Artificial Intelligence

Νίκος Τζιρίτας

These slides are based on the book "Reinforcement Learning, R. S. Sutton and A. G. Barto"

Τεχνητή Νοημοσύνη και Μηχανική Μάθηση

- Η τεχνητή νοημοσύνη είναι μία τεχνολογία που επιτρέπει μία μηχανή να λειτουργεί με τρόπο που προσεγγίζει την ανθρώπινη συμπεριφορά
- Με την τεχνητή νοημοσύνη λύνουμε οποιοδήποτε πρόβλημα λύνονται όπως θα τα έλυνε και ένας άνθρωπος
- Η Τεχνητή νοημοσύνη έχει ευρεία εμβέλεια
- Ο στόχος είναι να αυξηθεί η επιτυχία

- Η μηχανική μάθηση είναι ένα υποσύνολο της τεχνητής νοημοσύνης που στην ουσία μια μηχανή μαθαίνει από δεδομένα χωρίς ρητό προγραμματισμό
- Στη μηχανική μάθηση προσπαθούμε να μάθουμε να εκπαιδεύσουμε μία μηχανή μέσω χρήσης δεδομένων να λύσει ένα συγκεκριμένο πρόβλημα και να έχουμε ένα ακριβές αποτέλεσμα.
- Η μηχανική μάθηση έχει περιορισμένη εμβέλεια.
- Στόχος είναι να αυξηθεί η ακρίβεια.

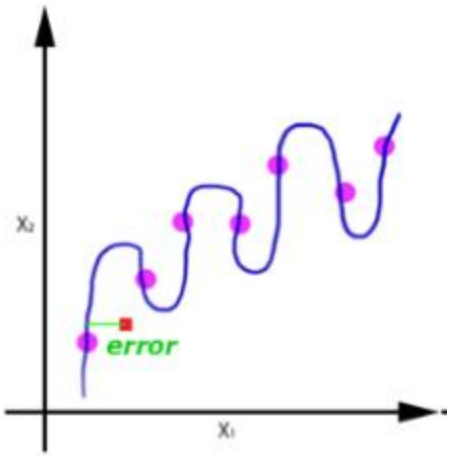
Τεχνητή Νοημοσύνη

- Τρείς τύποι
  - Ασθενής τεχνητή νοημοσύνη (weak AI)
    - Επίλυση προβλημάτων που ανήκουν σε μία περιοχή
  - Γενική τεχνητή νοημοσύνη (general AI)
    - Κατανόηση ή μάθηση οποιασδήποτε πνευματικής εργασίας που μπορεί να κάνει ένας άνθρωπος
  - Ισχυρή τεχνητή νοημοσύνη (strong AI)
    - Στόχος είναι οι υπολογιστές να μην μπορούν να ξεχωρίσουν από το ανθρώπινο μυαλό

Μηχανική Μάθηση

- Τρεις τύποι
  - Εποπτευόμενη μάθηση
    - Μάθηση μέσω δεδομένων με ετικέτα
    - Υπάρχει καθοδήγηση
    - Δεδομένα με ετικέτα
    - Αντιστοιχίζει την είσοδο με ετικέτες σε γνωστή έξοδο
    - Τύποι προβλημάτων όπως regression και classification
  - Μη εποπτευόμενη μάθηση
    - Μάθηση μέσω δεδομένων χωρίς ετικέτα
    - Δεν υπάρχει καθοδήγηση
    - Δεδομένα χωρίς ετικέτα
    - Καταλαβαίνει πρότυπα και ανακαλύπτει το αποτέλεσμα
    - Τύποι προβλημάτων όπως συσχέτιση και συσταδοποίηση
  - Ενισχυμένη μάθηση
    - Ο πράκτορας αλληλεπιδρά με το περιβάλλον πραγματοποιώντας ενέργειες και μαθαίνοντας μέσω trial and error.
    - Δεν υπάρχει καθοδήγηση
    - Δεδομένα που δεν έχουν προσδιοριστεί
    - Ακολουθεί την μέθοδο trial and error
    - Τύποι προβλημάτων που βασίζονται στην ανταμοιβή.

Overfitting και Underfitting



overfitting

optimum

underfitting

Ενισχυμένη Μάθηση

- Οι πράκτορες μαθαίνουμε μέσω της αλληλεπίδρασης με το περιβάλλον
- Πρέπει να είμαστε ενήμεροι πως το περιβάλλον ανταποκρίνεται όταν εκτελούνται ενέργειες.
- Πρέπει να αντιστοιχίζουμε καταστάσεις σε ενέργειες με στόχο να μεγιστοποιήσουμε την ανταμοιβή όταν θα εκτελέσουμε τις ενέργειες.

Εποπτευόμενη μάθηση (1/2)

- Στην εποπτευόμενη μάθηση ένας πράκτορας μπορεί να μαθαίνει από τα παραδείγματα.
  - Αυτά τα παραδείγματα παρέχονται από έναν επόπτη ο οποίος είναι ενήμερος για την αντιστοίχιση από εισόδους σε εξόδους.
  - Κάθε παράδειγμα μπορεί να θεωρηθεί ένα ζευγάρι που αποτελείται από ένα διάνυσμα εισόδου και μία επιθυμητή τιμή εξόδου
  - Ο πράκτορας αναλύει τα δεδομένα εκπαίδευσης και δημιουργεί μία συνάρτηση η οποία θα χρησιμοποιείται για την αντιστοίχιση νέων παραδειγμάτων

