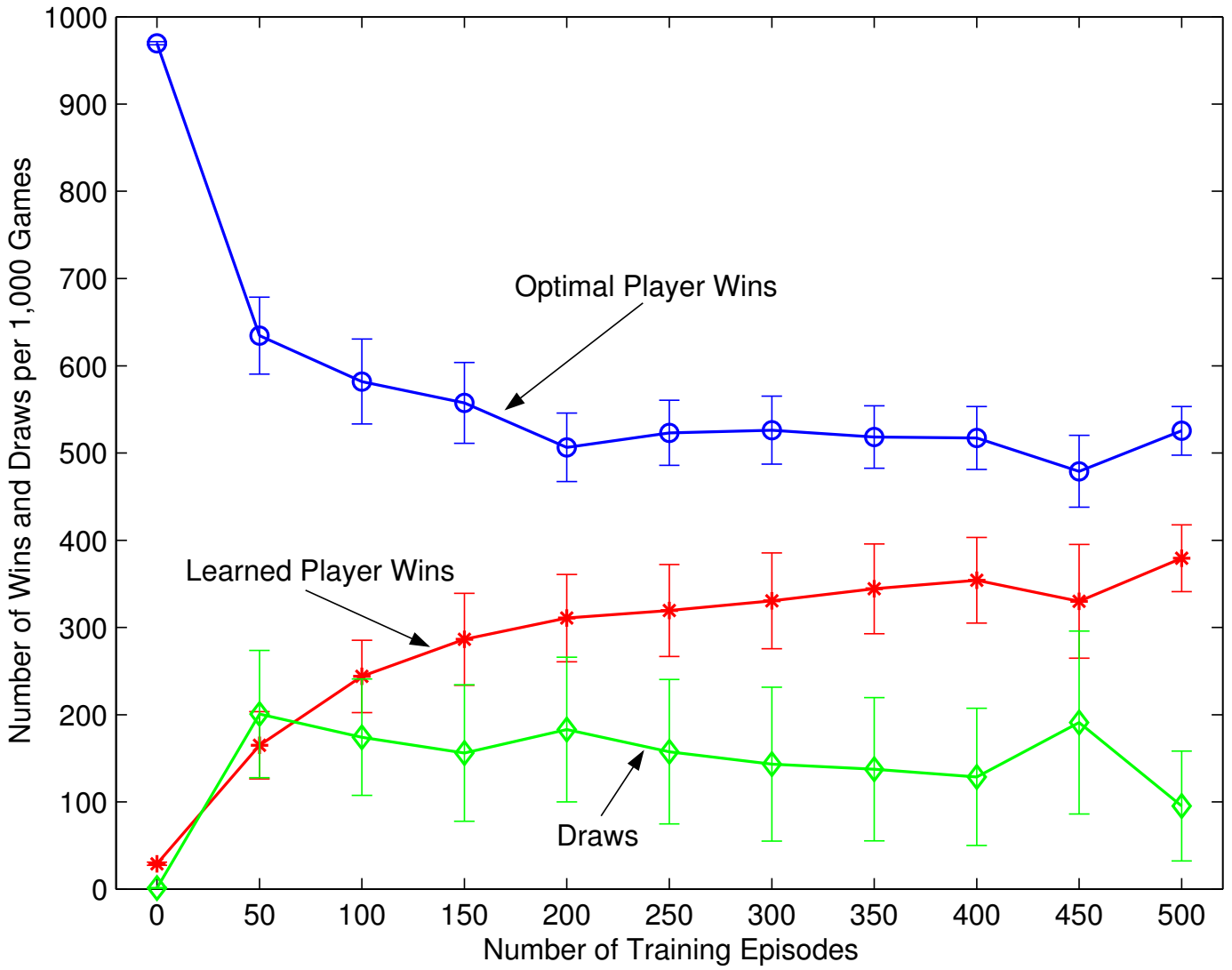
[Lagoudakis and Parr, 2002 (submitted)]

- Ο ***max*** παίκτης παίζει εναντίον του ***min*** παίκτη.
- Η βέλτιστη πολιτική είναι κατά κανόνα στοχαστική και ...
- *μεγιστοποιεί την ανταμοιβή στη χειρότερη περίπτωση* (minimax).

Παράμετροι Μάθησης

- Προσέγγιση: $k = 900$ (34 συναρτήσεις βάσης για κάθε ενέργεια)
- Δείγματα: Συλλογή απο 'τυχαία' παιχνίδια

Ποδόσφαιρο: Νίκες, Ισοπαλίες και Ήττες



Διάγραμμα

- *Τεχνικό Υπόβαθρο*
- *Προσέγγιση*
- *Μάθηση*
- *Αποτελέσματα*
- **Συμπεράσματα**

Εν Κατακλείδι ...

Τι παρατηρήσαμε

- Οι αλγόριθμοι για έλεγχο είναι αργοί
- Δυσκολεύονται σε πραγματικά προβλήματα

Τι επινοήσαμε

- LSQL: Least-Squares Q-Learning
- LSPI: Least-Squares Policy Iteration

Τι πετύχαμε

- Γρήγορη ενισχυτική μάθηση για έλεγχο
- Αντιμετώπιση προβλημάτων μεγάλης κλίμακας

Ευχαριστίες

Ευχαριστώ

όλους όσους συνετέλεσαν στην εκπλήρωση αυτής της εργασίας:

- Carlos Guestrin, Stanford University (συνεργασία)
- Αριστείδη Λίκα, Πανεπιστήμιο Ιωαννίνων (μεταφραστική βοήθεια)
- Γιάννη Ρεφανίδη, Αριστοτέλειο Πανεπιστήμιο (πρόσκληση)
- Ίδρυμα Λίλιαν Βουδούρη (οικονομική ενίσχυση)

και όλους εσάς που παρακολουθήσατε!

Περαιτέρω Πληροφορίες

- Michail G. Lagoudakis and Michael L. Littman
  Algorithm Selection using Reinforcement Learning
  ICML-2000: International Conference on Machine Learning,
  Stanford University, Palo Alto, CA, June 2000.
- Michail G. Lagoudakis and Ronald E. Parr
  Model-Free Least-Squares Policy Iteration
  *NIPS*2001: Neural Information Processing Systems*,
  Vancouver, BC, December 2001.

Διαθέσιμα on-line: `http://www.cs.duke.edu/~mgl`