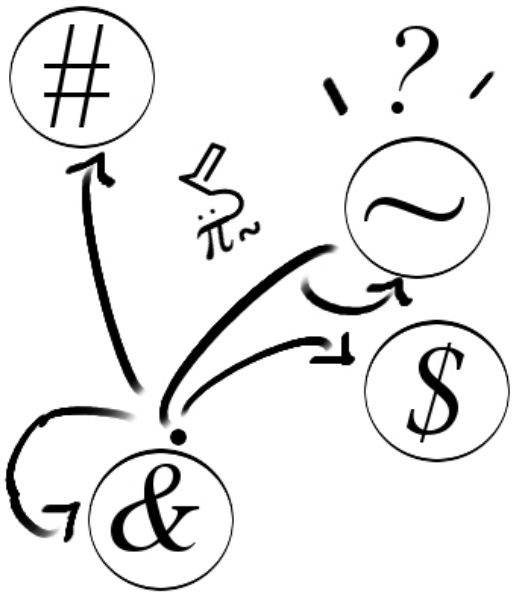
Βελτιστοποίηση Ελεγκτών MDP με τη χρήση τροχιών Μέγιστης Πιθανότητας



Παύλος Ανδρεάδης

Βελτιστοποίηση Ελεγκτών MDP με τη χρήση τροχιών Μέγιστης Πιθανότητας

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

Παύλος Ανδρεάδης

A.M: 2008019031, Email: andreadis.paul@yahoo.com

Εξεταστική Επιτροπή:

Ν. Βλάσσης, Επίκουρος Καθηγητής, Τμήμα Μηχανικών Παραγωγής και Διοίκησης, Πολυτεχνείο Κρήτης (Επιβλέπων)

Ι. Νικολός, Επίκουρος Καθηγητής, Τμήμα Μηχανικών Παραγωγής και Διοίκησης, Πολυτεχνείο Κρήτης

Μ. Λαγουδάκης, Επίκουρος Καθηγητής, Τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών, Πολυτεχνείο Κρήτης

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ
ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Εικόνα εξωφύλλου: *'KaPiΕλάς'*,

Βελτιστοποίηση Ελεγκτών MDP με τη χρήση τροχιών Μέγιστης Πιθανότητας

Περίληψη

Σκοπός της εργασίας είναι η υλοποίηση ενός νέου αλγορίθμου ενισχυτικής μάθησης (Reinforcement Learning, RL) με τη χρήση τροχιών μέγιστης πιθανότητας. Έχει δειχθεί ότι ένα πρόβλημα Βέλτιστου Ελέγχου Διακριτού Χρόνου (Discrete time optimal Control) μπορεί να αναχθεί σε ένα πρόβλημα Πιθανοτικού Συμπερασμού (Probabilistic Inference) και να λυθεί με αντίστοιχες τεχνικές. Πρόσφατες εργασίες έχουν επιλύσει το πρόβλημα χρησιμοποιώντας αλγορίθμους Προσδοκίας-Μεγιστοποίησης (Expectation - Maximization, EM) σε μοντέλα μικτής κατανομής πιθανότητας (probabilistic mixture model). Στηριζόμενοι στην ανάλυση αυτή, επιλύουμε το πρόβλημα Βέλτιστου Ελέγχου Διακριτού Χρόνου με τη χρήση αλγορίθμου Διάδοσης Γνώμης μέγιστου-γινομένου (max-product Belief Propagation). Παραθέτουμε αποτελέσματα από την εφαρμογή του προτεινόμενου αλγορίθμου και τη σύγκριση του με προϋπάρχουσες μεθόδους.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ενισχυτική μάθηση, Reinforcement Learning, RL, Πιθανοτικός Συμπερασμός, Probabilistic Inference, Αλγόριθμος προσδοκίας-μεγιστοποίησης, Expectation maximization, EM, Τροχιά μέγιστης πιθανότητας, Viterbi, Δυναμικά δίκτυα Bayes, Dynamic bayesian networks, DBN, Διαδικασίες αποφάσεων Markov, Markov Decision Process, MDP, Μοντέλα μίξης, Mixture models

Πρόλογος

Η εργασία που κρατάτε στα χέρια σας αποτελεί την διατριβή μου, για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στο Τμήμα Μηχανικών Παραγωγής και Διοίκησης, με κατεύθυνση σπουδών, τα Συστήματα Παραγωγής. Η εκπόνηση της ξεκίνησε στις αρχές του 2010 και ολοκληρώθηκε τον Σεπτέμβρη του ίδιου έτους, υπό την επίβλεψη και καθοδήγηση του καθηγητή κ. Νικόλαου Βλάσση.

Ο Νίκος είναι και ο πρώτος άνθρωπος που θα ήθελα να ευχαριστήσω. Για τις συναρπαστικές ώρες στην αίθουσα διδασκαλίας, την υποκίνηση, το ενδιαφέρον και την υπομονή του καθ' όλη τη διάρκεια της συνεργασίας μας.

Ένα μεγάλο ευχαριστώ στους φίλους και συναδέλφους μου κ. Ανδρέα Σιώλο και κ. Νίκο Θραψανιωτάκη για την στήριξη τους στα δύσκολα και την ώθηση τους στα ωραία και, φυσικά, ένα μεγάλο ευχαριστώ στην οικογένεια και τους φίλους μου, αν και θα ήταν άχαρο το να αρχίσω να κατονομάζω όλα τα αγαπημένα μου πρόσωπα.

στην οικογένεια

στους φίλους

στον έρωτα

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Στο paper τους 'Probabilistic Inference for solving discrete and continuous state Markov decision processes' (Toussaint and Storkey (2006)), οι Toussaint και Storkey δείξαν πως η συνάρτηση αξίας μιας διαδικασίας αποφάσεων Markov είναι ανάλογη της συνάρτησης πιθανότητας ενός κατάλληλα δομημένου μοντέλου μίξης. Ενός πιθανοτικού, δηλαδή, μοντέλου εκτίμησης πυκνότητας με τη χρήση μικτών κατανομών. Αυτό έκανε δυνατή τη χρήση πιθανοτικών τεχνικών συμπερασμού για την έμμεση μεγιστοποίηση της συνάρτησης αξίας. Μία από τις τεχνικές αυτές, ο αλγόριθμος Προσδοκίας-Μεγιστοποίησης (Expectation-Maximization, EM) είναι και η πιο συχνά χρησιμοποιούμενη. Στην παρακάτω εργασία αναπτύσσεται μία παραλλαγή του EM με κύριο άξονα την αξιοποίηση της τροχιάς μεγίστου πιθανότητας, ή όπως θα αποκαλείται στη συνέχεια του κειμένου, το Viterbi path, για την επίλυση προβλημάτων από το τομέα της ενισχυτικής μάθησης.

Ενισχυτική μάθηση (Reinforcement Learning, RL) είναι ένας τομέας της μηχανικής μάθησης στην επιστήμη υπολογιστών. Ασχολείται με το πως ένας πράκτορας πρέπει να επιλέγει τις δράσεις του στο περιβάλλον έτσι ώστε να μεγιστοποιήσει κάποιο είδος σωρευτικής αμοιβής. Οι αλγόριθμοι RL αποπειρώνται την εύρεση μιας πολιτικής η οποία καθορίζει τη δράση που ο πράκτορας οφείλει να πάρει σε κάθε κατάσταση. Το περιβάλλον μοντελοποιείται τυπικά ως μία πεπερασμένου αριθμού καταστάσεων (finite-state) διαδικασία αποφάσεων Markov (Markov decision process, MDP). Τέτοια θα είναι και τα προβλήματα που θα εξεταστούν στην εργασία αυτή.

Για να επιτύχουμε το σκοπό της εργασίας θα χρησιμοποιήσουμε εργαλεία από την επιστήμη της Στατιστικής και πιο συγκεκριμένα από τον τομέα του Στατιστικού Συμπερασμού. Ο τελευταίος ασχολείται με την εξαγωγή συμπερασμάτων από δεδομένα που υπόκεινται σε τυχαίες διακυμάνσεις. Στον τομέα αυτό συγκαταλέγονται εργαλεία όπως ο EM και αλγόριθμος Viterbi. Ο EM είναι μια επαναληπτική διαδικασία εύρεσης εκτιμήσεων μέγιστης πιθανότητας για παραμέτρους από στατιστικά μοντέλα, όπου το μοντέλο εξαρτάται από μη παρατηρήσιμες λανθάνουσες μεταβλητές. Ο αλγόριθμος Viterbi είναι ένας αλγόριθμος δυναμικού προγραμματισμού για την εύρεση της πιο πιθανής ακολουθίας κρυφών καταστάσεων - το Viterbi path - η οποία έχει σαν αποτέλεσμα μια ακολουθία παρατηρημένων γεγονότων.