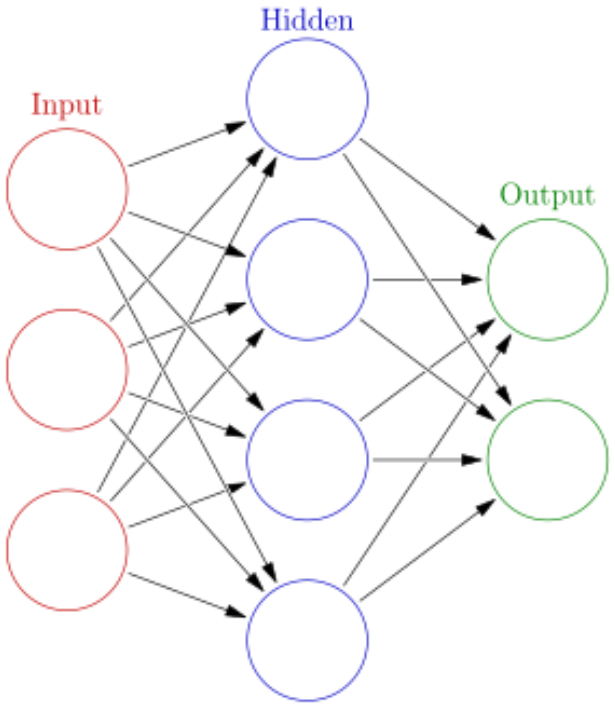
Εποπτευόμενη μάθηση (2/2)

- Σε προβλήματα αλληλεπίδρασης είναι σχεδόν αδύνατον να αποκτήσουμε παραδείγματα επιθυμητής συμπεριφοράς τα οποία είναι ταυτόχρονα σωστά και αναπαραστατικά σε σχέση με όλες τις καταστάσεις που ένας πράκτορας θα αντιδράσει.
- Σε μία αχαρτογράφητη περιοχή, ένας πράκτορα πρέπει να μάθει μέσω των δοκιμών και σφαλμάτων (trial and error). Κανένας επόπτης δεν είναι διαθέσιμος.

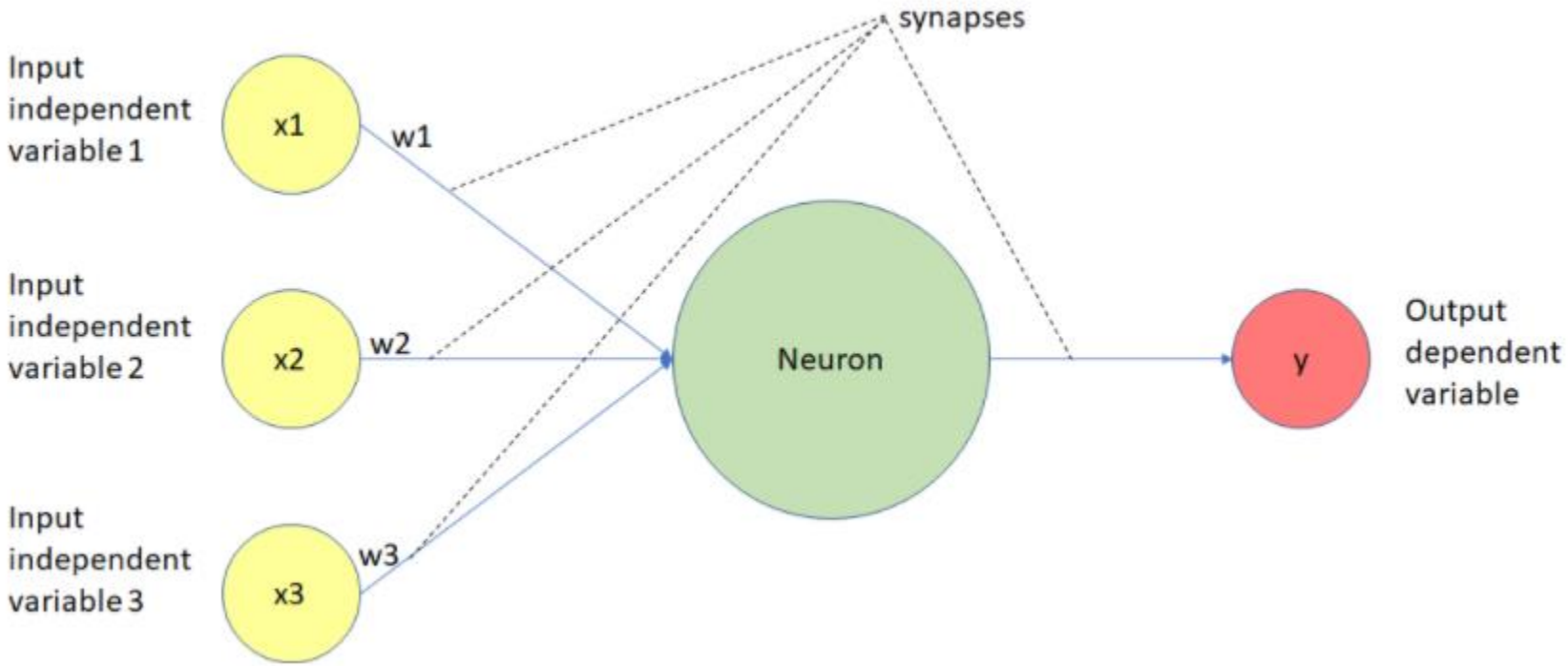
Neural Networks

- Neural networks are composed of a number of layers of nodes
  - Input layer
  - Hidden layers
  - Output layer

