Μελέτη και Υλοποίηση Ελεγκτών Ρομποτικών Συστημάτων με χρήση Αλγορίθμων Ενισχυτικής Μάθησης

Κόντες Γεώργιος

Πολυτεχνείο Κρήτης

22 Ιουλίου, 2009

Διάρθρωση

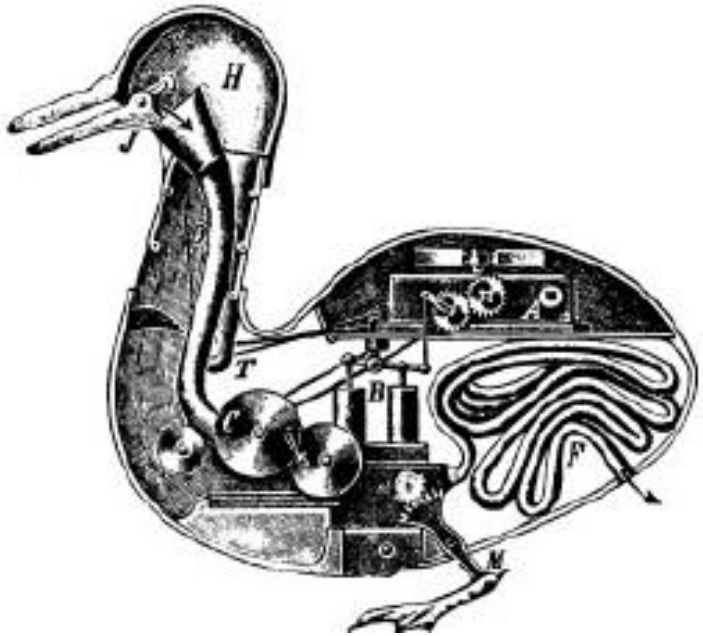
- Εισαγωγή
- Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)
- Ενισχυτική Μάθηση
- Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης
- Αλγόριθμοι Πολιτικής Κλίσης
- Ο αλγόριθμος ΕΜ
- Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού
- Πειράματα
- Συμπεράσματα

Διάρθρωση

- Εισαγωγή
- Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)
- Ενισχυτική Μάθηση
- Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης
- Αλγόριθμοι Πολιτικής Κλίσης
- Ο αλγόριθμος ΕΜ
- Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού
- Πειράματα
- Συμπεράσματα

Εισαγωγή

- Ρομπότ στην αρχαιότητα $\Longrightarrow$ εργασία, ψυχαγωγία



Εισαγωγή

- Σήμερα υπάρχει πληθώρα ρομπότ $\Longrightarrow$ οχήματα, ιπτάμενα, υποβρύχια, βιομηχανικά...

Εισαγωγή

- Τελικός στόχος $\Longrightarrow$ Πραγματικά αυτόνομα ρομπότ
- Αυτονομία $\Longrightarrow$ Προσαρμοστικότητα $\Longrightarrow$ Μάθηση
- Ενισχυτική Μάθηση

Διάρθρωση

- Εισαγωγή
- Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)
- Ενισχυτική Μάθηση
- Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης
- Αλγόριθμοι Πολιτικής Κλίσης
- Ο αλγόριθμος ΕΜ
- Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού
- Πειράματα
- Συμπεράσματα

Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)

- Υπόθεση Μάρκοβ (π.χ. αυτοκίνητο)
  - αν $x_t \longrightarrow$ είναι η τρέχουσα κατάσταση
  - αν $u_t \longrightarrow$ τρέχουσα ενέργεια που εφαρμόζεται
  - $Pr\{x_{t+1} = x' | x_t, u_t, x_{t-1}, u_{t-1}, \ldots, x_0, u_0\} = Pr\{x_{t+1} = x' | x_t, u_t\}$

Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)

- Στοχαστική διεργασία διακριτού χρόνου
- Περιγράφεται από το διάνυσμα (X, U, T, R, γ, D)
  - $X \longrightarrow$ Ο χώρος καταστάσεων του προβλήματος
  - $U \longrightarrow$ Ο χώρος των ενεργειών ή αποφάσεων της διεργασίας
  - $T \longrightarrow$ Το Μοντέλο Μετάβασης (Transition Model) της διεργασίας $\longrightarrow T(\boldsymbol{x}, u, \boldsymbol{x}') = T(\boldsymbol{x}' | \boldsymbol{x}, u)$

Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)

- Στοχαστική διεργασία διακριτού χρόνου
- Περιγράφεται από το διάνυσμα (X, U, T, R, γ, D)
  - $R \longrightarrow$ Η συνάρτηση Ανταμοιβής (Reward Function) ή Συνάρτηση Κόστους (Cost Function) $\longrightarrow R : X \times U \times X \rightarrow \mathbb{R}$. Συνηθέστερες μορφές: $r_t = R(\boldsymbol{x}, u, \boldsymbol{x}')$,$r_t = R(\boldsymbol{x}, u)$, $r_t = R(\boldsymbol{x})$
  - $\gamma \in (0, 1] \longrightarrow$ Ο παράγοντας έκπτωσης (discount factor) της διεργασίας
  - $D \longrightarrow$ Καθορίζει την αρχική κατάσταση της διεργασίας.

Διάρθρωση

- Εισαγωγή
- Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)
- Ενισχυτική Μάθηση
- Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης
- Αλγόριθμοι Πολιτικής Κλίσης
- Ο αλγόριθμος ΕΜ
- Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού
- Πειράματα
- Συμπεράσματα

Ενισχυτική Μάθηση

- Πολιτική (Policy) ή ελεγκτής (controller) = $\pi(\boldsymbol{x}) : X \longrightarrow U$ από καταστάσεις σε ενέργειες

- Μαρκοβιανό Σύστημα

$$\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0), \tag{1}$$
$$\boldsymbol{x}_{t+1} \sim p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{u}_t, t), \tag{2}$$
$$r_{t+1} = R(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{x}_{t+1}, t+1), \tag{3}$$
$$\boldsymbol{u}_t \sim \pi_{\boldsymbol{\theta}}(\boldsymbol{u}_t|\boldsymbol{x}_t, t) = p(\boldsymbol{u}_t|x_t, t, \boldsymbol{\theta}), \tag{4}$$

- Τροχιές

$$\boldsymbol{x}_0 \xrightarrow[r_0]{u_0} \boldsymbol{x}_1 \xrightarrow[r_1]{u_1} \boldsymbol{x}_2 \xrightarrow[r_2]{u_2} \boldsymbol{x}_3 \xrightarrow[r_3]{u_3} \boldsymbol{x}_4 \xrightarrow[r_4]{u_4} \quad \dots \quad \boldsymbol{x}_{H-1} \xrightarrow[r_{H-1}]{u_{H-1}} \boldsymbol{x}_H \tag{5}$$

Ενισχυτική Μάθηση

- Στόχος $\Longrightarrow$ Εύρεση πολιτικής $\pi^*_{\boldsymbol{\theta}}$, βέλτιστη προς:

$$J(\boldsymbol{\theta}) = \mathbb{E}\left\{\sum_{t=0}^{H} \gamma^t r_t; \boldsymbol{\theta}\right\} \tag{6}$$

Διάρθρωση

- Εισαγωγή
- Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)
- Ενισχυτική Μάθηση
- Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης
- Αλγόριθμοι Πολιτικής Κλίσης
- Ο αλγόριθμος ΕΜ
- Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού
- Πειράματα
- Συμπεράσματα

Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης

- Συναρτήσεις Αξιολόγησης:

$$V^{\pi}(x) = \mathbb{E}_{\pi}\left\{\sum_{t=0}^{H} \gamma^{t} r_t | x_0 = x\right\} \tag{7}$$

και

$$Q^{\pi}(x,u) = \mathbb{E}_{\pi}\left\{\sum_{t=0}^{H} \gamma^{t} r_t | x_0 = x, u_0 = u\right\} \tag{8}$$

- Άπληστη Πολιτική

$$\pi'(x) = \arg\max_{u \in U} Q^{\pi}(x,u) \tag{9}$$

με

$$\forall x \in X, \quad Q^{\pi'}(x, \pi'(x)) \geq Q^{\pi}(x, \pi(x)) \tag{10}$$

Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης

- Γραμμική Εξίσωση Bellman

$$Q^{\pi}(x,u) = r(x,u) + \gamma \sum_{x' \in X} T(x,u,x') Q^{\pi}(x', \pi(x')) \tag{11}$$

- Bellman Optimality Equation

$$Q^{*}(x,u) = r(x,u) + \gamma \sum_{x' \in X} T(x,u,x') \max_{u' \in U} Q^{*}(x',u') \tag{12}$$

- Value Iteration $\Longrightarrow$ Bellman Optimality Equation

- Policy Iteration $\Longrightarrow$ Γραμμική Εξίσωση Bellman

Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης

- Προβλήματα:
  - Γνώση του Μοντέλου
  - Διακριτές Ενέργειες
- Λύσεις:
  - Συναρτήσεις Βάσης

$$Q^{\pi}(\boldsymbol{x}, \boldsymbol{u}) = \phi(\boldsymbol{x}, \boldsymbol{u})^{T} \boldsymbol{w} \tag{13}$$

  - Neural Networks
  - Βέλτιστες - Δυναμικές Διακριτοποιήσεις

Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης

- Αστάθεια Μάθησης
- Διακριτές Ενέργειες

Διάρθρωση

- Εισαγωγή
- Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)
- Ενισχυτική Μάθηση
- Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης
- Αλγόριθμοι Πολιτικής Κλίσης
- Ο αλγόριθμος ΕΜ
- Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού
- Πειράματα
- Συμπεράσματα