Το υπόλοιπο κείμενο ακολουθεί την παρακάτω δομή:

Στο **Κ.2 - «Ενισχυτική μάθηση»**, αποπειράται μια εισαγωγή στην θεωρία του RL, με τις βασικές έννοιες και εργαλεία του τομέα καθώς και μια σύντομη ιστορική αναδρομή.

Στο **Κ.3 - «Αλγόριθμοι Προσδοκίας-Μεγιστοποίησης και Viterbi»**, δίνονται οι απαιτούμενες γνώσεις από την επιστήμη της Στατιστικής, για την κατανόηση των επόμενων κεφαλαίων.

Στο **Κ.4 - «Επίλυση διαδικασιών απόφασης Markov μέσω πιθανοτικού συμπερασμού»**, παρουσιάζεται ο αλγόριθμος EM για την εύρεση βέλτιστων πολιτικών σε MDP όπως πρωτοεκδόθηκε από τους Toussaint και Storkey, καθώς και ο νέος αλγόριθμος Viterbi-EM, για την επίλυση του ίδιου προβλήματος.

Στο **Κ.5 - «Πειραματικά Αποτελέσματα»**, παρατίθενται τα αποτελέσματα από την επίλυση των δοκιμαστικών προβλημάτων με τις μεθόδους EM και Viterbi - EM, καθώς και με Policy και Value iteration.

Τέλος, το **Κ.6 - «Συμπεράσματα»**, συγκεντρώνει τα αποτελέσματα της εργασίας.

Κεφάλαιο 2

Ενισχυτική Μάθηση

2.1 Εισαγωγή

Η πιο φυσική διαδικασία μάθησης, είναι ίσως η μάθηση μέσω της αλληλεπίδρασης με το περιβάλλον. Η Ενισχυτική Μάθηση (Reinforcement Learning, RL), εστιάζει στην κατευθυνόμενη από το στόχο μάθηση μέσω αλληλεπίδρασης, σε βαθμό πολύ μεγαλύτερο από ότι άλλες μέθοδοι μηχανικής μάθησης. Ενισχυτική Μάθηση είναι η μάθηση του απαιτούμενου τρόπου δράσης, μια αντιστοίχηση καταστάσεων σε δράσεις, έτσι ώστε να μεγιστοποιήσουμε ένα αριθμητικό σήμα αμοιβής. Ο πράκτορας δεν ενημερώνεται για το ποιες δράσεις να πάρει, όπως συνηθίζεται στις περισσότερες μεθόδους μηχανικής μάθησης, αλλά αντί αυτού πρέπει να ανακαλύψει ποιες δράσεις θα αποφέρουν την μεγαλύτερη αμοιβή μέσω δοκιμής. Στις πιο ενδιαφέρουσες και απαιτητικές περιπτώσεις, οι δράσεις μπορούν να επηρεάζουν όχι μόνο την άμεση αμοιβή αλλά και την επόμενη κατάσταση, και διαμέσου αυτής, όλες τις μελλοντικές αμοιβές. Τα δύο αυτά χαρακτηριστικά, αναζήτηση μέσω δοκιμής-σφάλματος (trial-and-error) και καθυστερημένη απόδοση αμοιβής, είναι τα πιο σημαντικά χαρακτηριστικά γνωρίσματα του RL.

Η ενισχυτική μάθηση ορίζεται όχι από τον χαρακτηρισμό μεθόδων μάθησης, αλλά από τον χαρακτηρισμό ενός προβλήματος μάθησης. Οποιαδήποτε μέθοδος αποτελεί ικανό εργαλείο για την επίλυση του προβλήματος αυτού, είναι μια μέθοδος RL. Η βασική ιδέα είναι απλά η σύλληψη των πιο σημαντικών πλευρών του πραγματικού προβλήματος, που ο μαθητευόμενος πράκτορας αντιμετωπίζει καθώς αλληλεπιδρά με το περιβάλλον του, για την επίτευξη ενός στόχου. Φυσικά, ένας τέτοιος πράκτορας θα πρέπει να είναι σε θέση να αντιληφθεί την κατάσταση του περιβάλλοντος του σε κάποιο βαθμό και να λάβει δράσεις για την αλλαγή του. Επίσης, πρέπει να διαθέτει έναν ή περισσότερους στόχους σχετιζόμενους με την κατάσταση του περιβάλλοντος. Η σύνθεση αυτή έχει ως σκοπό να συμπεριλάβει τα τρία μόλις αυτά στοιχεία – αίσθηση, δράση, στόχος – στην πιο δυνατή απλή τους μορφή, διατηρώντας όμως την εγκυρότητα του μοντέλου.

Η ενισχυτική μάθηση διαφέρει από την επιβλεπόμενη μάθηση (supervised learning), την πιο συχνά ίσως μελετημένη μορφή μηχανικής μάθησης. Η επιβλεπόμενη μάθηση τελείται μέσω της παροχής παραδειγμάτων από έναν εξωτερικό επιβλέποντα με γνώση στο εξεταζόμενο αντικείμενο. Αν και αποτελεί επίσης

ένα σημαντικό τομέα μάθησης, δεν είναι από μόνος του αρκετός για την μάθηση μέσω αλληλεπίδρασης. Σε διαδραστικά προβλήματα είναι συχνά μη πρακτική η απόκτηση παραδειγμάτων επιθυμητής συμπεριφοράς που να είναι και έγκυρα και αντιπροσωπευτικά όλων των καταστάσεων στις οποίες ο πράκτορας θα πρέπει να αντιδράσει. Σε ανεξερεύνητες καταστάσεις, όπου κάποιος θα περίμενε η μάθηση να είναι πιο ωφέλιμη, ένας πράκτορας θα πρέπει να μπορεί να μαθαίνει από τις προσωπικές του εμπειρίες.

Μία από τις προκλήσεις που εμφανίζονται στην ενισχυτική μάθηση και όχι σε άλλες μορφές μηχανικής μάθησης είναι η αντιστάθμιση μεταξύ εξερεύνησης και εκμετάλλευσης. Για την συγκέντρωση υψηλής αμοιβής, ο πράκτορας RL πρέπει να προτιμά δράσεις που στο παρελθόν αποδείχθηκαν αποτελεσματικές στην παραγωγή αμοιβής. Η ανακάλυψη όμως τέτοιων δράσεων απαιτεί τη δοκιμή δράσεων που δεν έχουν επιλεγεί ως τώρα. Ο πράκτορας πρέπει δηλαδή να εκμεταλλευτεί την ήδη αποκτηθείσα γνώση, αλλά ταυτόχρονα να εξερευνήσει ώστε να επιλέξει καλύτερες δράσεις στο μέλλον. Το κρίσιμο στοιχείο είναι ότι ο πράκτορας δεν μπορεί να ασχοληθεί κατά αποκλειστικότητα με την εξερεύνηση ή την εκμετάλλευση χωρίς να αποτύχει στο έργο του. Θα πρέπει να δοκιμάζει μια ποικιλία δράσεων και στην πορεία να προτιμά αυτές που εμφανίζονται καλύτερες. Σε μια στοχαστική διαδικασία, κάθε δράση πρέπει να επιλέγεται αρκετές φορές ώστε να παρθεί μία ρεαλιστική εκτίμηση της αναμενόμενης αμοιβής.

Ένα άλλο σημαντικό στοιχείο της ενισχυτικής μάθησης είναι η άμεση θεώρηση ολόκληρου του προβλήματος αλληλεπίδρασης ενός καθοδηγούμενου από το στόχο πράκτορα, με το περιβάλλον. Αυτό έρχεται σε αντίθεση με πολλές προσεγγίσεις που ορίζουν υποπροβλήματα χωρίς να επιλύουν το πρόβλημα σύνθεσης τους. Η ενισχυτική μάθηση ξεκινά με έναν πλήρη, αλληλεπιδραστικό πράκτορα σε αναζήτηση στόχου. Όλοι οι πράκτορες RL διαθέτουν προκαθορισμένους στόχους, δέχονται πληροφορίες για το περιβάλλον τους, και επιλέγουν δράσεις για τον επηρεασμό του. Επιπλέον, γίνεται συνήθως από την αρχή η υπόθεση ότι ο πράκτορας θα πρέπει να λειτουργεί ανεξαρτήτως σημαντικής αβεβαιότητας για το περιβάλλον του. Όταν η ενισχυτική μάθηση περιλαμβάνει την κατάστρωση σχεδίου (planning), θα πρέπει να ορίζεται η αλληλεπίδραση μεταξύ της σχεδίασης και της επιλογής δράσης σε πραγματικό χρόνο, καθώς και το πως τα μοντέλα του περιβάλλοντος αποκτώνται και βελτιώνονται.