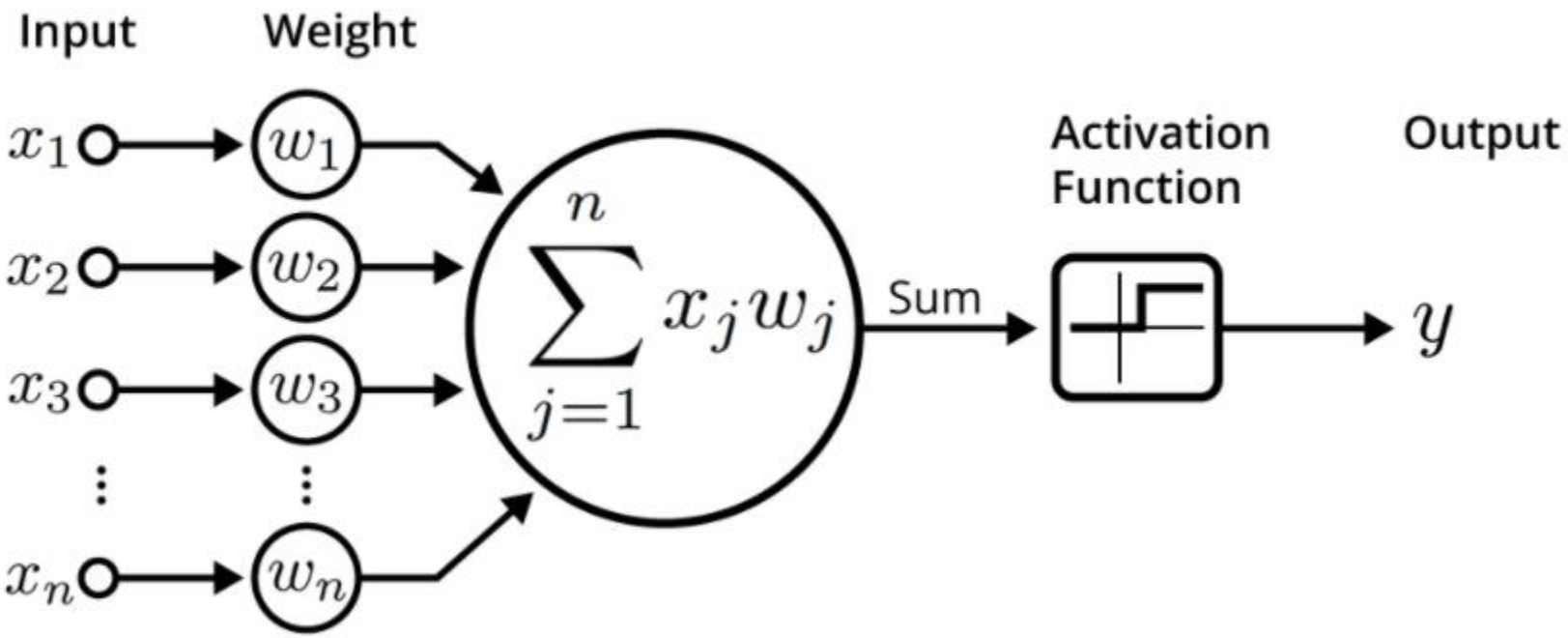
Neural Network Visualization



Neuron (1/2)


synapses
Input independent variable 1
x1
w1
Input independent variable 2
x2
w2
Input independent variable 3
x3
w3
Neuron
y
Output dependent variable

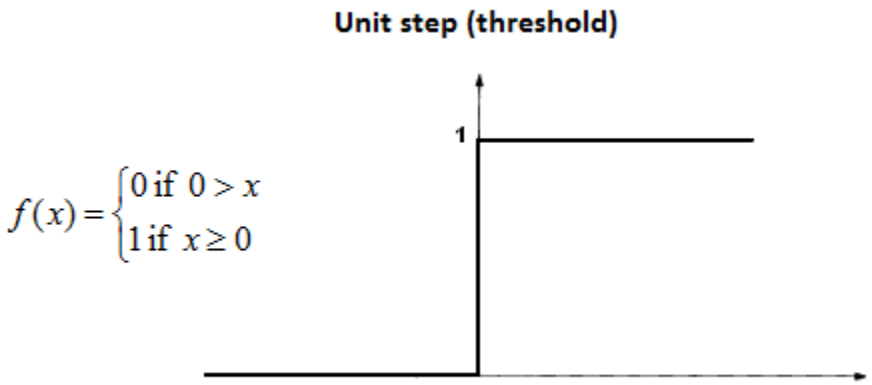
Neuron (2/2)


Input
Weight
x_1
x_2
x_3
x_n
w_1
w_2
w_3
w_n
$\sum_{j=1}^{n} x_j w_j$
Sum
Activation Function
Output
y