Πρόβλημα

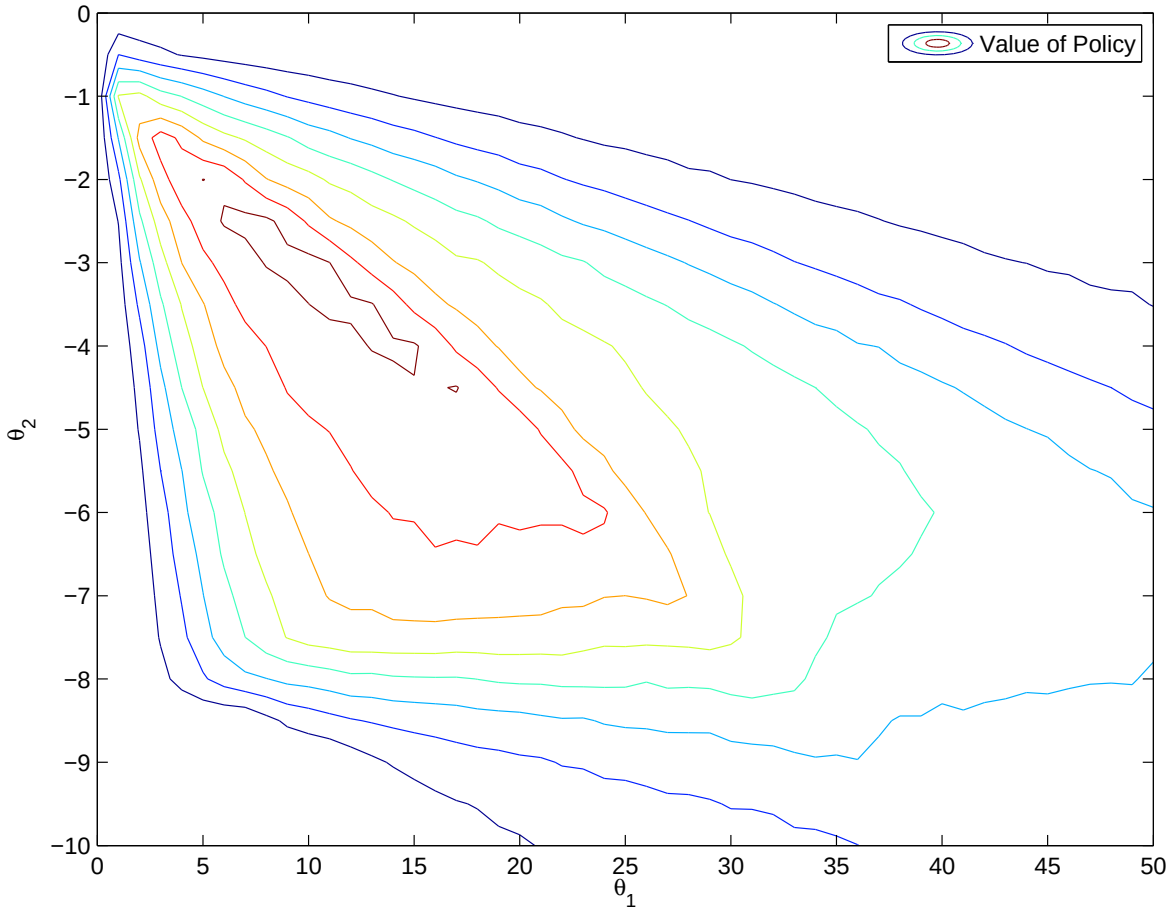
- Διδιάστατο Συνθετικό Πρόβλημα
- x_1=θέση, x_2=ταχύτητα, u=επιτάχυνση
- Δυναμικό Μοντέλο

$$x_2(t+1) = x_2(t) - u(t) + \kappa, \tag{14}$$
$$x_1(t+1) = x_1(t) - 0.1x_2(t+1) + \kappa, \tag{15}$$

- $\sigma_\kappa = 0.003$
- $u_t = 1/[1 + \exp(-(\boldsymbol{\theta} + \varepsilon_t)^T \boldsymbol{x})] - 0.5$
- $\boldsymbol{x}_0 = [1, 0] + noise$, $x_{goal} = [0, 0]$
- $r(t) = \exp(-0.5||\boldsymbol{x}(t)||_2/(0.03)^2)$

Πρόβλημα

- Η συνάρτηση προς βελτιστοποίηση



Αλγόριθμοι Πολιτικής Κλίσης

- Απ' ευθείας στο χώρο τον πολιτικών

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k \bigtriangledown_{\boldsymbol{\theta}} J(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k} \tag{16}$$

- Μέθοδος Διακριτών Διαφορών

$$\Delta\hat{J} \approx J(\boldsymbol{\theta}_k + \Delta\boldsymbol{\theta}) - J_{ref}{}^{1} \tag{17}$$

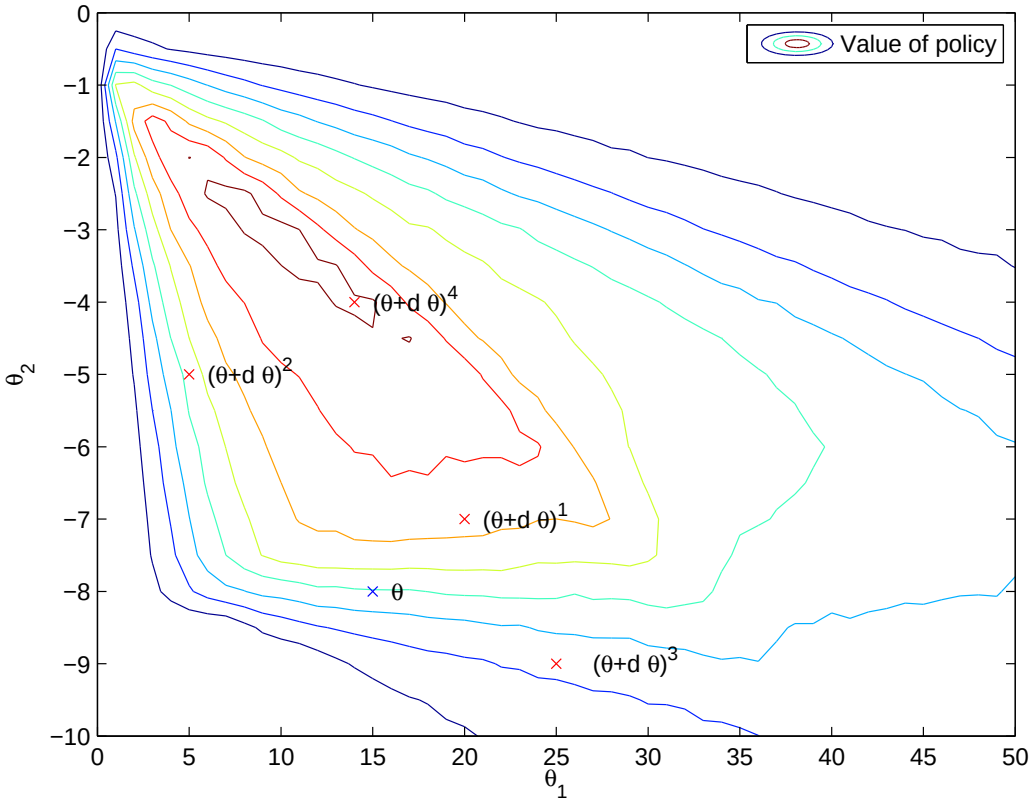
και στην συνέχεια

$$\boldsymbol{g}_{FD} = (\Delta\boldsymbol{\theta}^T \Delta\boldsymbol{\theta})^{-1} \Delta\boldsymbol{\theta}^T \Delta\hat{J} \tag{18}$$

[1] Ως J_{ref} χρησιμοποιείται συνήθως το $J(\boldsymbol{\theta}_k)$ ή το $J(\boldsymbol{\theta}_k - \Delta\boldsymbol{\theta})$

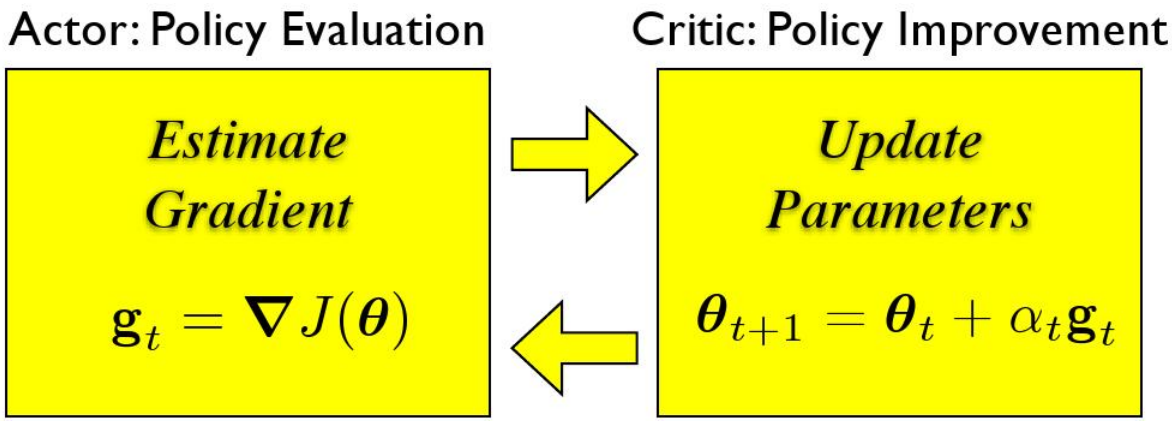
Πρόβλημα

- Η μέθοδος διακριτών διαφορών



Αλγόριθμοι Πολιτικής Κλίσης

- Actor - Critic



Αλγόριθμοι Πολιτικής Κλίσης

- Θετικά:
  - Συνεχείς καταστάσεις - ενέργειες
  - Ελεγκτές $\longrightarrow$ Φυσική σημασία για το σύστημα
- Αρνητικά:
  - Ρύθμιση των ρυθμών μάθησης
  - Τοπικά ελάχιστα

Διάρθρωση

- Εισαγωγή
- Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)
- Ενισχυτική Μάθηση
- Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης
- Αλγόριθμοι Πολιτικής Κλίσης
- Ο αλγόριθμος EM
- Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού
- Πειράματα
- Συμπεράσματα

Ο αλγόριθμος ΕΜ

- Έστω παρατηρήσιμες μεταβλητές $\boldsymbol{X}$, κρυφές μεταβλητές $\boldsymbol{Z}$, παράμετροι $\boldsymbol{\theta}$
- Σκοπός:

$$\max p(\boldsymbol{X}|\boldsymbol{\theta}) = \max \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{19}$$

- Ορίζουμε $q(\boldsymbol{Z})$

Ο αλγόριθμος ΕΜ

- $\forall \quad q(\boldsymbol{Z})$ ισχύει:

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p), \tag{20}$$

όπου

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} \tag{21}$$

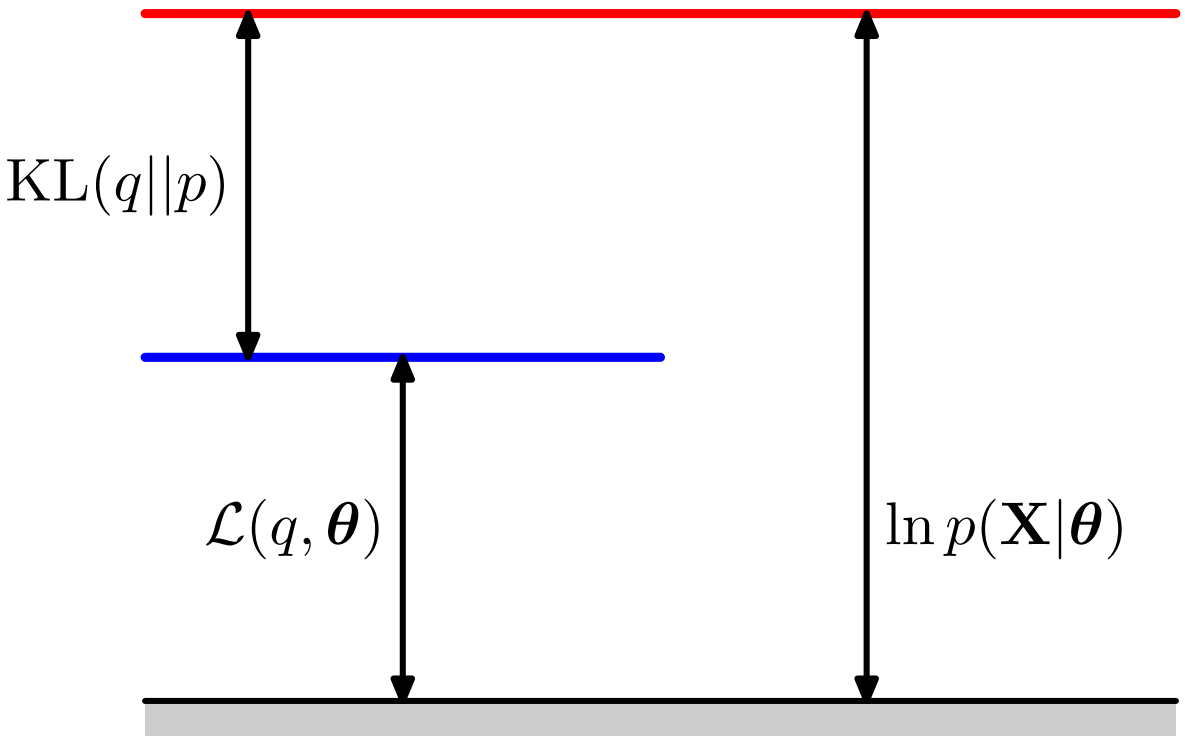
και

$$KL(q||p) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} \tag{22}$$

- $KL(q||p) = 0 \Longleftrightarrow q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$

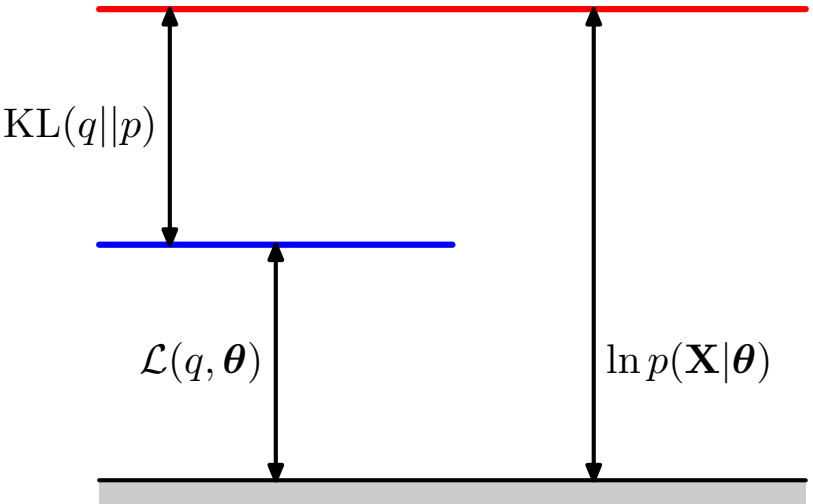
Ο αλγόριθμος ΕΜ

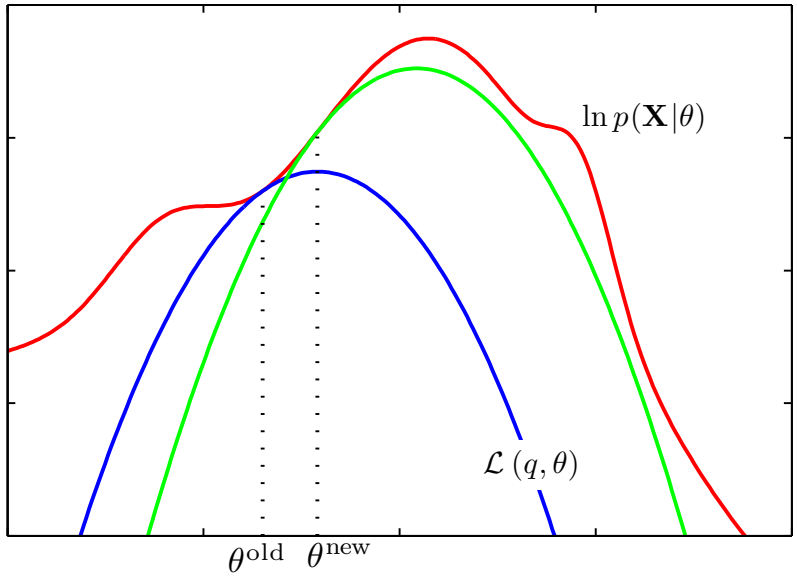
$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p), \tag{23}$$



Ο αλγόριθμος ΕΜ

- E - step:
  - Παράμετροι $\boldsymbol{\theta}_{old}$
  - Μεγιστοποίηση $\mathcal{L}(q, \boldsymbol{\theta}_{old})$ ως προς $q(\boldsymbol{Z})$, $\boldsymbol{\theta}_{old}$ σταθερό
  - max $\mathcal{L}(q, \boldsymbol{\theta}_{old}) \rightarrow KL(q||p) = 0 \rightarrow q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}_{old})$
- M - step:
  - $q(\boldsymbol{Z})$ σταθερή
  - max $\mathcal{L}(q, \boldsymbol{\theta})$ ως προς $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}_{new}$





Διάρθρωση

- Εισαγωγή
- Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)
- Ενισχυτική Μάθηση
- Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης
- Αλγόριθμοι Πολιτικής Κλίσης
- Ο αλγόριθμος ΕΜ
- Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού
- Πειράματα
- Συμπεράσματα

Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού

- Monte Carlo Expectation - Maximization για Ενισχυτική Μάθηση
- Ο ορίζοντας της ΜΔΑ είναι τυχαία μεταβλητή T
- Οι ανταμοιβές θεωρούνται πιθανότητες φανταστικών γεγονότων R
- Χρησιμοποιούμε όλες τις υποτροχιές μήκους t της κάθε τροχιάς
- Οι λανθάνουσες μεταβλητές είναι T, t, ξ

Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού

- Για το E-step αποδεικνύεται πως:

$$q^*(\xi, T, t) = p(\xi, T, t|R; \boldsymbol{\theta}_{old}) \tag{24}$$

$$\propto \alpha(T)b(t)p(\xi|t; \boldsymbol{\theta}_{old})p(R|\xi, T) \tag{25}$$

- Μοντέλο μη διαθέσιμο $\longrightarrow$ δειγματοληψία
  - $\alpha(t) = (1-\delta)\delta^t$, $\gamma < \delta < 1$

Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού

- Στο M-step μεγιστοποιούμε και προκύπτει:

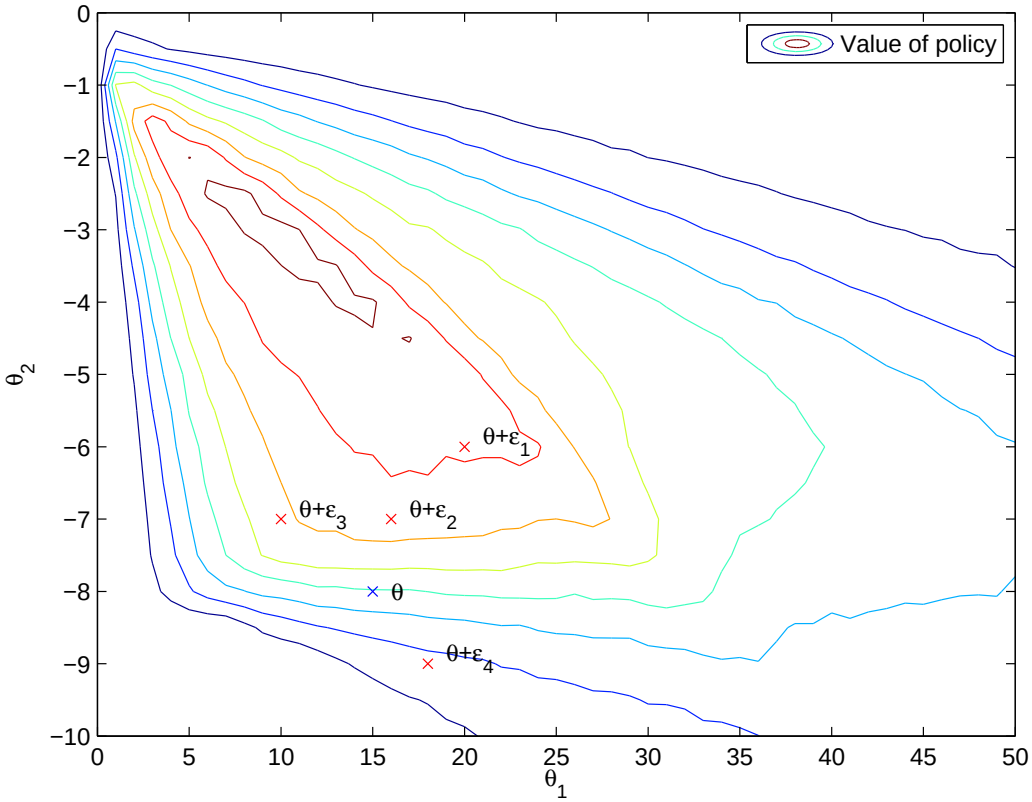
$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \frac{\sum_{i=1}^{m} \frac{1}{|\xi_i|+1} \sum_{t=0}^{|\xi_i|} Q_{it} \boldsymbol{\varepsilon}_{it}}{\sum_{i=1}^{m} \frac{1}{|\xi_i|+1} \sum_{t=0}^{|\xi_i|} Q_{it}} \tag{26}$$