2.1.1 Ιστορία της Ενισχυτικής μάθησης

Η ιστορία της ενισχυτικής μάθησης έχει δύο βασικούς άξονες οι οποίοι εξελίχθηκαν χωριστά προτού συνενωθούν στη μοντέρνα μορφή της. Ο ένας άξονας αφορά τη μάθηση μέσω δοκιμής και σφάλματος και ξεκίνησε στη ψυχολογία της ζωικής μάθησης. Ο άξονας αυτός διαπερνά κάποια από τα παλαιότερα έργα στην τεχνητή νοημοσύνη και οδήγησε στην αναγέννηση του RL στις αρχές του 1980. Ο άλλος άξονας αφορά το πρόβλημα του βέλτιστου ελέγχου και τη λύση του με χρήση συναρτήσεων αξίας και δυναμικού προγραμματισμού. Για το μεγαλύτερο διάστημα, ο άξονας αυτός δεν περιλάμβανε μάθηση. Παρόλη την γενική ανεξαρτησία των δύο αξόνων, εξαιρέσεις εμφανίζονται γύρω από έναν τρίτο, δυσκολότερα διακριτό άξονα, που αφορά τις μεθόδους χρονικών διαφορών (temporal difference). Οι

τρεις άξονες συνενώθηκαν κατά τα τέλη του 1980 και δημιούργησαν το σύγχρονο κλάδο της ενισχυτικής μάθησης.

Ο όρος 'βέλτιστος έλεγχος' βρήκε χρήση στα τέλη του 1950 ώστε να περιγράψει το πρόβλημα σχεδίασης ενός ελεγκτή για την ελαχιστοποίηση κάποιου μέτρου της συμπεριφοράς ενός δυναμικού συστήματος στο χρόνο. Μία από τις προσεγγίσεις σε αυτό το πρόβλημα αναπτύχθηκε στα μέσα του 1950 από τον Richard Bellman και άλλους μέσω της επέκτασης μιας θεωρίας του δεκάτου ενάτου αιώνα από τους Hamilton και Jacobi. Η προσέγγιση αυτή κάνει χρήση των εννοιών της κατάστασης ενός δυναμικού συστήματος και της συνάρτησης αξίας, ή 'συνάρτηση βέλτιστης απόδοσης', για τον ορισμό μιας εξίσωσης που συχνά στις μέρες μας αποκαλείται εξίσωση Bellman. Η κατηγορία μεθόδων για την επίλυση προβλημάτων βέλτιστου ελέγχου μέσω της επίλυσης αυτής της εξίσωσης έγινε γνωστή ως δυναμικός προγραμματισμός (Bellman (1957a)). Ο (Bellman (1957b)) εισήγαγε επίσης την διακριτή στοχαστική έκδοση του προβλήματος βέλτιστου ελέγχου, γνωστή ως διαδικασία αποφάσεων Markov (Markovian decision processes, MDPs), και ο Ron Howard (Howard (1960)) εφηύρε τη μέθοδο 'policy iteration' για MDPs. Όλα αυτά αποτελούν σημαντικά στοιχεία στήριξης της σημερινής θεωρίας και των αλγορίθμων της ενισχυτικής μάθησης.

Στα 1960 οι όροι 'ενίσχυση' (reinforcement) και 'ενισχυτική μάθηση' χρησιμοποιήθηκαν για πρώτη φορά σε κείμενα μηχανικής (όπως, Waltz and Fu (1965), Mendel (1966), Fu (1970), Mendel and McLaren (1970)). Ιδιαίτερη επιρροή είχε το κείμενο του Minsky 'Βήματα προς τη Τεχνητή Νοημοσύνη' (Minsky (1961)), όπου αναλύθηκαν πολλά θέματα σχετικά με το RL, συμπεριλαμβανομένου και του προβλήματος ανάθεσης ευσήμων, όπως το αποκάλεσε: Πως μοιράζεται η ευθύνη για την επιτυχία ανάμεσα στις πολλές αποφάσεις που μπορεί να οδήγησαν στην παραγωγή της; Το ερώτημα αυτό αποτελεί κατά ένα τρόπο την ουσία των μεθόδων RL.

2.2 Βασικές Έννοιες

Εκτός του πράκτορα και του περιβάλλοντος, μπορούν να εντοπιστούν τέσσερα κύρια στοιχεία ενός συστήματος ενισχυτικής μάθησης: η πολιτική, η συνάρτηση αμοιβών, η συνάρτηση αξίας, και προαιρετικά, ένα μοντέλο του περιβάλλοντος.

Μια πολιτική ορίζει την συμπεριφορά του πράκτορα για κάθε δεδομένη στιγμή. Πιο συγκεκριμένα, η πολιτική είναι μια αντιστοίχηση των αντιληφθέντων καταστάσεων του περιβάλλοντος σε δράσεις που πρέπει να εκτελεστούν στις καταστάσεις αυτές. Σε κάποιες περιπτώσεις η πολιτική μπορεί να είναι μια απλή συνάρτηση ή πίνακας αντιστοιχίας, ενώ σε άλλες μπορεί να περιλαμβάνει εκτενείς υπολογισμούς, όπως μια διαδικασία εύρεσης. Πυρήνας ενός πράκτορα RL, η πολιτική αρκεί για τον καθορισμό της συμπεριφοράς του. Γενικά, μια πολιτική μπορεί να είναι στοχαστική.

Η συνάρτηση αμοιβών ορίζει το στόχο σε ένα πρόβλημα RL. Το επιτυγχάνει αυτό αντιστοιχώντας κάθε αντιληφθείσα κατάσταση (ή ζεύγος δράσης-κατάστασης) του περιβάλλοντος σε μια αμοιβή, έναν αριθμό, μέτρο του επιθυμητού της κατάστασης (ή ζεύγους). Μοναδικός σκοπός ενός πράκτορα RL είναι η μεγιστοποίηση της συνολικά ληφθείσας αμοιβής. Η συνάρτηση αμοιβών ορίζει, ουσιαστικά,

τα θετικά και τα αρνητικά για τον πράκτορα γεγονότα και αποτελεί άμεσο και καθοριστικό στοιχείο του εξεταζόμενου προβλήματος. Ως τέτοια, η συνάρτηση αμοιβών πρέπει να παραμένει αμετάβλητη από τον πράκτορα, λειτουργώντας όμως ως βάση για τυχών αλλαγές στην πολιτική. Γενικώς, οι συναρτήσεις αμοιβών μπορεί να είναι στοχαστικές.

Ενώ μια συνάρτηση αμοιβών δείχνει τι είναι καλό κατά μία άμεση έννοια, η συνάρτηση αξίας καθορίζει τι είναι καλό μακροπρόθεσμα. Η αξία μιας κατάστασης είναι η συνολική αμοιβή που ένας πράκτορας αναμένεται να συσσωρεύσει στο μέλλον, ξεκινώντας από αυτή τη κατάσταση. Η συνάρτηση αξίας αποτελεί, έτσι, ένδειξη του πόσο επιθυμητή είναι μία κατάσταση σε βάθος χρόνου, έχοντας λάβει υπόψη της, τις αναμενόμενες επόμενες καταστάσεις και τις αντίστοιχες αμοιβές. Οι αξίες, λοιπόν, προκύπτουν από τις αμοιβές, και ο μόνος λόγος υπολογισμού τους είναι για την συσσώρευση περισσότερης αμοιβής. Παρόλα αυτά, οι αξίες αποτελούν το βασικό κριτήριο λήψης και αξιολόγησης αποφάσεων. Οι δράσεις επιλέγονται έτσι ώστε να βρεθούμε σε καταστάσεις μεγαλύτερης αξίας, όχι αμοιβής. Ενώ όμως οι αμοιβές καθορίζονται από το περιβάλλον, οι αξίες πρέπει να εκτιμούνται συνεχώς από την αλληλουχία παρατηρήσεων του πράκτορα κατά τη διάρκεια της ζωής του. Μάλιστα, το σημαντικότερο στοιχείο της πλειοψηφίας των αλγορίθμων ενισχυτικής μάθησης είναι μια μέθοδος για την αποτελεσματική εκτίμηση αξιών.