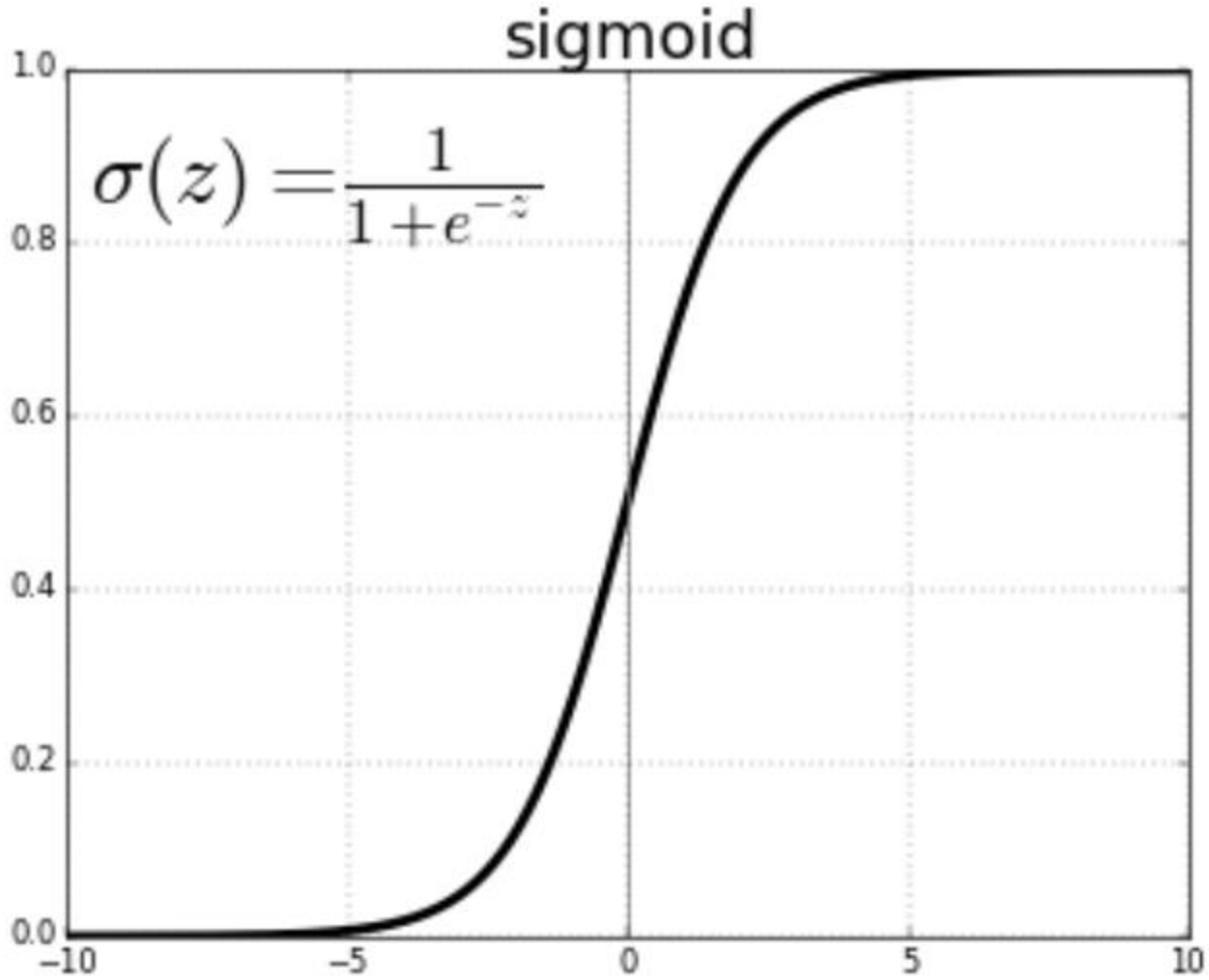
Threshold Functions

These slides are based on the book
"Reinforcement Learning, R. S. Sutton and
A. G. Barto"