  - $Q_{it} = \sum_{\tau=t}^{|\xi_i|} b(\tau) r_{i\tau}$
  - $b(t) = (1 - \gamma/\delta)(\gamma/\delta)^t$

- Ελεγκτής: $u_t = (\boldsymbol{\theta} + \boldsymbol{\varepsilon}_t)\boldsymbol{\phi}(\boldsymbol{x}_t)$, $\varepsilon_t \sim \mathcal{N}(\varepsilon_t; 0, \sigma^2 I_d)$

Πρόβλημα

- MCEM

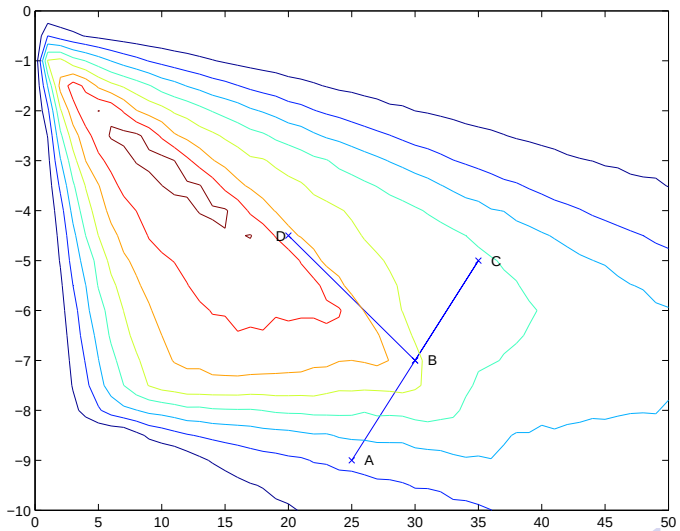


Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού

- Ο αλγόριθμος PoWER
- Σταθερό μήκος H
- Θέτουμε $b(t) = 1/(1+H)$

Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού

- Χρήση των $p\%$ καλύτερων τροχιών
- Αναζήτηση επί γραμμής
  - Κλασική θεώρηση: αναζήτηση στην κατεύθυνση του διανύσματος κλίσης $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \mu\alpha_k \boldsymbol{g}$
  - Θεωρούμε: $\mu\alpha_k \boldsymbol{g} = \boldsymbol{\theta}_{k+1}^{MCEM} - \boldsymbol{\theta}_k^{MCEM}$
  - $n \ll m$ τροχιές σταθερού μήκους H



Διάρθρωση

- Εισαγωγή
- Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)
- Ενισχυτική Μάθηση
- Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης
- Αλγόριθμοι Πολιτικής Κλίσης
- Ο αλγόριθμος ΕΜ
- Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού
- Πειράματα
- Συμπεράσματα

Πειράματα

- Διδιάστατο Συνθετικό Πρόβλημα
- x_1=θέση, x_2=ταχύτητα, u=επιτάχυνση
- Δυναμικό Μοντέλο

$$x_2(t+1) = x_2(t) - u(t) + \kappa, \tag{27}$$
$$x_1(t+1) = x_1(t) - 0.1x_2(t+1) + \kappa, \tag{28}$$

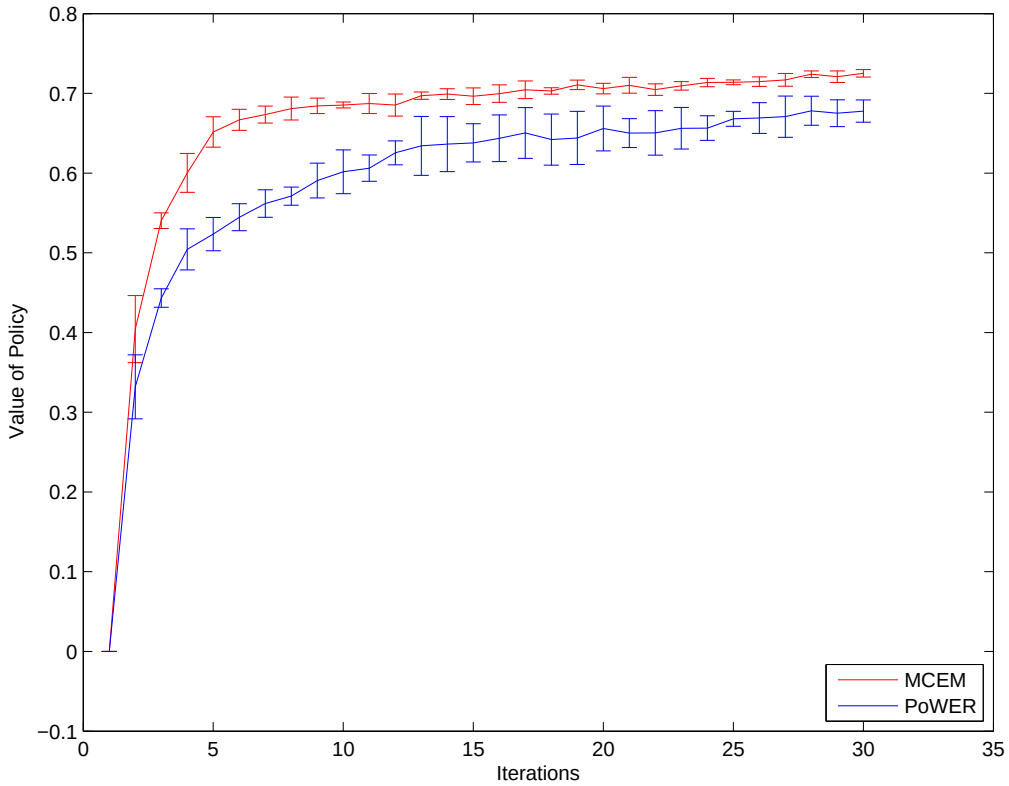
- $\sigma_\kappa = 0.003$
- $u_t = 1/[1 + \exp(-(\boldsymbol{\theta} + \varepsilon_t)^T \boldsymbol{x})] - 0.5$

Πειράματα

- $\boldsymbol{x}_0 = [1, 0] + \textit{noise}$, $x_{goal} = [0, 0]$
- $r(t) = \exp(-0.5||\boldsymbol{x}(t)||_2/(0.03)^2)$
- MCEM
  - $\gamma = 0.95$, $\delta = 0.99$, $\textit{Tmax} = 100$
- PoWER
  - $H = 100$
- Αναζήτηση επί γραμμής
  - 10 τροχιές
- Χρήση καλύτερων τροχιών
  - 2 τροχιές