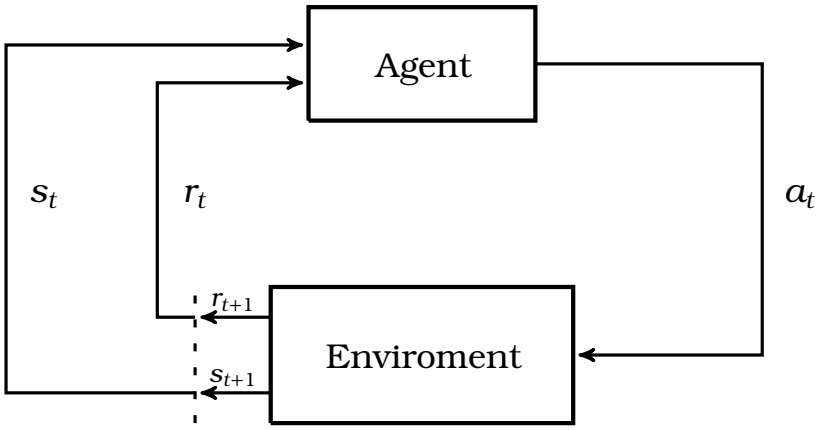
Το τέταρτο και τελευταίο στοιχείο κάποιων προβλημάτων RL είναι ένα μοντέλο του περιβάλλοντος. Για παράδειγμα, δεδομένης κατάστασης και δράσης, το μοντέλο μπορεί να προβλέπει την προκύπτουσα νέα κατάσταση και την αμοιβή. Τα μοντέλα χρησιμοποιούνται για το σχεδιασμό, με το οποίο νοείται οποιοσδήποτε τρόπος επιλογής δράσης μέσω της θεώρησης μελλοντικών καταστάσεων, προτού αυτές βιωθούν. Η ενσωμάτωση μοντέλων και σχεδιασμού σε προβλήματα RL είναι σχετικά καινούργια εξέλιξη. Τα πρώτα συστήματα RL ασχολούνταν αποκλειστικά με την μάθηση δοκιμής-σφάλματος (trial-and-error). Η δράση τους ήταν αντιληπτή ως σχεδόν αντίθετη του σχεδιασμού. Παρόλο αυτό, έγινε γρήγορα αντιληπτό πως οι μέθοδοι RL είναι στενά συνδεδεμένοι με τις μεθόδους δυναμικού προγραμματισμού, οι οποίες χρησιμοποιούν μοντέλα, και είναι με τη σειρά τους συγγενείς με τις μεθόδους σχεδιασμού χώρου καταστάσεων (state-space planning methods).

2.3 Το πρόβλημα Ενισχυτικής Μάθησης

Παρακάτω θα οριστεί το πρόβλημα της ενισχυτικής μάθησης. Οποιαδήποτε μέθοδος είναι σε θέση να επιλύσει το πρόβλημα αυτό είναι μια μέθοδος ενισχυτικής μάθησης.

2.3.1 Διεπαφή Πράκτορα-Περιβάλλοντος

Το πρόβλημα της ενισχυτικής μάθησης είναι μια καθαρή υλοποίηση του προβλήματος μάθησης μέσω αλληλεπίδρασης με το περιβάλλον για την εκπλήρωση κάποιου στόχου. Ο μαθητής και αποφασίζον αποκαλείται πράκτορας. Το σύνολο από αλληλεπιδρώντα στοιχεία, εκτός του πράκτορα, αποκαλείται περιβάλλον. Η αλληλεπίδραση είναι συνεχής, με τον πράκτορα να επιλέγει δράσεις



Σχήμα 2.1: Δυναμικό Δίκτυο Bayes (DBN) για διαδικασία αποφάσεων Markov (MDP). Με x συμβολίζονται οι μεταβλητές κατάστασης, a οι δράσεις και r οι αμοιβές.

και το περιβάλλον να ανταποκρίνεται σε αυτές παρουσιάζοντας νέες καταστάσεις. Παράλληλα, το περιβάλλον αποδίδει αμοιβές, αριθμητικές τιμές τις οποίες ο πράκτορας προσπαθεί να μεγιστοποιήσει σε βάθος χρόνου.

Πιο συγκεκριμένα, ο πράκτορας αλληλεπιδρά με το περιβάλλον σε κάθε μία από μια σειρά διακριτών χρονικών στιγμών, $t = 0, 1, 2, 3 \ldots$. Σε κάθε βήμα t, ο πράκτορας δέχεται κάποια ένδειξη της *κατάστασης* του περιβάλλοντος, $s_t \in S$, όπου S είναι το σύνολο των δυνατών καταστάσεων. Υπό αυτή, επιλέγει μία *δράση*, $a_t \in A(s_t)$, όπου $A(s_t)$ είναι το σύνολο των δυνατών δράσεων στην κατάσταση s_t. Την επόμενη χρονική στιγμή, εν μέρη ως αποτέλεσμα της δράσης του, ο πράκτορας λαμβάνει μια αμοιβή $r_{t+1} \in \mathbb{R}$, και βρίσκεται σε μια νέα κατάσταση s_{t+1}. Το Σχήμα (2.1) διαγράφει την αλληλεπίδραση πράκτορα-περιβάλλοντος.

Σε κάθε βήμα, ο πράκτορας υλοποιεί μια αντιστοιχία από κατάσταση σε πιθανότητες επιλογής δράσεων. Η αντιστοιχία αυτή αποκαλείται *πολιτική* του πράκτορα π_t, όπου $\pi_t(s, a)$ είναι η πιθανότητα $a_t = a$ αν $s_t = s$. Οι μέθοδοι RL ορίζουν το πως ένας πράκτορας αλλάζει την πολιτική του ως αποτέλεσμα των εμπειριών του. Στόχος του πράκτορα είναι η μεγιστοποίηση της συνολικά συλλεγόμενης αμοιβής σε βάθος χρόνου, και όχι της άμεσης αμοιβής.

2.3.2 Αμοιβή και Συσσωρευμένη Αμοιβή

Ο πράκτορας πάντα μαθαίνει να μεγιστοποιεί την αμοιβή του. Οι αμοιβές πρέπει λοιπόν να δοθούν με τέτοιο τρόπο, ώστε η μεγιστοποίηση τους να σημαίνει και την εκπλήρωση των στόχων μας από τον πράκτορα. Η τοποθέτηση των αμοιβών πρέπει να είναι τέτοια ώστε να αποτελεί πραγματική ένδειξη του επιθυμητού αποτελέσματος. Συγκεκριμένα, το σήμα αμοιβών δεν είναι κατάλληλο για την εκχώρηση γνώσης στον πράκτορα του *πως* να εκπληρώσει τους στόχους που του τέθηκαν.

Συμβολίζοντας την αλληλουχία λαμβανόμενων αμοιβών μετά το βήμα t ως $r_{t+1}, r_{t+2}, r_{t+3}, \ldots$, είναι επιθυμητή η μεγιστοποίηση της αναμενόμενης απόδοσης, όπου η απόδοση R_t ορίζετε ως συνάρτηση της αλληλουχίας αυτής. Στην πιο απλή

περίπτωση η απόδοση είναι το άθροισμα των αμοιβών

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \ldots + r_T, \tag{2.1}$$

όπου T η τελευταία χρονική στιγμή. Η προσέγγιση αυτή είναι βάσιμη σε εφαρμογές με διακριτή *κατάσταση τερματισμού*, εφαρμογές, δηλαδή, όπου η αλληλεπίδραση πράκτορα-περιβάλλοντος μπορεί να χωριστεί σε *επεισόδια*, όπως π.χ παρτίδες σκακιού.

Οι περιπτώσεις όμως όπου η αλληλεπίδραση πράκτορα-περιβάλλοντος δεν μπορεί να χωριστεί σε επεισόδια αλλά συνεχίζεται χωρίς όριο αποκαλούνται *συνεχείς*. Λόγω του ότι η τελική χρονική στιγμή είναι $T = \infty$, η απόδοση εύκολα απειρίζεται. Για αυτό εισάγεται η έννοια της *'μείωσης' (discounting)*. Ο πράκτορας επιλέγει τις δράσεις έτσι ώστε να μεγιστοποιήσει την *'μειωμένη' απόδοση*

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \tag{2.2}$$

όπου το *παράγοντας μείωσης (discount rate)* γ είναι παράμετρος με τιμές $0 \leq \gamma \leq 1$.

2.3.3 Η ιδιότητα Markov

Στη δομή του RL, ο πράκτορας αποφασίζει ως συνάρτηση ενός σήματος που λαμβάνει από το περιβάλλον, της κατάστασης. Παρακάτω αναλύονται κάποια χαρακτηριστικά του σήματος αυτού ενώ ορίζεται ένα ιδιαίτερα σημαντικό χαρακτηριστικό, η *ιδιότητα Markov*.