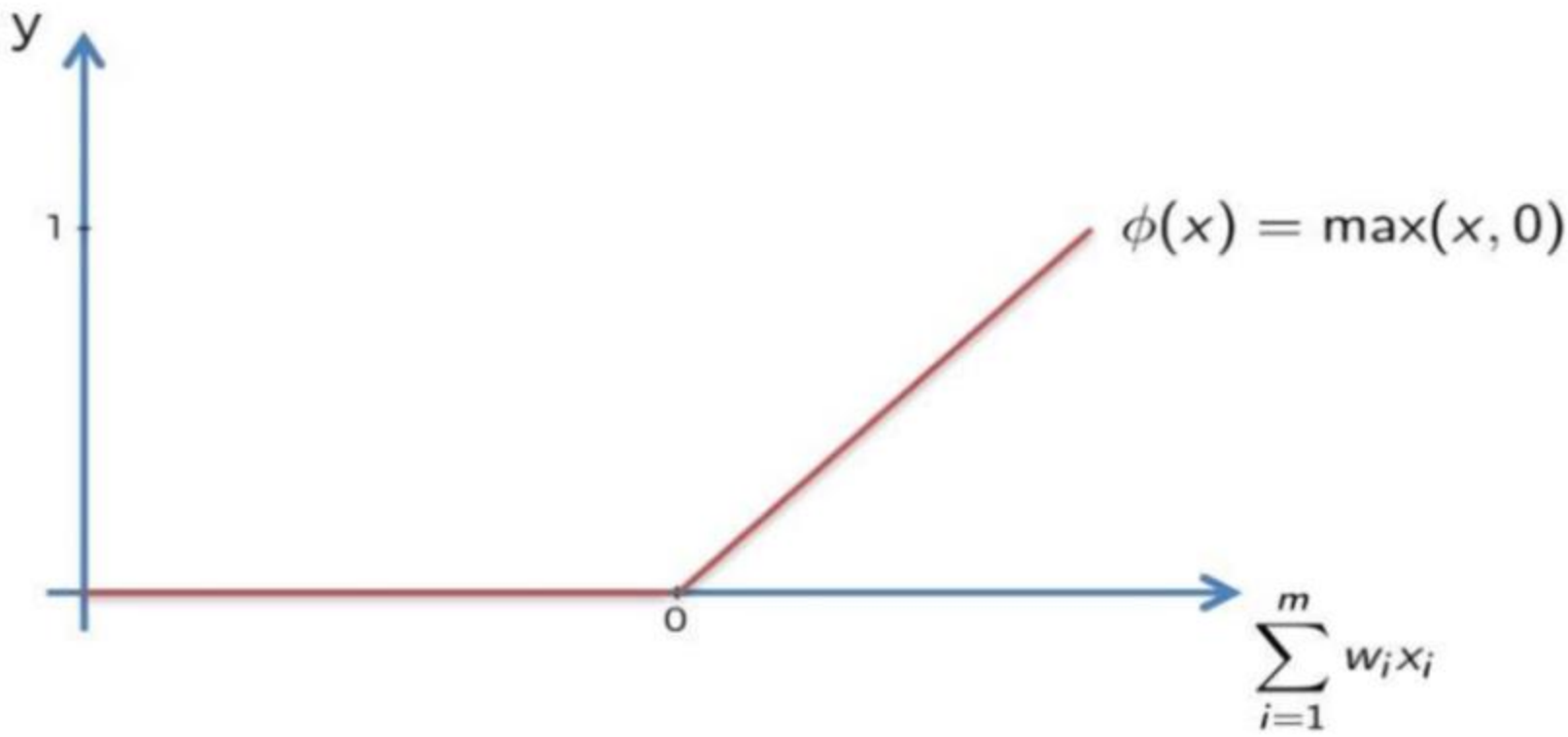
Unit Step Function



Sigmoid Function



Rectifier Functions or Rectifier Linear Unit Activation Functions (ReLu)



Προκλήσεις της Ενισχυμένης Μάθησης

- Εξισορρόπηση μεταξύ εξερεύνησης (exploration) και εκμετάλλευσης (exploitation)
  - Για να αποκτήσουμε μεγάλη ανταμοιβή, ένας πράκτορας ενισχυμένης μάθησης θα πρέπει να προτιμάει ενέργειες που έχει δοκιμάσει στο παρελθόν και του έχουν αποφέρει τη μέγιστη ανταμοιβή.
  - Για να ανακαλύψει ενέργειες με μεγάλη ανταμοιβή, ο πράκτορας πρέπει να δοκιμάσει ενέργειες που δεν έχει επιλέξει προηγουμένως και δεν ξέρει το αποτέλεσμα αυτών των ενεργειών.
  - Σε γενικές γραμμές ο πράκτορας θα πρέπει να εκμεταλλεύεται ότι γνωρίζει για να αποκτήσει μεγάλες ανταμοιβές, αλλά θα πρέπει να εξερευνεί για να βρει καλύτερες ενέργειες στο μέλλον.
  - Σε μία στοχαστική εργασία, ένας πράκτορας πρέπει να επιλέγει μία ενέργεια πολλές φορές για να βρει κάνει μία αξιόπιστη εκτίμηση της αναμενόμενης ανταμοιβής (expected reward)

Παράδειγμα Ενισχυμένης Μάθησης

- Ένα κινούμενο ρομπότ αποφασίζει αν θα πρέπει να μπει σε ένα νέο δωμάτιο για να ψάξει για περισσότερα σκουπίδια ή να βρει τον δρόμο πίσω που βρίσκεται ο σταθμός φόρτισης.
- Παίρνει αποφάσεις με βάση το πόσο γρήγορα και εύκολα στο παρελθόν βρήκε τον σταθμό φόρτισης.

Τα κύρια υπο-στοιχεία ενός συστήματος ενισχυμένης μάθησης

- Πολιτική
- Συνάρτηση ανταμοιβήςReward function
- Συνάρτηση τιμής
- Προαιρετικά ένα μοντέλο του περιβάλλοντος

Πολιτική

- Καθορίζει τον τρόπο που συμπεριφέρεται ένας πράκτορας μάθησης σε μία δοθείσα στιγμή.
- Μία πολιτική αντιστοιχίζει τις καταστάσεις του περιβάλλοντος (όπως τις αντιλαμβάνεται ο πράκτορας) σε ενέργειες που θα πρέπει να ληφθούν όταν ο πράκτορας βρεθεί σε αυτές τις καταστάσεις
- Η πολιτική είναι ο πυρήνας ενός πράκτορα ενισχυμένης μάθησης

Συνάρτηση Ανταμοιβής

- Καθορίζει τον στόχο σε ένα πρόβλημα ενισχυμένης μάθησης
- Αντιστοιχίζει μία κατάσταση (ή ζευγάρι κατάστασης-ενέργειας) του περιβάλλοντος σε έναν αριθμό, την ανταμοιβή, δείχνοντας την έμφυτη επιθυμία του πράκτορα να βρίσκεται στην συγκεκριμένη κατάσταση.
- Ο στόχος του πράκτορα είναι να μεγιστοποιήσει την συνολική ανταμοιβή που λαμβάνει μακροπρόθεσμα.
- Η συνάρτηση ανταμοιβής καθορίζει ποια είναι καλά και ποια κακά γεγονότα για έναν πράκτορα.