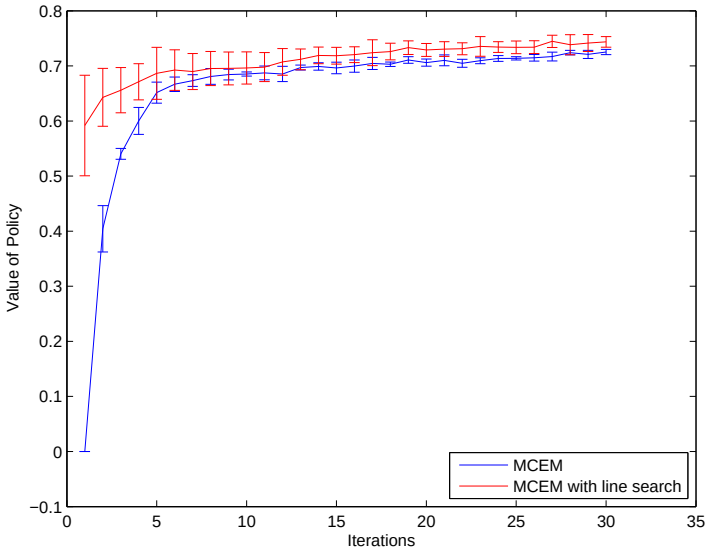
Πειράματα

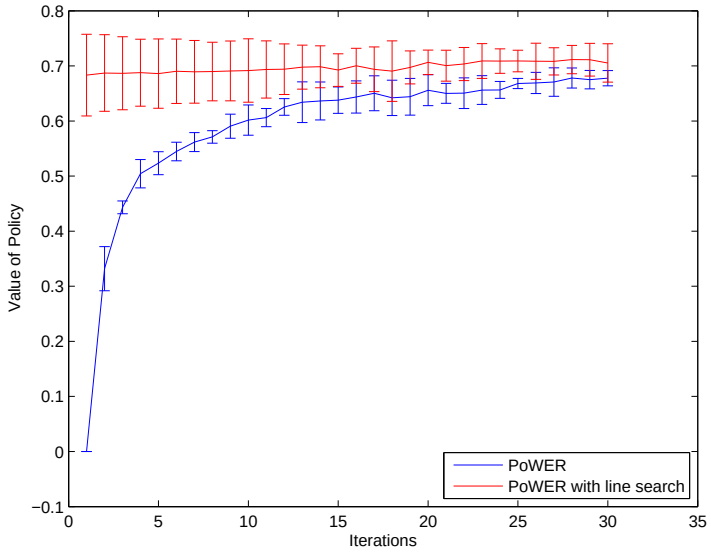
- MCEM vs PoWER



Πειράματα

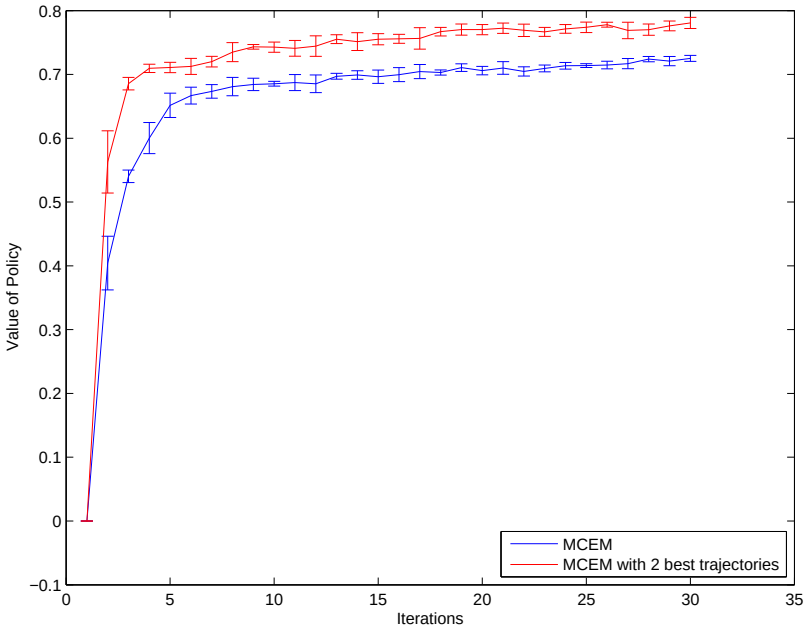
- Με και χωρίς αναζήτηση επί γραμμής

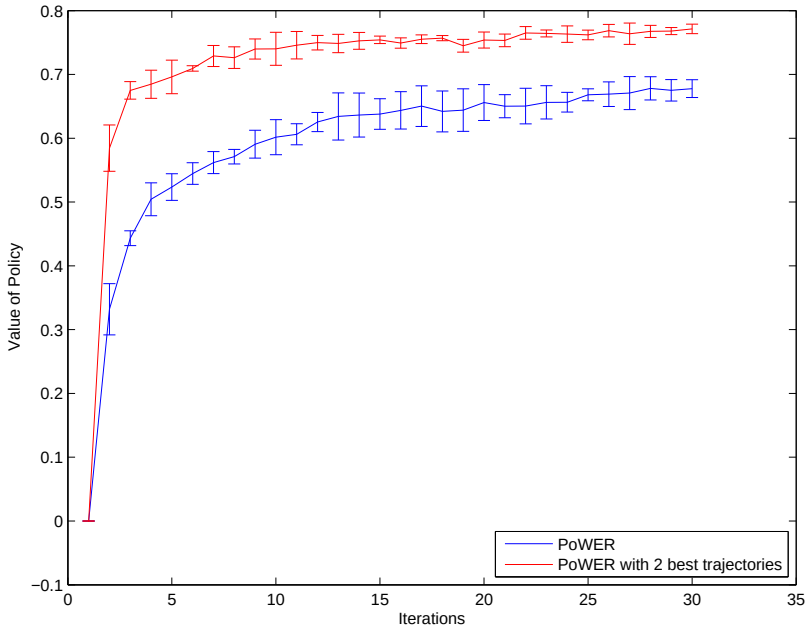




Πειράματα

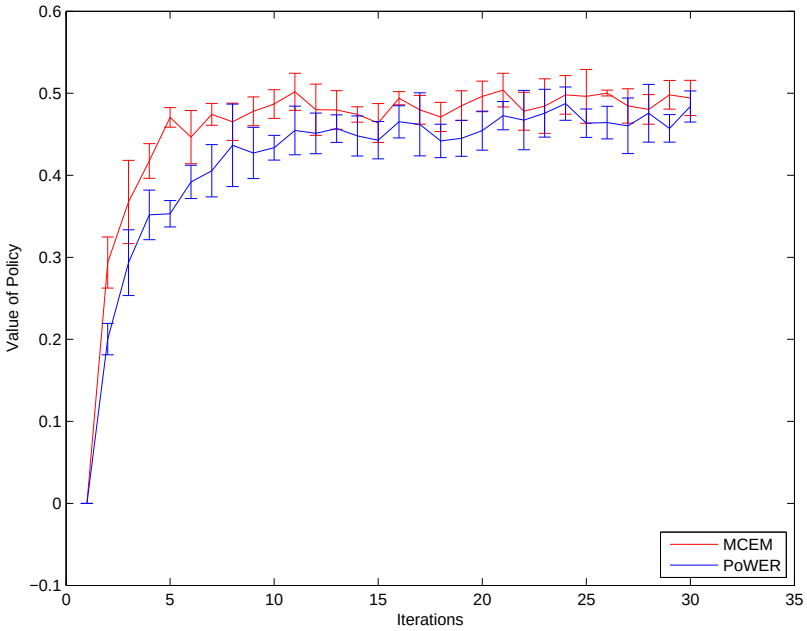
- Με και χωρίς τη χρήση των καλύτερων τροχιών

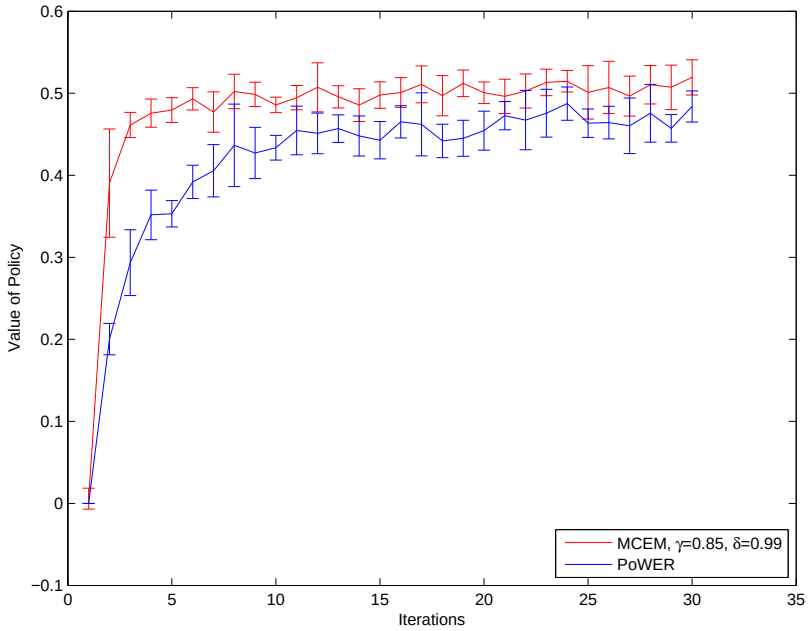




Πειράματα

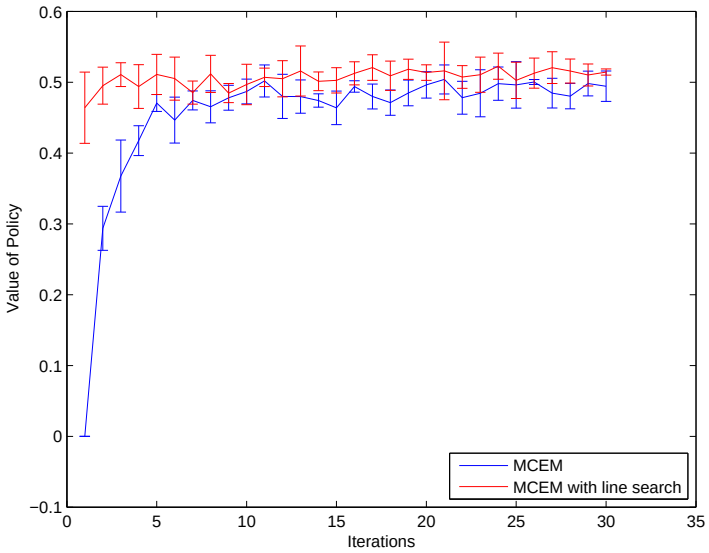
- $\sigma_\kappa = 0.01$
- MCEM vs PoWER

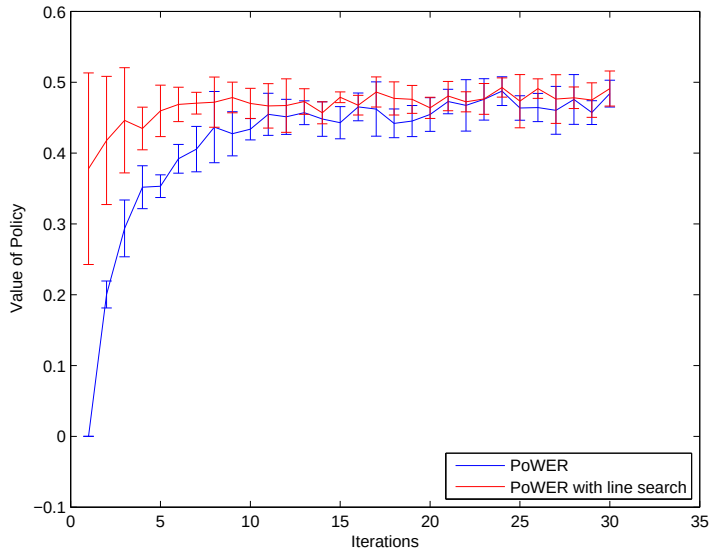




Πειράματα

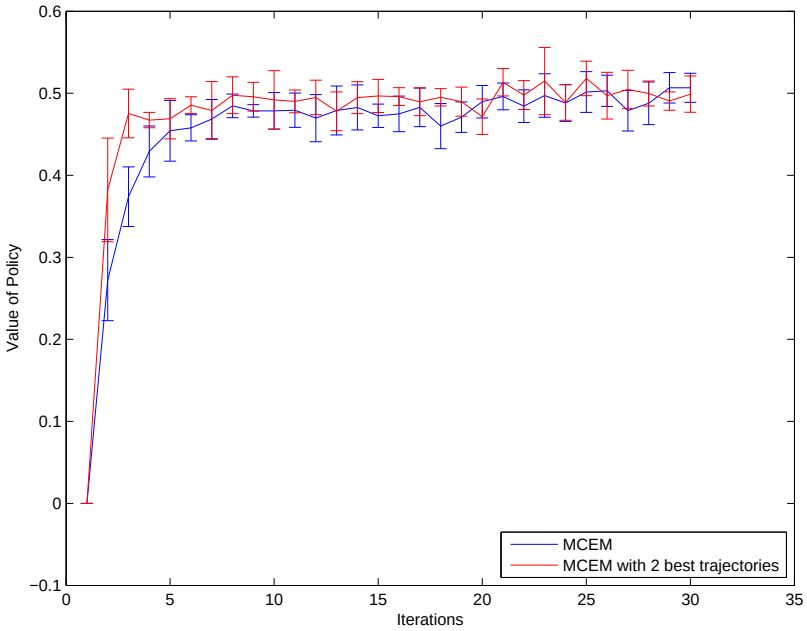
- $\sigma_\kappa = 0.01$
- Με και χωρίς αναζήτηση επί γραμμής