Με τον όρο 'κατάσταση' νοείται η πληροφορία που είναι διαθέσιμη στο πράκτορα. Όμως, το σήμα κατάστασης δεν ενημερώνει τον πράκτορα για το σύνολο του περιβάλλοντος αλλά ούτε και το σύνολο της χρήσιμης για την λήψη αποφάσεων πληροφορίας. Είναι δυνατή η ύπαρξη κρυφής πληροφορίας κατάστασης. Γνώσης, δηλαδή χρήσιμης για τον πράκτορα, αλλά για την οποία δεν έχει δεχτεί κάποιο σχετικό ερέθισμα.

Το επιθυμητό είναι ένα σήμα κατάστασης που θα συμπυκνώνει το παρελθόν, αλλά με τέτοιο τρόπο ώστε το σύνολο της χρήσιμης πληροφορίας να διατηρείτε. Ένα σήμα κατάστασης που επιτυγχάνει το τελευταίο αποκαλείται *Markov*, ή ότι έχει την *ιδιότητα Markov*. Κατά αυτό τον τρόπο, ένα σκακιστικό πρόβλημα – με τη θέση των κομματιών στη σκακιέρα και τον ενεργό παίχτη – θα αποτελούσε μια κατάσταση Markov επειδή συμπυκνώνει το σύνολο της χρήσιμης πληροφορίας από την πλήρη ακολουθία κινήσεων που οδήγησαν σε αυτό. Η έννοια της τρέχουσας κατάστασης είναι ανεξάρτητη της ιστορίας των σημάτων κατάστασης που οδήγησαν σε αυτή. Θα δοθεί τώρα ο ορισμός της ιδιότητας Markov για την περίπτωση διακριτών καταστάσεων, αμοιβών και χρονικών στιγμών:

Ένα σήμα κατάστασης λέγεται ότι διαθέτει την ιδιότητα Markov αν, για κάθε s', r, s_t και a_t, ισχύει

$$\begin{aligned} &Pr\{s_{t+1} = s', r_{t+1} = r | s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots, r_1, s_0, a_0\} = \\ &Pr\{s_{t+1} = s', r_{t+1} = r | s_t, a_t\}. \end{aligned} \tag{2.3}$$

Στη περίπτωση αυτή λέγεται πως το περιβάλλον και το πρόβλημα στο σύνολο του, έχουν την ιδιότητα Markov.

Εάν ένα περιβάλλον έχει την ιδιότητα Markov, τότε μπορούμε να προβλέψουμε την επόμενη κατάσταση και αμοιβή δεδομένων της τρέχουσας κατάστασης και δράσης. Η επανάληψη της πάνω εξίσωσης μάλιστα μπορεί να χρησιμοποιηθεί για την πρόβλεψη όλων των μελλοντικών καταστάσεων και αμοιβών, χρησιμοποιώντας μόνο την τρέχουσα κατάσταση, και μάλιστα εξίσου καλά με το αν γινόταν χρήση του συνόλου της ιστορίας ως αυτή τη στιγμή.

2.3.4 Διαδικασίες Απόφασης Markov

Ένα πρόβλημα ενισχυτικής μάθησης το οποίο ικανοποιεί την ιδιότητα Markov ονομάζεται *διαδικασία απόφασης Markov (Markov decision process, MDP)*. Αν ο χώρος των καταστάσεων και των δράσεων είναι πεπερασμένος, καλείται *πεπερασμένη διαδικασία απόφασης Markov (finite MDP)*. Ένα finite MDP ορίζεται από τα σύνολα καταστάσεων και δράσεων, καθώς και από την ενός-βήματος δυναμική του συστήματος. Οι *πιθανότητες μετάβασης (transition propabilities)* ορίζονται

$$P^a_{ss'} = Pr\{s_{t+1} = s' | s_t = s, a_t = a\}, \tag{2.4}$$

και είναι η πιθανότητα, δεδομένης κατάστασης s και δράσης a, η επόμενη κατάσταση να είναι s'. Παρομοίως, δεδομένης κατάστασης s, δράσης a και επόμενης κατάστασης s', η αναμενόμενη τιμή της επόμενης αμοιβής είναι

$$R^a_{ss'} = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\}. \tag{2.5}$$

2.3.5 Συναρτήσεις αξίας

Σχεδόν το σύνολο των αλγορίθμων RL βασίζονται στον υπολογισμό *συναρτήσεων αξίας*. Οι τελευταίες αποτελούν μια συνάρτηση της κατάστασης, ή ζεύγους κατάστασης-δράσης, που εκτιμούν το *πόσο καλή* είναι για τον πράκτορα μια δεδομένη κατάσταση, ή *πόσο καλή* είναι η επιλογή μιας δεδομένης δράσης στην κατάσταση αυτή. Το 'πόσο καλή' είναι η κατάσταση νοείται σε όρους προσδοκώμενης απόδοσης. Οι μελλοντικές αμοιβές του πράκτορα εξαρτώνται φυσικά από τις δράσεις που θα επιλέξει, με τις συναρτήσεις αξίας να καθορίζονται έτσι σε σχέση με συγκεκριμένες πολιτικές.

Η αξία μιας κατάστασης s υπό μια πολιτική π, $V^\pi(s)$, είναι η αναμενόμενη απόδοση ξεκινώντας από την κατάσταση s και ακολουθώντας την πολιτική π από εκεί και πέρα. Για MDP ορίζεται

$$V^\pi(s) = E_\pi\{R_t | s_t = s\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s\right\}, \tag{2.6}$$

όπου το $E_\pi\{\}$ δηλώνει την προσδοκώμενη τιμή δεδομένου ότι ο πράκτορας ακολουθεί την πολιτική π, και t είναι οποιαδήποτε χρονική στιγμή. Σημειώνεται ότι η αξία μιας τερματικής κατάστασης, εφόσον υπάρχει, είναι πάντα μηδέν. Η V^π αποκαλείται *συνάρτηση αξίας κατάστασης της πολιτικής* π.

Αντίστοιχα ορίζεται η αξία επιλογής μιας δράσης a, στην κατάσταση s υπό την πολιτική π, $Q^\pi(s,a)$, ως την αναμενόμενη απόδοση έναρξης στην s, επιλογής a, και στη συνέχεια ακολούθηση της πολιτικής π:

$$Q^\pi(s,a) = E_\pi\{R_t|s_t = s, a_t = a\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\middle| s_t = s, a_t = a\right\}. \tag{2.7}$$

Η Q^π αποκαλείται *συνάρτηση αξίας δράσης της πολιτικής* π.

Μια θεμελιώδης ιδιότητα των συναρτήσεων αξίας που χρησιμοποιείται τόσο στην ενισχυτική μάθηση όσο και στον δυναμικό προγραμματισμό είναι ότι ικανοποιούν συγκεκριμένες αναδρομικές σχέσεις. Για οποιαδήποτε πολιτική π και κατάσταση s ισχύει η παρακάτω συνθήκη συνοχής μεταξύ της αξίας του s και της αξίας των πιθανών μελλοντικών καταστάσεων:

$$\begin{aligned}
V^\pi(s) &= E_\pi\{R_t|s_t = s\} \\
&= E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\middle| s_t = s\right\} \\
&= E_\pi\left\{r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2}\middle| s_t = s\right\} \\
&= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'}\left[R^a_{ss'} + \gamma E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+2}\middle| s_{t+1} = s'\right\}\right] \\
&= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'}\left[R^a_{ss'} + \gamma V^\pi(s')\right],
\end{aligned} \tag{2.8}$$

όπου νοείται πως οι δράσεις, a, παίρνονται από το σύνολο $A(s)$, και ότι οι επόμενες καταστάσεις, s', παίρνονται από το σύνολο S. Η (2.8) είναι η εξίσωση *Bellman για το* V^π. Εκφράζει μια σχέση μεταξύ της αξίας μιας κατάστασης και της αξίας των διαδεχομένων αυτής καταστάσεων. Ουσιαστικά, παίρνει την μέση τιμή για όλες τις δυνατές καταστάσεις, σταθμίζοντας κάθε μία με την πιθανότητα της να προκύψει. Η συνάρτηση αξίας V^π αποτελεί την μοναδική λύση της εξίσωσης Bellman.