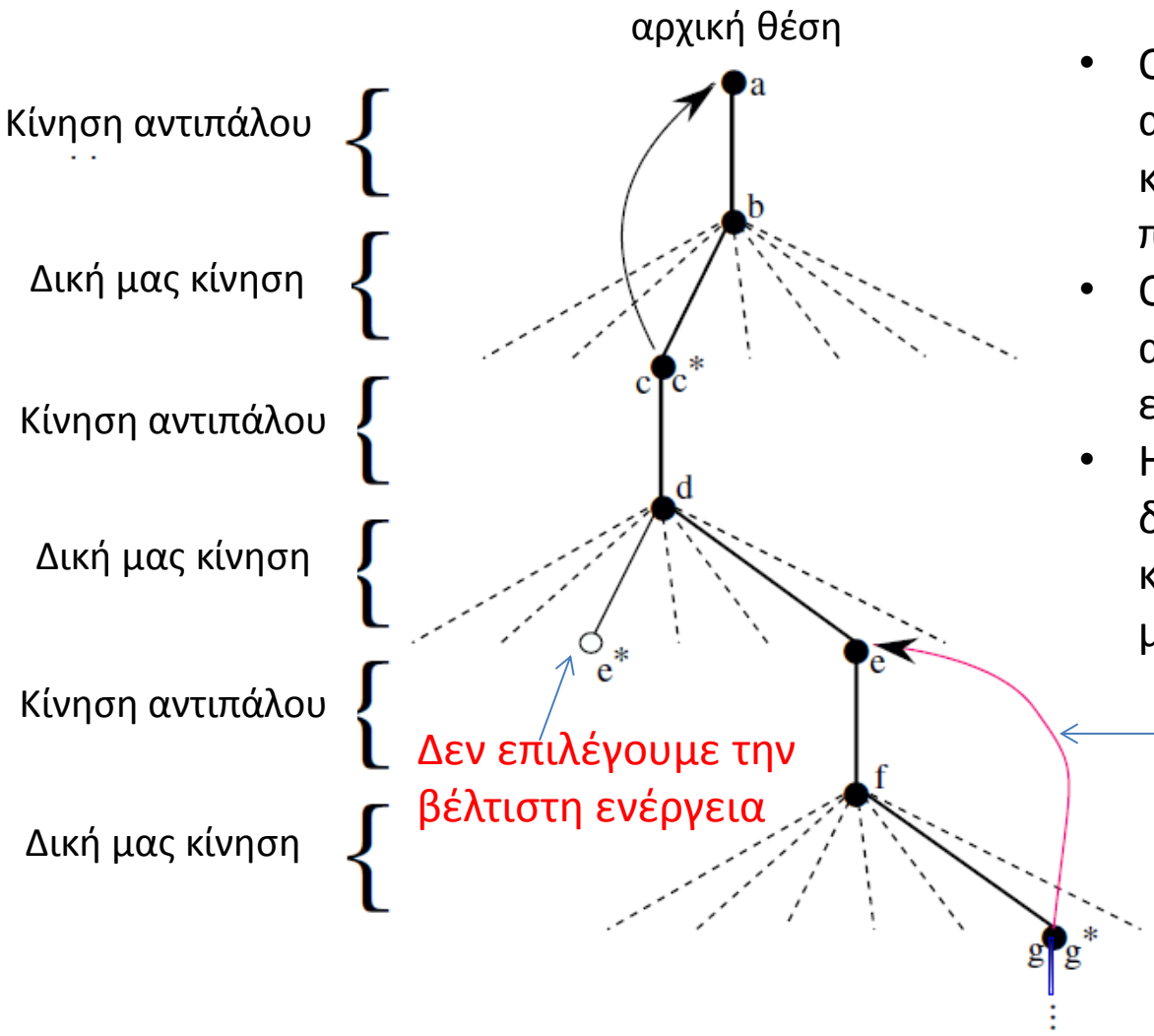
Συνάρτηση Τιμής

- Ενώ μία συνάρτηση ανταμοιβής μας δείχνει τι είναι καλό στην τρέχουσα στιγμή, μία συνάρτηση τιμής καθορίζει τι είναι καλό μακροπρόθεσμα.
- Η τιμή μιας κατάστασης είναι το σύνολο των ανταμοιβών που ένας πράκτορας αναμένει να συσσωρεύσει στο μέλλον, ξεκινώντας από την συγκεκριμένη κατάσταση.
- Για παράδειγμα, μία κατάσταση μπορεί πάντα να αποφέρει πολύ μικρή ανταμοιβή αλλά να έχει μεγάλη τιμή επειδή ακολουθείται από καταστάσεις που αποφέρουν μεγάλες ανταμοιβές.

Μοντέλο Περιβάλλοντος

- Το μοντέλο περιβάλλοντος είναι κάτι που μιμείται την συμπεριφορά του περιβάλλοντος.
  - Για παράδειγμα, δοθείσης μίας κατάστασης και μίας ενέργειας, το μοντέλο θα πρέπει να προβλέψει την επόμενη κατάσταση και την επόμενη ανταμοιβή.
- Τα μοντέλα χρησιμοποιούνται για σχεδιασμό, όπου εννοούμε οποιοδήποτε τρόπο να αποφασίσουμε μία αλληλουχία από ενέργειες εξετάζοντας πιθανές μελλοντικές καταστάσεις, χωρίς αυτές να έχουν ποτέ συμβεί στην πραγματικότητα.

Backup Διάγραμμα



- Οι συνεχόμενες γραμμές αναπαριστούν τις κινήσεις που κάνουμε κατά την διάρκεια του παιχνιδιού
- Οι διακεκομμένες γραμμές αναπαριστούν κινήσεις που εξετάσαμε αλλά δεν επιλέξαμε
- Η δεύτερη κίνηση είναι διερευνητική (οι διερευνητικές κινήσεις δεν καταλήγουν σε μάθηση)

Backups

$$V(s) \leftarrow V(s) + \alpha \left[V(s') - V(s) \right]$$

- Κάνουμε “back up” την τιμή μιας κατάστασης μετά από μία άπληστη κίνηση
- Συγκεκριμένα, με το backup προσαρμόζουμε την τρέχουσα τιμή της προηγούμενης κατάστασης να είναι πιο κοντά στην τιμή της πρόσφατης κατάστασης.
- Το s υποδηλώνει την κατάσταση πριν την άπληστη κίνηση, ενώ s’ υποδηλώνει την κατάσταση μετά την κίνηση.
- Το α είναι μία πολύ μικρή τιμή και ονομάζεται βηματική παράμετρος (step-size parameter), η οποία επηρεάζει τον ρυθμό μάθησης.

n-Armed Bandit Problem (1/2)