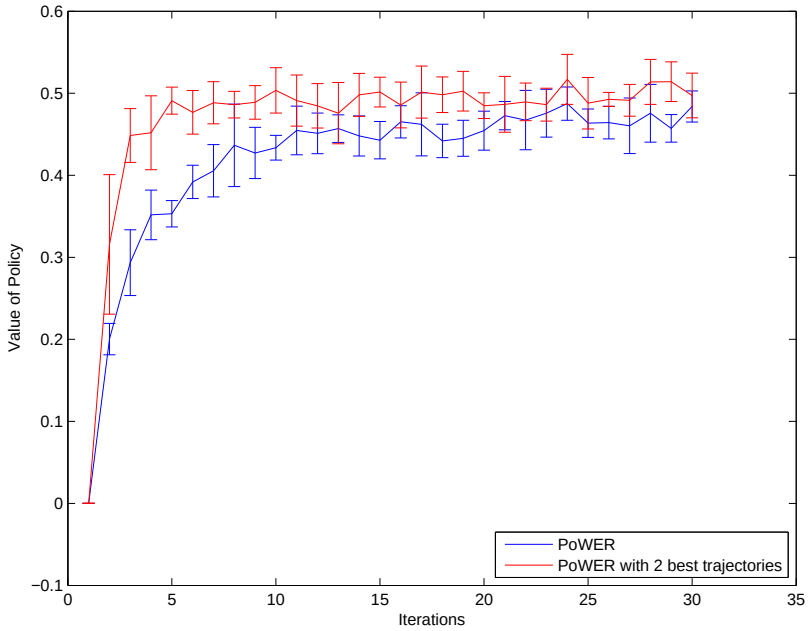




Πειράματα

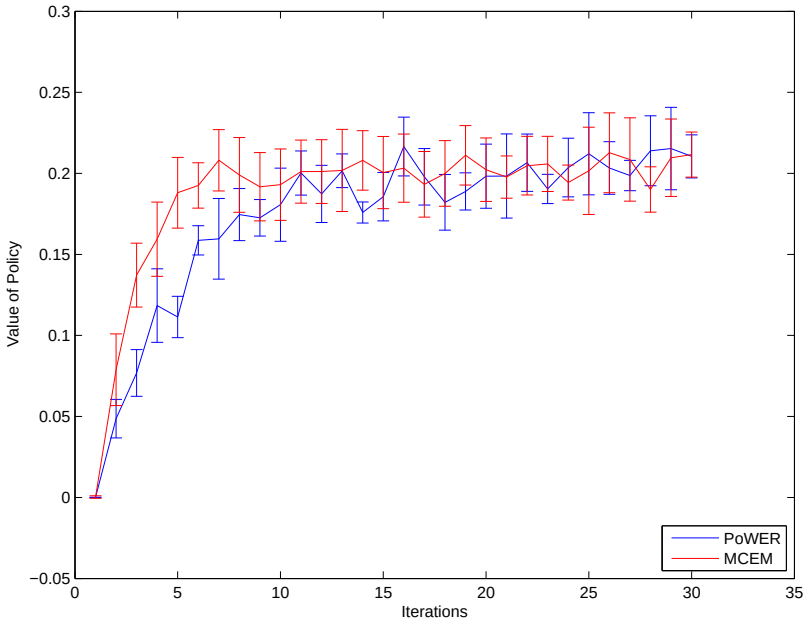
- $\sigma_\kappa = 0.01$
- Με και χωρίς τη χρήση των καλύτερων τροχιών

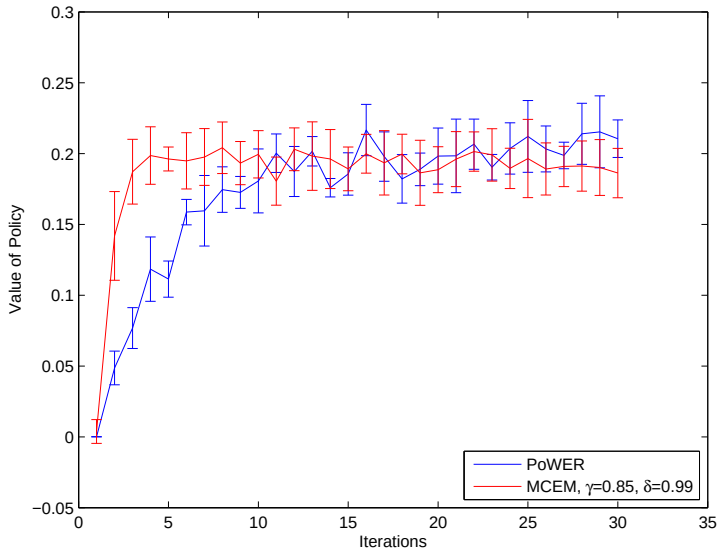




Πειράματα

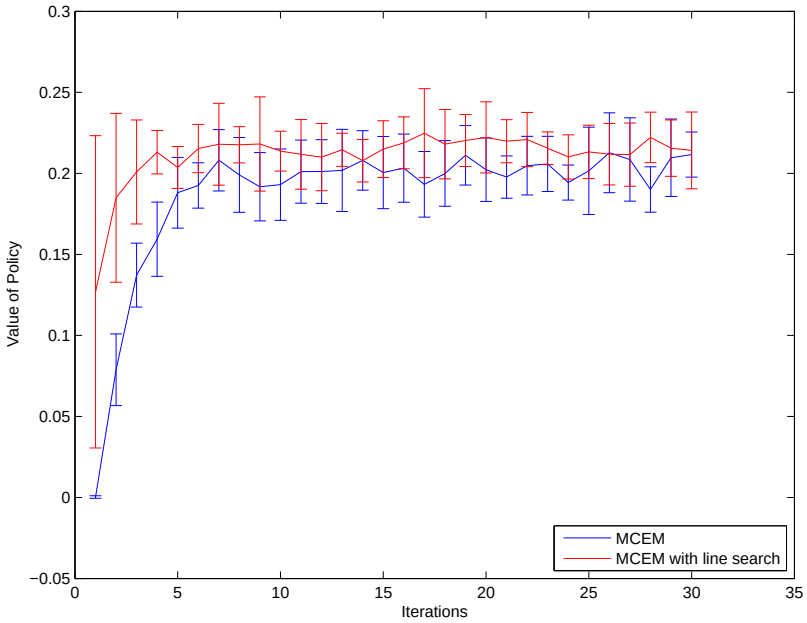
- $\sigma_\kappa = 0.03$
- MCEM vs PoWER

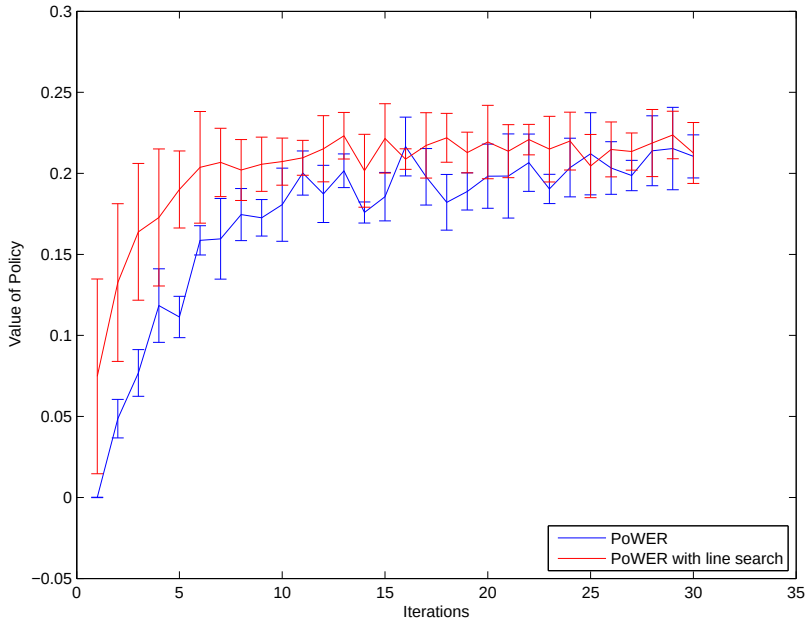




Πειράματα

- $\sigma_\kappa = 0.03$
- Με και χωρίς αναζήτηση επί γραμμής

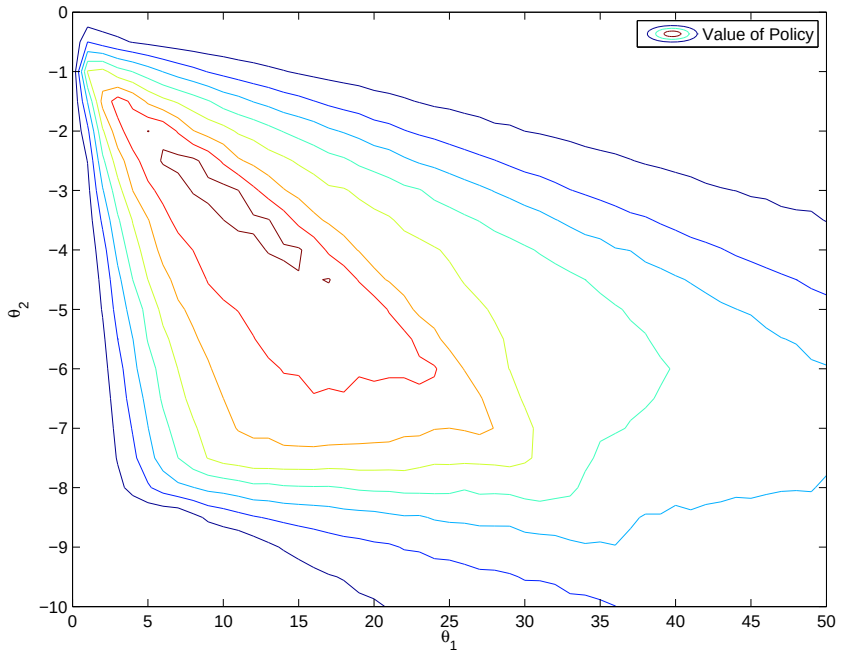




Πειράματα

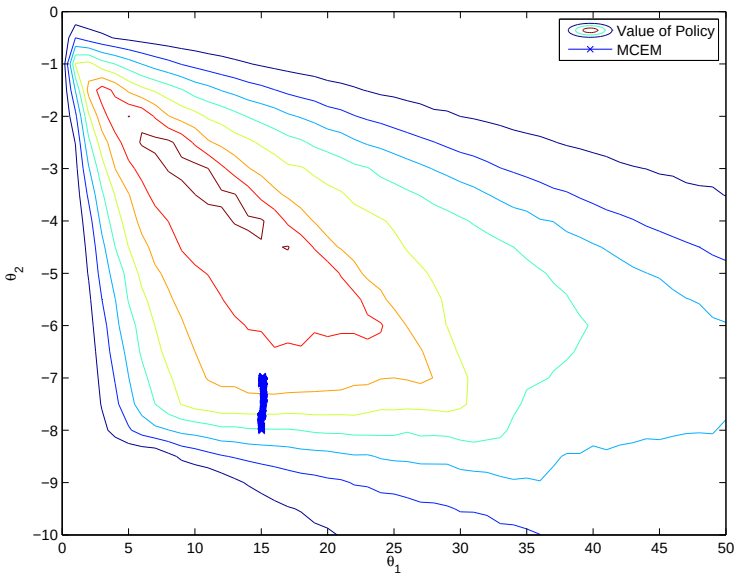
- $\sigma_\kappa = 0.02$
- Συνάρτηση Ανταμοιβής