2.3.6 Βελτιστότητα

Μία πολιτική π ορίζεται ως καλύτερη ή ίση μιας άλλης π' αν η αναμενόμενη απόδοση της είναι μεγαλύτερη ή ίση της π' για όλες τις καταστάσεις. Με άλλα λόγια $\pi \geq \pi'$ αν και μόνο αν $V^\pi(s) \geq V^{\pi'}(s)$ για κάθε $s \in S$. Υπάρχει πάντα τουλάχιστον μια πολιτική η οποία είναι καλύτερη ή ίση των υπολοίπων, η *βέλτιστη πολιτική*. Αν και δεν είναι κατά ανάγκη μοναδικές, οι βέλτιστες πολιτικές σημειώνονται ως π^*. Μοιράζονται την ίδια συνάρτηση αξίας κατάστασης, V^*, που αποκαλείται *βέλτιστη συνάρτηση αξίας κατάστασης* και ορίζεται ως

$$V^*(s) = \max_\pi V^\pi(s), \tag{2.9}$$

για κάθε $s \in S$.

Οι βέλτιστες πολιτικές μοιράζονται επίσης την ίδια *βέλτιστη συνάρτηση αξίας δράσης*, Q^*, που ορίζεται ως

$$Q^*(s,a) = \max_\pi Q^\pi(s,a), \tag{2.10}$$

για κάθε $s \in S$ και $a \in A(s)$. Για κάθε ζεύγος κατάστασης-δράσης, η συνάρτηση αυτή δίνει την αναμενόμενη απόδοση του να εκτελεστεί η δράση στη κατάσταση, και στη συνέχεια να ακολουθηθεί η βέλτιστη πολιτική. Έτσι

$$Q^*(s,a) = E\{r_{t+1} + \gamma V^*(s_{t+1}) | s_t = s, a_t = a\}. \tag{2.11}$$

Η V^* ως συνάρτηση αξίας θα πρέπει να τηρεί τον περιορισμό (2.8). Επειδή όμως είναι η βέλτιστη, μπορεί να γραφεί με μία ειδική μορφή, χωρίς αναφορά σε κάποια πολιτική. Αυτή είναι η συνάρτηση Bellman για V^*, ή *συνάρτηση βελτιστότητας Bellman*:

$$\begin{aligned}
V^*(s) &= \max_{a \in A(s)} Q^{\pi^*}(s,a) \\
&= \max_a E_{\pi^*}\{R_t | s_t = s, a_t = a\} \\
&= \max_a E_{\pi^*}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s, a_t = a\right\} \\
&= \max_a E_{\pi^*}\left\{r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \middle| s_t = s, a_t = a\right\} \\
&= \max_a E\{r_{t+1} + \gamma V^*(s_{t+1}) | s_t = s, a_t = a\} \qquad (2.12) \\
&= \max_{a \in A(s)} \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^*(s')\right]. \qquad (2.13)
\end{aligned}$$

Οι δύο τελευταίες εξισώσεις είναι μορφές της εξίσωσης βελτιστότητας Bellman για V^*. Η εξίσωση βελτιστότητας Bellman για Q^* είναι

$$\begin{aligned}
Q^*(s,a) &= E\{r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a') | s_t = s, a_t = a\} \\
&= \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma \max_{a'} Q^*(s', a')]. \qquad (2.14)
\end{aligned}$$

Για finite MDPs, η εξίσωση βελτιστότητας Bellman (2.13) έχει μια μοναδική λύση ανεξάρτητη της πολιτικής. Η εξίσωση βελτιστότητας Bellman είναι ουσιαστικά ένα σύστημα εξισώσεων, μία για κάθε κατάσταση, οπότε αν υπάρχουν N καταστάσεις, τότε υπάρχουν N εξισώσεις πάνω σε N αγνώστους. Αν είναι γνωστή η δυναμική του συστήματος ($R^a_{ss'}$ και $P^a_{ss'}$), τότε γενικά μπορεί να λυθεί το μη-γραμμικό σύστημα προς V^*. Αναλόγως μπορεί να επιλυθεί και ένα σύνολο εξισώσεων προς Q^*.

Η βέλτιστη πολιτική προκύπτει εύκολα από το V^*, αν για κάθε κατάσταση βρεθούν μία ή περισσότερες δράσεις για τις οποίες αποκτάται το μέγιστο στη εξίσωση βελτιστότητας Bellman. Αυτό βρίσκεται με αναζήτηση των αποτελεσμάτων κάθε δράσης για ένα βήμα στο μέλλον. Οποιαδήποτε πολιτική ορίζει μη-μηδενικές πιθανότητες *μόνο* στις βέλτιστες δράσεις είναι μια βέλτιστη πολιτική. Η γνώση των Q^* κάνει ακόμη ευκολότερη την εύρεση των βέλτιστων δράσεων, αφού αρκεί, για κάθε κατάσταση, η επιλογή της δράσης που μεγιστοποιεί το $Q^*(s,a)$.

Η απευθείας επίλυση της εξίσωσης βελτιστότητας Bellman είναι ένας τρόπος εύρεσης της βέλτιστης πολιτικής, και έτσι, την επίλυση του προβλήματος ενισχυτικής μάθησης. Όμως αυτή η λύση σπάνια έχει άμεση χρησιμότητα. Δε διαφέρει από μία εξονυχιστική αναζήτηση. Μια τέτοια λύση βασίζεται σε τρεις τουλάχιστον υποθέσεις οι οποίες σπάνια ισχύουν στην πράξη:

1. Υπάρχει πλήρη γνώση της δυναμικής του περιβάλλοντος.

2. Υπάρχει αρκετή υπολογιστική ισχύς και μνήμη για τον πλήρη υπολογισμό της λύσης.

3. Ισχύει η ιδιότητα Markov.

Στα περισσότερα πρακτικά προβλήματα, η παραπάνω λύση δεν είναι δυνατή λόγω της αναίρεσης κάποιου συνδυασμού από τις παραπάνω υποθέσεις. Στην ενισχυτική μάθηση συνηθίζεται η εύρεση προσεγγιστικών λύσεων.

2.4 Μεθόδοι Ενισχυτικής Μάθησης

Υπάρχουν δύο βασικές στρατηγικές για την επίλυση προβλημάτων ενισχυτικής μάθησης. Η πρώτη είναι η εξερεύνηση του χώρου πολιτικών για την εύρεση μίας που να αποδίδει καλά στο περιβάλλον. Αυτή είναι η προσέγγιση των γενετικών αλγορίθμων και του γενετικού προγραμματισμού, καθώς και κάποιων πιο καινοφανών τεχνικών εύρεσης (Schmidhuber (1997)). Η δεύτερη είναι η χρήση στατιστικών τεχνικών και μεθόδων δυναμικού προγραμματισμού για την εκτίμηση της αξίας επιλογής της κάθε δράσης στο πρόβλημα. Η δεύτερη προσέγγιση, που θα αναλυθεί παρακάτω, είναι και η περισσότερο συνυφασμένη με το τομέα, καθώς μόνο οι τεχνικές αυτές εκμεταλλεύονται την ειδική δομή των προβλημάτων RL.

2.4.1 Δυναμικός Προγραμματισμός

Ο όρος δυναμικός προγραμματισμός (dynamic programming, DP) αναφέρεται σε μια συλλογή αλγορίθμων που μπορούν να χρησιμοποιηθούν για τον υπολογισμό βέλτιστων πολιτικών δεδομένου ενός τέλειου μοντέλου του περιβάλλοντος ως ένα MDP. Οι κλασικοί αλγόριθμοι DP έχουν περιορισμένη αξία στην ενισχυτική μάθηση, τόσο λόγω της υπόθεσης ενός τέλειου μοντέλου, όσο και λόγω των εξαιρετικά μεγάλων υπολογιστικών απαιτήσεων. Έχουν ωστόσο, σημαντική θεωρητική αξία.

Θα δειχθεί πως ο δυναμικός προγραμματισμός μπορεί να χρησιμοποιηθεί για τον υπολογισμό των συναρτήσεων αξίας, όπως ορίστηκαν στο (2.3.5). Όπως εκεί δείχθηκε, η απόκτηση βέλτιστων πολιτικών γίνεται εύκολα εφόσον έχουν βρεθεί οι βέλτιστες συναρτήσεις αξίας οι οποίες ικανοποιούν τις εξισώσεις βελτιστότητας Bellman (2.13, 2.14). Οι αλγορίθμοι DP αποκτώνται με την μετατροπή των εξισώσεων Bellman σε αναθέσεις, δηλαδή, σε κανόνες ενημέρωσης για συνεχώς βελτιούμενες προσεγγίσεις των επιθυμητών συναρτήσεων αξίας.

Παρακάτω υποθέτεται πως το περιβάλλον είναι finite MDP, όπως ορίστηκε στο (2.3.4).