- Είμαστε αντιμέτωποι συνεχώς με την επιλογή n διαφορετικών επιλογών/ενεργειών.
- Μετά από κάθε επιλογή λαμβάνουμε μία ανταμοιβή που επιλέγεται από μία σταθερή κατανομή πιθανότητας.
- Ο στόχος είναι να μεγιστοποιήσουμε τη συνολική αναμενόμενη ανταμοιβή για κάποιο χρονικό διάστημα.

n-Armed Bandit Problem (2/2)

- Στο *n*-armed bandit πρόβλημα, κάθε ενέργεια έχει μία αναμενόμενη ή μέση ανταμοιβή δοθείσης της ενέργειας που επιλέξαμε (τιμή από την ενέργεια).
- Αν γνωρίζουμε την τιμή από κάθε ενέργεια τότε το πρόβλημα λύνεται εύκολα.
- Διατηρούμε εκτιμήσεις από τις τιμές των ενεργειών.
- Όταν επιλέγουμε μία ενέργεια με την μέγιστη τιμή (άπληστη) τότε έχουμε εκμετάλλευση (**exploitation**)
- Όταν επιλέγουμε μία μη άπληστη λύση ενέργεια τότε έχουμε εξερεύνηση (**exploration**)

Εξερεύνηση vs Εκμετάλλευση

- Εκμετάλλευση είναι αυτό που πρέπει να κάνουμε για να μεγιστοποιήσουμε την αναμενόμενη ανταμοιβή για ένα play
- Η εξερεύνηση μπορεί να παράγει μία μεγαλύτερη ανταμοιβή μακροπρόθεσμα

Μέθοδοι ενέργειας-τιμής Action-Value Methods (1/2)

- Μέθοδοι που εκτιμούν τις τιμές των ενεργειών
- Οι εκτιμήσεις χρησιμοποιούνται για να επιλέξουμε μία ενέργεια
- Η αληθινή (πραγματική τιμή) μιας ενέργεια δηλώνεται με το $Q^*(\alpha)$
  - Η αληθινή τιμή μιας ενέργειας είναι η μέση ανταμοιβή που λαμβάνεται όταν η συγκεκριμένη ενέργεια επιλέγεται
- Η εκτιμώμενη τιμή από μια ενέργεια στο t-οστο παιχνίδι (play) δηλώνεται με $Q_t(\alpha)$

Action-Value Methods (2/2)

- Πως κάποιος μπορεί να εκτιμήσει τη μέση ανταμοιβή μιας ενέργειας?
  - Μέση ανταμοιβή.

$$Q_t(a) = \frac{r_1 + r_2 + \cdots + r_{k_a}}{k_a}.$$

- Αν $k_\alpha = 0$, τότε επιλέγουμε $Q_0(\alpha)=0$
- Όταν $k_a \to \infty$, τότε ισχύει ότι $Q_t(\alpha)$ συγκλίνει στο $Q^*(\alpha)$.

Άπληστη Ενέργεια

- Επιλέγουμε την βέλτιστη ενέργεια, δηλαδή την ενέργεια με την μεγαλύτερη εκτιμώμενη τιμή