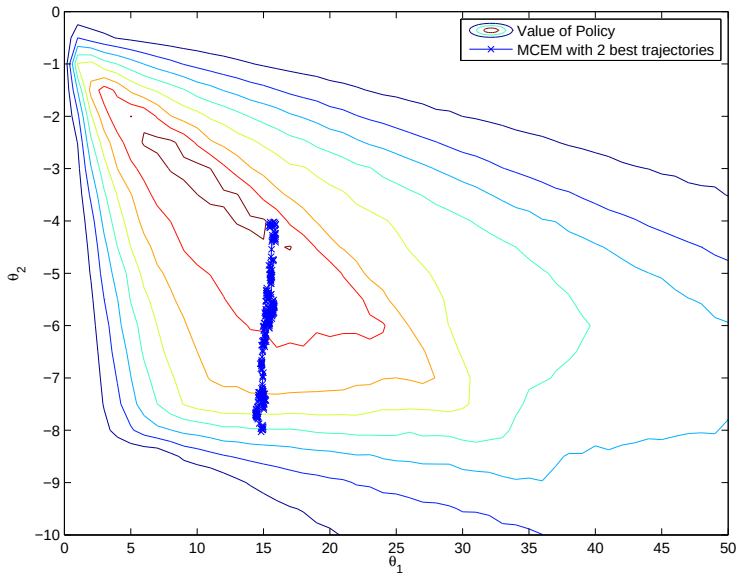
$$r(t) = \begin{cases} 1, & \text{αν } ||\boldsymbol{x}||_2 < 0.1 \\ 0, & \text{αλλού} \end{cases} \tag{29}$$



Πειράματα

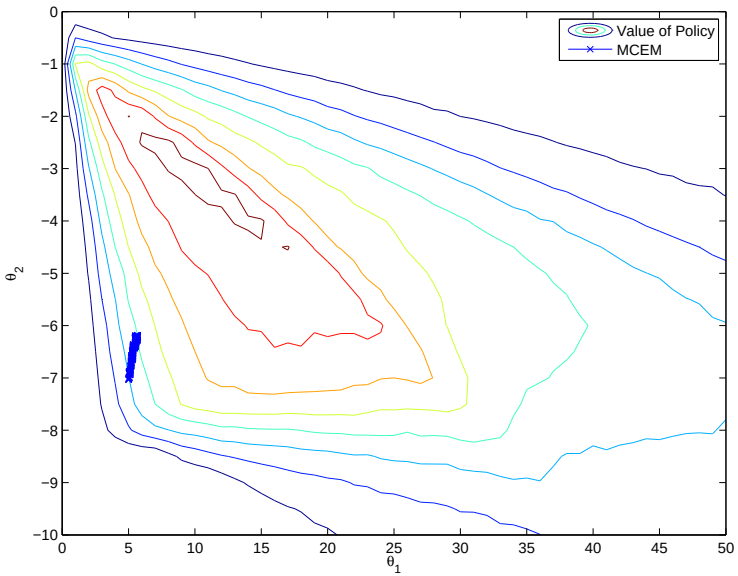
- 300 επαναλήψεις με και χωρίς τη χρήση των 2 καλύτερων τροχιών
- $\boldsymbol{x}_0 = [15, -8]$

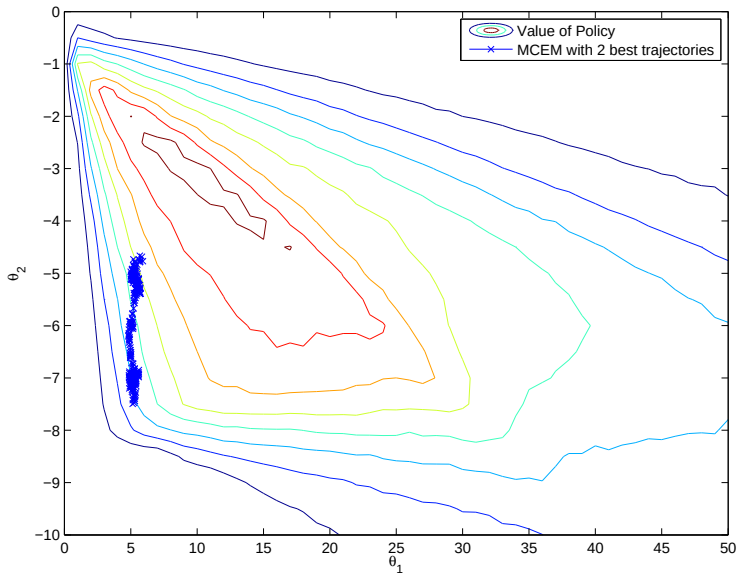




Πειράματα

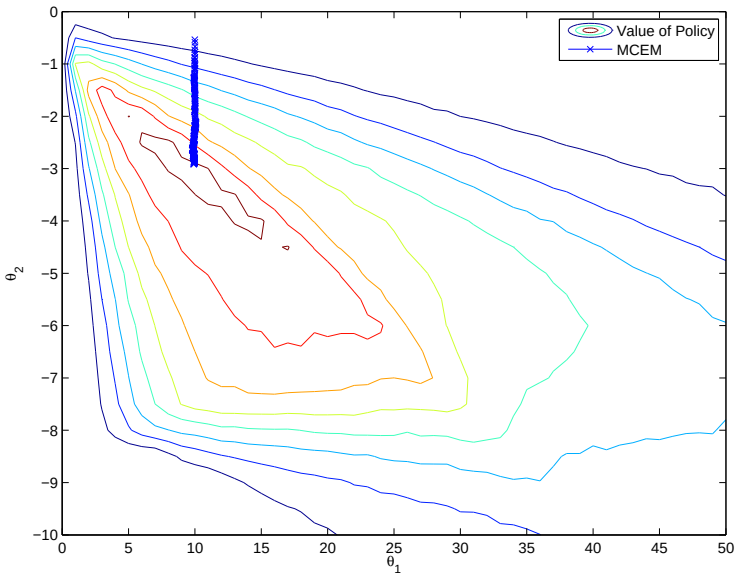
- 300 επαναλήψεις με και χωρίς τη χρήση των 2 καλύτερων τροχιών
- $\boldsymbol{x}_0 = [5, -7]$

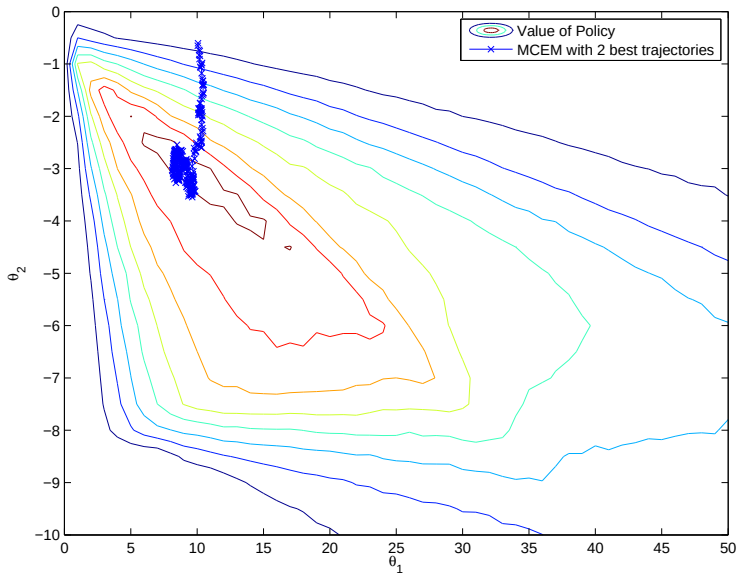




Πειράματα

- 300 επαναλήψεις με και χωρίς τη χρήση των 2 καλύτερων τροχιών
- $\boldsymbol{x}_0 = [10, -0,5]$





Πειράματα

- Εξομοίωση αιώρησης ελικοπτέρου
- Μη γραμμικό, υψηλής διάστασης, στοχαστικό
- Χώρος καταστάσεων
  - Η θέση του ελικοπτέρου (x, y, z)
  - Η διεύθυνσή του (roll ϕ, pitch θ, yaw ω)
  - Η ταχύτητά του σε κάθε διεύθυνση ($\dot{x}$, $\dot{y}$, $\dot{z}$)
- Χώρος ενεργειών
  - u_1: Έλεγχος μπρος - πίσω κυκλικής κλίσης
  - u_2: Έλεγχος αριστερά - δεξιά κυκλικής κλίσης
  - u_3: Έλεγχο κλίσης ρότορα της κύριας έλικας
  - u_4: Έλεγχος κλίσης ρότορα της ουραίας έλικας