Αξιολόγηση Πολιτικής

Αξιολόγιση πολιτικής (policy evaluation) είναι ο υπολογισμός της συνάρτησης αξίας κατάστασης για μια πολιτική. Στο (2.8) είδαμε ότι το $V^{\pi}(s)$ εξαρτάται από

την ακολουθούμενη πολιτική π. Η ύπαρξη και μοναδικότητα του $V^\pi(s)$ εξασφαλίζεται εφόσον είτε $\gamma < 1$ είτε ο τερματισμός είναι δεδομένος υπό από όλες τις καταστάσεις υπό την πολιτική π.

Αν η δυναμική του περιβάλλοντος είναι πλήρως γνωστή, τότε η (2.8) είναι ένα σύστημα $|S|$ ταυτόχρονων γραμμικών εξισώσεων $|S|$ αγνώστους. Γενικά η επίλυση του είναι άμεση αλλά επίπονη. Προτιμώνται επαναληπτικές μέθοδοι λύσης. Μετατρέποντας την εξίσωση Bellman (2.8) σε κανόνα ενημέρωσης έχουμε:

$$\begin{aligned} V_{k+1}(s) &= E_\pi\{r_{t+1} + \gamma V_k(s_{t+1}) | s_t = s\} \\ &= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'} \Big[R^a_{ss'} + \gamma V_k(s') \Big]. \end{aligned} \tag{2.15}$$

για κάθε $s \in S$. Είναι εμφανές πως το $V_k = V^\pi$ αποτελεί σταθερό σημείο του κανόνα ενημέρωσης λόγω της εξίσωσης του Bellman. Η πάνω ακολουθία συγκλίνει καθώς $k \to \infty$ και ο αντίστοιχος αλγόριθμος ονομάζεται *επαναληπτική αξιολόγηση πολιτικής (iterative policy evaluation)*.

Για την παραγωγή κάθε νέας προσέγγισης V_{k+1} από V_k, η επαναληπτική αξιολόγηση πολιτικής εφαρμόζει την ίδια διαδικασία σε κάθε κατάσταση s: αντικαθιστά την παλιά τιμή με μία νέα που έχει παραχθεί από τις παλιές τιμές των απογόνων του s, και τις προσδοκώμενες άμεσες αμοιβές, βάση όλων των μεταβάσεων που είναι δυνατές σε ένα χρονικό βήμα και κάτω από την αξιολογούμενη πολιτική. Μια τέτοια διαδικασία αποκαλείται 'full backup'. Κάθε επανάληψη του αλγορίθμου 'backs up' την αξία κάθε κατάστασης μία φορά, για την παραγωγή της νέας προσέγγισης V_{k+1}.

Βελτίωση Πολιτικής

Σύμφωνα με το *θεώρημα βελτίωσης πολιτικής (policy improvement theorem)*, αν για ζεύγος ντετερμινιστικών πολιτικών π και π', ισχύει για κάθε για κάθε $s \in S$,

$$Q^\pi(s, \pi'(s)) \geq V^\pi(s), \tag{2.16}$$

τότε η πολιτική π' πρέπει να είναι τουλάχιστον εξίσου καλή με την π, αν όχι καλύτερη. Δηλαδή, πρέπει να έχει μεγαλύτερη ή ίση προσδοκώμενη απόδοση για όλες τις καταστάσεις $s \in S$:

$$V^{\pi'}(s) \geq V^\pi(s), \tag{2.17}$$

Επιπλέον, αν ισχύει η ισχυρή ανισότητα της (2.16) για οποιαδήποτε κατάσταση, τότε θα πρέπει να υπάρχει αυστηρή ανισότητα της (2.17) για τουλάχιστον μία κατάσταση.

Με τα παραπάνω δεδομένα μπορούμε εύκολα να αξιολογήσουμε το αποτέλεσμα της αλλαγής της πολιτικής σε μία κατάσταση για μία συγκεκριμένη δράση. Φυσική προέκταση αυτού είναι η θεώρηση αλλαγών σε *όλες* τις καταστάσεις και για *όλες* τις δυνατές δράσεις, επιλέγοντας σε κάθε κατάσταση τη δράση που εμφανίζεται καλύτερη βάση της $Q^\pi(s,a)$. Με άλλα λόγια, την θεώρηση της νέας

'greedy' πολιτικής, π', όπως δίνεται από την εξίσωση

$$\begin{aligned}\pi'(s) &= \arg\max_a Q^\pi(s,a) \\ &= \arg\max_a E\{r_{t+1} + \gamma V^\pi(s_{t+1}) | s_t = s, a_t = a\} \\ &= \arg\max_a \sum_{s'} P^a_{ss'}\Big[R^a_{ss'} + \gamma V^\pi(s')\Big], \end{aligned} \tag{2.18}$$

όπου το $\arg\max_a$ δηλώνει την δράση a για την οποία η έκφραση που ακολουθεί μεγιστοποιείται. Η 'greedy' αυτή πολιτική παίρνει τη δράση που δείχνει καλύτερη βραχυπρόθεσμα, σε βάθος μιας χρονικής στιγμής, σύμφωνα με το V^π. Από την κατασκευή της, η 'greedy' αυτή πολιτική ικανοποιεί τις συνθήκες του θεωρήματος (2.16). Η διαδικασία βελτίωσης μιας πολιτικής κάνοντας την 'greedy' πάνω στην αρχική συνάρτηση αξίας της αποκαλείται *βελτίωση πολιτικής (policy improvement)*.

Policy Iteration

Μόλις μια πολιτική, π, βελτιωθεί χρήση V^π για την εύρεση μιας καλύτερης πολιτικής, π', μπορούμε να υπολογίσουμε το $V^{\pi'}$, βελτιώνοντας την ξανά για να βρούμε την π''. Δημιουργούμε έτσι μια διαδικασία μονοτονικής βελτίωσης πολιτικών και συναρτήσεων αξίας:

$$\pi_0 \xrightarrow{E} V^{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} V^{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} \dots \xrightarrow{I} \pi^* \xrightarrow{E} V^*, \tag{2.19}$$

όπου το $\xrightarrow{E}$ δηλώνει *αξιολόγιση* πολιτικής και το $\xrightarrow{I}$ δηλώνει *βελτίωση* πολιτικής. Κάθε πολιτική είναι εγγυημένα καλύτερη της προηγούμενης, εκτός και αν αυτή είναι η βέλτιστη. Λόγω του πεπερασμένου αριθμού πολιτικών ενός finite MDP, η διαδικασία πρέπει να συγκλίνει προς μία βέλτιστη πολιτική και συνάρτηση αξίας σε πεπερασμένο αριθμό επαναλήψεων.

Ο τρόπος αυτός εύρεσης μιας βέλτιστης πολιτικής καλείται *'policy iteration'*.

Value Iteration

Ένα μειονέκτημα του 'policy iteration' είναι ότι κάθε επανάληψη της περιλαμβάνει αξιολόγηση πολιτικής, που με τη σειρά της μπορεί να απαιτεί πληθώρα περασμάτων του συνόλου των καταστάσεων. Εάν η αξιολόγηση πολιτικής γίνει επαναληπτικά η ακριβής σύγκλιση στο V^π συμβαίνει μόνο στο όριο. Το βήμα αξιολόγησης της πολιτικής, όμως, μπορεί να συμπτυχθεί με αρκετούς τρόπους, χωρίς να χαθεί η εγγύηση σύγκλισης του 'policy iteration'. Μία σημαντική ειδική περίπτωση αυτού, είναι όταν η αξιολόγηση πολιτικής σταματά μετά από ένα μόλις βήμα. Ο αλγόριθμος αυτός αποκαλείται 'value iteration'. Μπορεί να γραφτεί σαν μία ιδιαίτερα απλή διαδικασία 'backup' που συνδυάζει την βελτίωση πολιτικής με συμπτυγμένα βήματα αξιολόγισης πολιτικής:

$$\begin{aligned} V_{k+1}(s) &= \max_a E\{r_{t+1} + \gamma V_k(s_{t+1}) | s_t = s, a_t = a\} \\ &= max_a \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V_k(s')\right], \end{aligned} \tag{2.20}$$

για κάθε $s \in S$. Για αυθαίρετο V_0, η ακολουθία $\{V_k\}$ μπορεί να δειχθεί ότι συγκλίνει στο V^*, υπό τις ίδιες συνθήκες που απαιτούνται για την ύπαρξη του V^*.