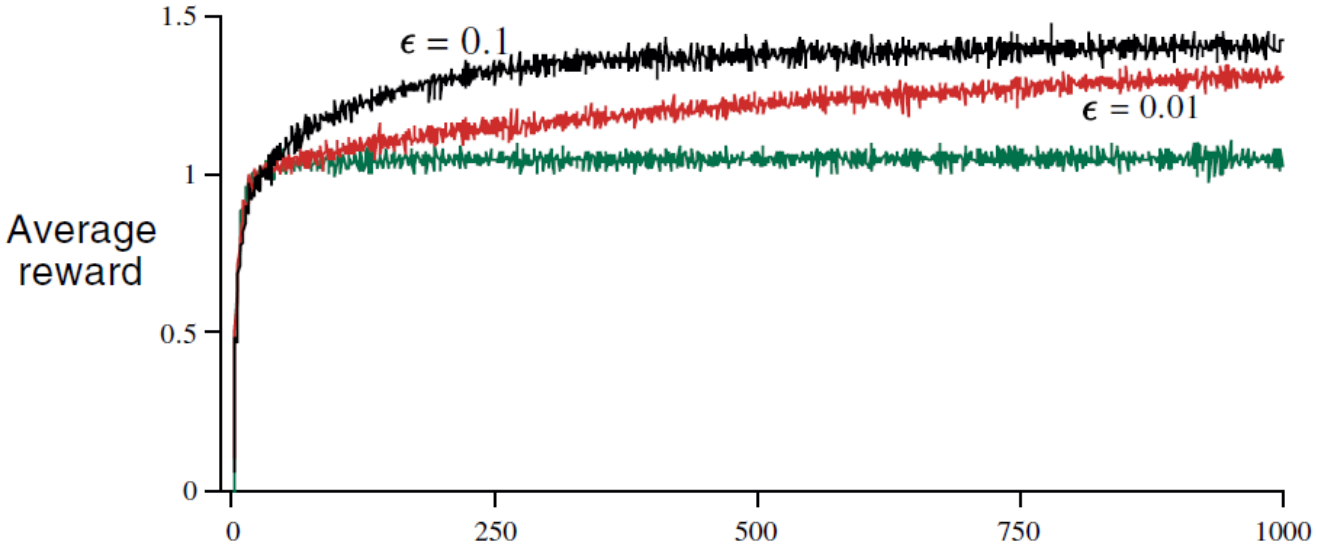
$$Q_t(a^*) = \max_a Q_t(a).$$

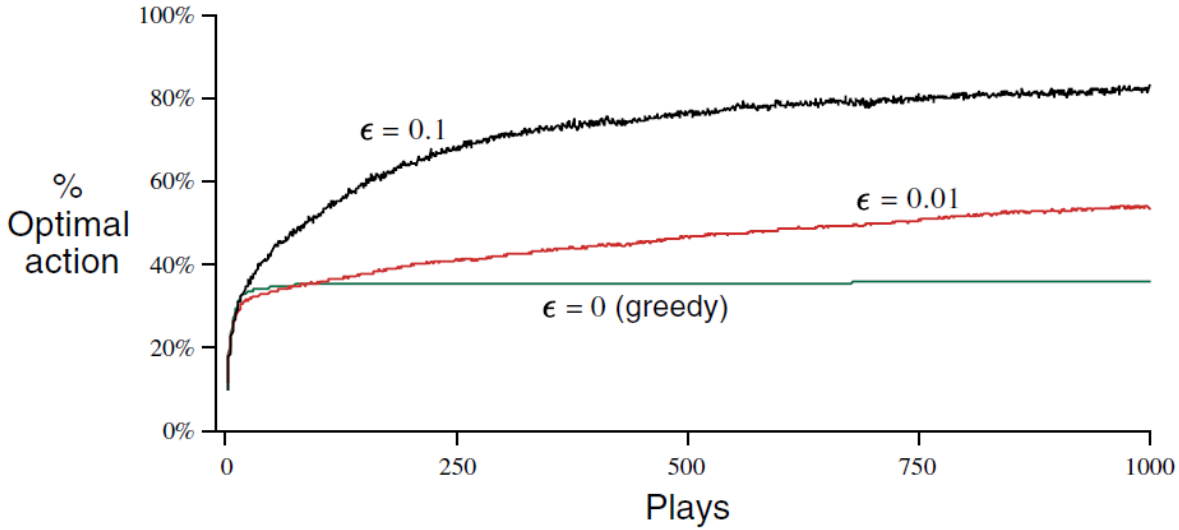
- Αυτή η μέθοδος εκμεταλλεύεται την τρέχουσα γνώση για να μεγιστοποιήσει την άμεση ανταμοιβή
- Δεν ξοδεύει καθόλου χρόνο για δειγματοληψία, καταλήγοντας σε μη βέλτιστες λύσεις για να δει αν μακροπρόθεσμα θα βοηθήσουν για καλύτερες άπληστες ενέργειες

ε-greedy

- Συμπεριφέρεται με άπληστο τρόπο τον περισσότερο χρόνο
- Σε κάποιες χρονικές στιγμές με μικρή πιθανότητα ε, επιλέγουμε μία ενέργεια τυχαία, ομοιόμορφα, ανεξάρτητα από τις εκτιμήσεις των τιμών-ενεργειών
- Το πλεονέκτημα αυτής της μεθόδου είναι ότι όσο ο αριθμός των παιχνιδιών αυξάνεται, για κάθε ενέργεια γίνεται δειγματοληψία άπειρες φορές
  - Θα ισχύει ότι $\mathrm{k}_a \to \infty$ για όλα τα α
  - $Q_t(\alpha)$ συγκλίνει στο $Q^*(\alpha)$

Experiments





Softmax Action Selection (1/2)

- Το μειονέκτημα της ε-greedy μεθόδου είναι ότι εξερευνεί με ισότιμο τρόπο όλες τις ενέργειες
  - Επομένως τέτοιες μέθοδοι επιλέγουν με τον ίδιο τρόπο ενέργειες που είναι πολύ καλές ή πολύ κακές
- Οι Softmax μέθοδοι ποικίλουν τις πιθανότητες ενεργειών σαν μία βαθμονομημένη συνάρτηση εκτιμώμενης τιμής
- Η άπληστη ενέργεια έχει την μεγαλύτερη πιθανότητα να επιλεχθεί.
- Οι υπόλοιπες ενέργειες ταξινομούνται με βάση το βάρος των εκτιμήσεων των τιμών.

Softmax Action Selection (2/2)

- Με την Softmax μέθοδο μία ενέργεια α επιλέγεται με πιθανότητα:

$$\frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}}$$

- Η r είναι μία θετική παράμετρος που ονομάζεται θερμοκρασία
  - Μεγάλες θερμοκρασίες οδηγούν σε σχεδόν ισοπίθανες ενέργειες
  - Χαμηλές θερμοκρασίες οδηγούν σε μεγαλύτερη διαφορά σχετικά με την πιθανότητα για κάθε ενέργεια όσον αφορά την εκτίμηση τιμών

Τμηματική υλοποίηση

$$
\begin{aligned}
Q_{k+1} &= \frac{1}{k+1} \sum_{i=1}^{k+1} r_i \\
&= \frac{1}{k+1} \left(r_{k+1} + \sum_{i=1}^{k} r_i \right) \\
&= \frac{1}{k+1} \left(r_{k+1} + k Q_k + Q_k - Q_k \right) \\
&= \frac{1}{k+1} \left(r_{k+1} + (k+1) Q_k - Q_k \right) \\
&= Q_k + \frac{1}{k+1} \left[r_{k+1} - Q_k \right],
\end{aligned}
$$

Κανόνας Ενημέρωσης (Update Rule)

$$NewEstimate \leftarrow OldEstimate + StepSize\left[Target - OldEstimate\right]$$

Η έκφραση [Target-OldEstimate] είναι το σφάλμα στην εκτίμηση

- Ο στόχος είναι να κάνουμε ένα βήμα προς τον στόχο.
- Η βηματική παράμετρος που χρησιμοποιείται στην τμηματική μέθοδο αλλάζει από βήμα σε βήμα