Πειράματα

- Διάνυσμα λάθους 9 διαστάσεων

- Ο ελεγκτής

$$u_1 = -w_1 x_{error} - w_2 \dot{x}_{error} - w_3 \theta_{error} + w_4 \quad (30)$$

$$u_2 = -w_5 y_{error} - w_6 \dot{y}_{error} - w_7 \phi_{error} + w_8 \quad (31)$$

$$u_3 = w_9 z_{error} - w_{10} \dot{z}_{error} + w_{11} \quad (32)$$

$$u_4 = -w_{12} \omega_{error} \quad (33)$$

- Η ανταμοιβή

$$r_1(t) = exp(-x_{error}^2) + exp(-\dot{x}_{error}^2) + exp(-\theta_{error}^2) \quad (34)$$

$$r_2(t) = exp(-y_{error}^2) + exp(-\dot{y}_{error}^2) + exp(-\phi_{error}^2) \quad (35)$$

$$r_3(t) = exp(-z_{error}^2) + exp(-\dot{z}_{error}^2) \quad (36)$$

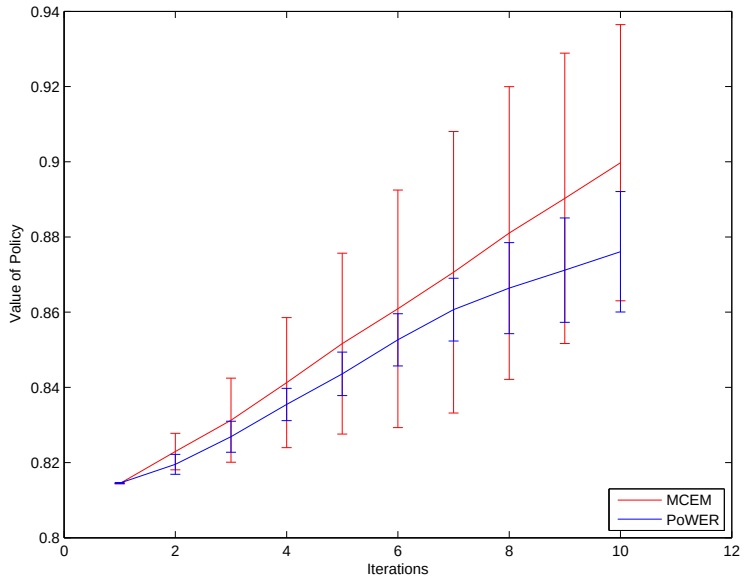
$$r_4(t) = exp(-\omega_{error}^2) \quad (37)$$

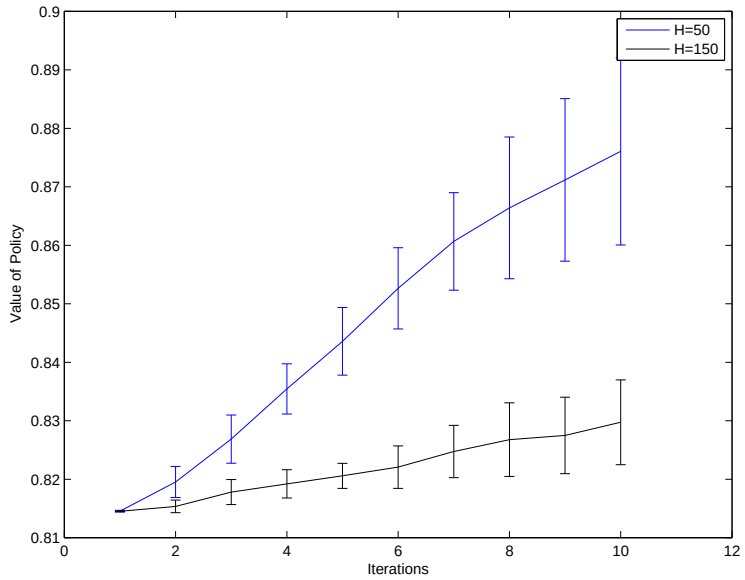
Πειράματα

- 6000 βήματα $\longrightarrow$ 10 λεπτά πραγματικού χρόνου πτήσης
- 100 επεισόδια σε κάθε επανάληψη
- Αναζήτηση επί γραμμής $\longrightarrow$ 10 επεισόδια
- Χρήση καλύτερων τροχιών $\longrightarrow$ 5 καλύτερες
- PoWER
  - $H = 50$
- MCEM
  - $\gamma = 0.95$
  - $\delta = 0.99$
  - $Tmax = 600$

Πειράματα

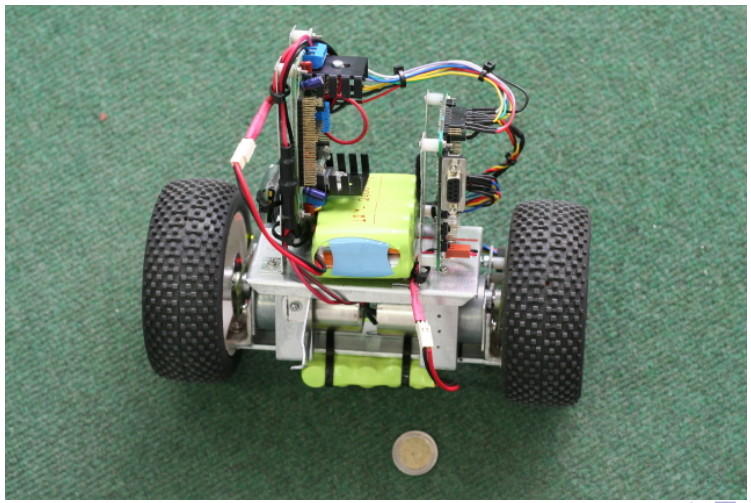
- Ευαισθησία του PoWER στο μήκος επεισοδίου





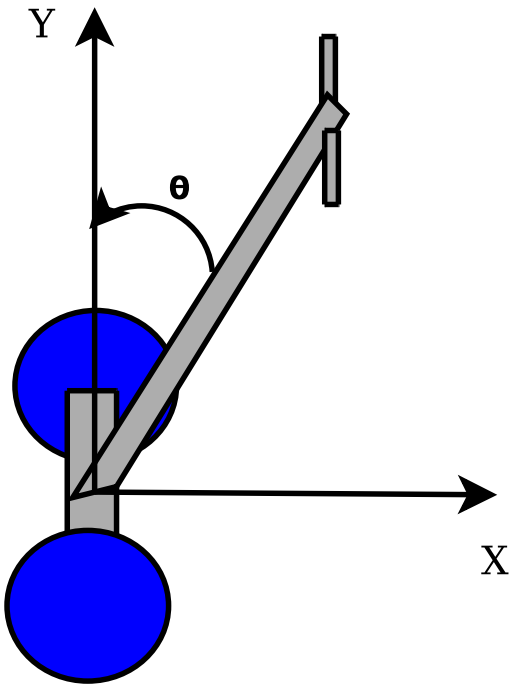
Πειράματα

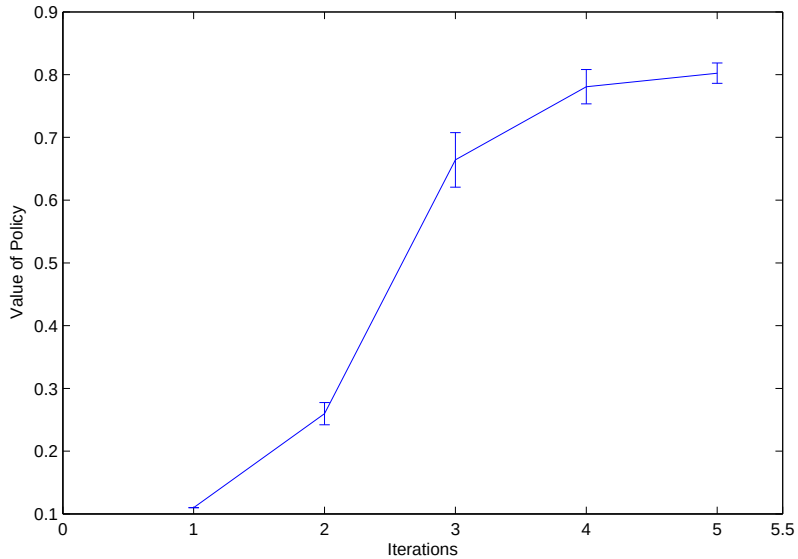
- Δύο 12 Vdc, 152 RpM spur geared κινητήρες
- Ένα ooPIC microcontroller
- Δύο 64 παλμών ανά περιστροφή οδόμετρα
- Ένα επιταχυνσιόμετρο 2 αξόνων και ένα γυροσκόπιο ενός άξονα
- Δύο πηγές ρεύματος



Πειράματα

- Χώρος καταστάσεων: $\boldsymbol{x} = [x_1, x_2, x_3, x_4]$
- Ελεγκτής: $u_t = \boldsymbol{\theta}^T \boldsymbol{x}$
- Ανταμοιβή: $r(t) = exp(-x_1^2(t)) + exp(-x_3^2(t))$





Διάρθρωση

- Εισαγωγή
- Μαρκοβιανές Διεργασίες Απόφασης (ΜΔΑ)
- Ενισχυτική Μάθηση
- Ενισχυτική Μάθηση με Συναρτήσεις Αξιολόγησης
- Αλγόριθμοι Πολιτικής Κλίσης
- Ο αλγόριθμος ΕΜ
- Ενισχυτική Μάθηση μέσω Πιθανοτικού Συμπερασμού
- Πειράματα
- Συμπεράσματα

Συμπεράσματα

- MCEM
  - (+) Απλή υλοποίηση
  - (+) Ελάχιστη χειροκίνητη παραμετροποίηση
  - (+) Εύρωστη συμπεριφορά σε στοχαστικά περιβάλλοντα
  - (+) Χαμηλή υπολογιστική πολυπλοκότητα $O(mT)$
  - (-) Τοπικά βέλτιστα
  - (-) "Τεθλασμένη" πορεία στο χώρο παραμέτρων
  - (-) Μεγάλες απαιτήσεις σε μνήμη

- Εμπειρικοί κανόνες χρήσης των μεθόδων επιτάχυνσης
  - Χρήση πάντα των $p\%$ καλύτερων τροχιών
  - Μικρή στοχαστικότητα $\longrightarrow$ Αναζήτηση επί γραμμής

Συμπεράσματα

1. Nikos Vlassis, Kontes Georgios and Savas Piperidis, “Reinforcement Learning of Robot Control via Probabilistic Inference”, 1st Hellenic Robotics Conference (HEROC), February 23-24, Athens, 2009

2. Nikos Vlassis, Marc Toussaint, Georgios Kontes and Savas Piperidis, "Learning Model - free Robot Control by a Monte Carlo EM Algorithm", Autonomous Robots, Special issue on robot learning (accepted)

Ερωτήσεις

Ερωτήσεις???