2.4.2 Μέθοδοι Monte Carlo

Οι μέθοδοι Monte Carlo είναι μέθοδοι μάθησης, για την εκτίμηση των συναρτήσεων αξίας και την ανακάλυψη βέλτιστων πολιτικών. Το μοντέλο του περιβάλλοντος δεν θεωρείτε γνωστό καθώς οι μεθόδοι Monte Carlo απαιτούν μόνο *εμπειρία*. Εμπειρία που αποκτάται μέσω δοκιμαστικών τροχιών καταστάσεων, δράσεων και αμοιβών είτε κατά την εκτέλεση της εργασίας ('on-line') είτε μέσω της προσομοιωμένης αλληλεπίδρασης με το περιβάλλον. Η on-line μάθηση παρουσιάζει ενδιαφέρων καθώς δεν απαιτεί προηγούμενη γνώση του συστήματος, αλλά παρόλα αυτά μπορεί να οδηγήσει σε βέλτιστη πολιτική. Χρήσιμη είναι και η μάθηση από προσομοιωμένα συστήματα γιατί αν και απαιτείται μοντέλο, αυτό χρειάζεται απλά να είναι σε θέση να παράγει δοκιμαστικές τροχιές, και όχι πλήρες κατανομές πιθανοτήτων του συνόλου των δυνατών μεταβάσεων, όπως στις μεθόδους δυναμικού προγραμματισμού.

2.4.3 Μάθηση Χρονικών Διαφορών

Η μάθηση χρονικών διαφορών (temporal-difference (TD) learning) είναι ένας συνδυασμός ιδεών από τις μεθόδους Monte Carlo δυναμικού προγραμματισμού. Όπως και στις Monte Carlo, οι μέθοδοι TD μπορούν να μάθουν κατευθείαν από εμπειρία, χωρίς κάποιο μοντέλο του περιβάλλοντος. Όπως στο δυναμικό προγραμματισμό, έτσι και οι μέθοδοι TD ενημερώνουν τις εκτιμήσεις τους βασισμένες εν μέρη στις υπάρχουσες εκτιμήσεις, χωρίς την αναμονή ενός τελικού αποτελέσματος. Η σχέση μεταξύ της μάθησης χρονικών διαφορών, του δυναμικού προγραμματισμού και των μεθόδων Monte Carlo είναι ένα επαναλαμβανόμενο θέμα στην θεωρία της ενισχυτικής μάθησης.

2.5 Εξελίξεις στην ενισχυτική μάθηση

Ο τομέας της ενισχυτικής μάθησης βρίσκεται σε συνεχή εξέλιξη και το παραπάνω κείμενο, που σκοπό είχε την εμπέδωση βασικών εννοιών, δεν θα μπορούσε να καλύψει μεγάλο κομμάτι της υπάρχουσας θεωρίας. Τρέχοντα θέματα έρευνας είναι, μεταξύ άλλων: Οι εναλλακτικές αντιπροσωπεύσεις χώρου κατάστασης, αλγόριθμοι κατάβασης πλαγιάς στο χώρο πολιτικών (gradient descent in policy space) (Baird and Moore (1998), Hansen (1998)), αλγόριθμοι και αποτελέσματα σύγκλισης για μερικώς παρατηρήσιμες διαδικασίες απόφασης Markov (Partially observable Markov decision process, POMDP) (Vlassis and Toussaint (2009), Spaan and Vlassis (2005), Hsu (Lee), Cassandra (Littman and Zhang)), αρθρωτή και ιεραρχική ενισχυτική μάθηση (Modular, hierarchical reinforcement learning) (Barto and Mahadevan (2003)). Η πολυπρακτορική ή κατανεμημένη ενισχυτική μάθηση (Multiagent / Distributed reinforcement learning) (Claus and Boutilier (1998)) αποτελεί επίσης σημαντικό θέμα έρευνας στον τομέα.

Κεφάλαιο 3

Αλγόριθμοι Προσδοκίας-Μεγιστοποίησης και Viterbi

Παρακάτω περιγράφεται το πρόβλημα εκτίμησης παραμέτρων μέγιστης πιθανότητας (maximum-likelihood parameter estimation problem) και πως ο αλγόριθμος προσδοκίας-μεγιστοποίησης (Expectation-Maximization algorithm, EM) μπορεί να χρησιμοποιηθεί για την λυση του. Πρώτα περιγράφεται η αφηρημένη μορφή του αλγορίθμου EM όπως συχνά συναντάται στην βιβλιογραφία. Έπειτα παράγεται η διαδικασία εκτίμησης παραμέτρων EM για δύο εφαρμογές:

1. Εύρεση των παραμέτρων μιας μίξης πυκνοτήτων Gauss (mixture of Gaussian densities) και

2. Εύρεση των παραμέτρων ενός κρυφού μοντέλου Markov (hidden Markov model, HMM), δηλαδή τον αλγόριθμο Baum-Welch, για διακριτά και μίξης πυκνοτήτων Gauss μοντέλα παρατηρήσεων.

Επίσης θα παρουσιαστεί ο αλγόριθμος Viterbi, ο οποίος, δεδομένου ενός συνόλου παρατηρήσεων και των παραμέτρων ενός μοντέλου, βρίσκει την καλύτερη ακολουθία καταστάσεων που εξηγεί τις παρατηρήσεις.

3.1 Μέγιστη Πιθανότητα

Στην παράγραφο αυτή ορίζεται το πρόβλημα εκτίμησης παραμέτρων μέγιστης πιθανότητας. Έστω συνάρτηση πυκνότητας $p(x|\Theta)$ που ορίζεται από ένα σύνολο παραμέτρων Θ (π.χ., p μπορεί να είναι ένα σύνολο κατανομών Gauss και Θ οι μέσες τιμές και οι συνδιασπορές). Έχουμε επίσης ένα σύνολο δεδομένων μεγέθους N, υποθετικά τραβηγμένα από την κατανομή αυτή, δηλαδή, $X = \{x_1, \ldots, x_N\}$. Υποθέτουμε έτσι ότι τα διανύσματα δεδομένων αυτά είναι ανεξάρτητα και ομοιόμορφα κατανεμημένα με κατανομή p. Για αυτό, η τελική πυκνότητα του δείγματος είναι

$$p(X|\Theta) = \prod_{i=1}^{N} p(x_i|\Theta) = L(\Theta|X). \tag{3.1}$$

Η συνάρτηση $L(\Theta|X)$ καλείται πιθανότητα των παραμέτρων βάση των δεδομένων, ή απλά συνάρτηση πιθανότητας. Η πιθανότητα νοείται ως μία συνάρτηση των παραμέτρων Θ όπου τα δεδομένα X είναι σταθερά. Στο πρόβλημα μέγιστης πιθανότητας, στόχος είναι η εύρεση του Θ που να μεγιστοποιεί το L. Η εύρεση, δηλαδή, του Θ^* όπου

$$\Theta^* = \arg\max_{\Theta} L(\Theta|X). \tag{3.2}$$

Συχνά προτιμάται η μεγιστοποίηση του $\log L(\Theta|X)$ ως πιο εύκολη αναλυτικά.

Η δυσκολία του παραπάνω προβλήματος είναι ανάλογη της μορφής του $p(x|\Theta)$. Για παράδειγμα, εάν το $p(x|\Theta)$ είναι απλά μία κατανομή Gauss όπου $\Theta = (\mu, \sigma^2)$, τότε μπορεί να τεθεί η παράγωγος του $\log L(\Theta|X)$ στο μηδέν, και να λυθεί άμεσα προς μ και σ^2. Το τελευταίο μάλιστα οδηγεί στις τυποποιημένες εξισώσεις για την μέση τιμή και τη συνδιασπορά ενός συνόλου δεδομένων. Για πολλά προβλήματα, όμως, δεν είναι δυνατή η εύρεση μιας αναλυτικής έκφρασης, και απαιτούνται πιο περίπλοκες μέθοδοι.

3.2 Βασικός αλγόριθμος Προσδοκίας-Μεγιστοποίησης

Ο αλγόριθμος EM (Dempster (Laird and Rubin), Rabiner and Juang (1993)) είναι μία γενική μέθοδος εύρεσης της μεγίστου πιθανότητας εκτίμησης των παραμέτρων μίας κατανομής από ένα δοθέν σύνολο δεδομένων, όταν το τελευταίο είναι μη πλήρες ή του λείπουν τιμές. Δύο είναι οι βασικές εφαρμογές του αλγορίθμου EM. Η πρώτη προκύπτει όταν τα δεδομένα πράγματι λείπονται τιμών, λόγω προβλημάτων ή περιορισμών της διαδικασίας παρατηρήσεων. Η δεύτερη προκύπτει όταν η βελτιστοποίηση μιας συνάρτησης πιθανότητας είναι αναλυτικά δύσχρηστη αλλά η συνάρτηση μπορεί να απλοποιηθεί υποθέτοντας την ύπαρξη ελλειπόντων ή κρυμμένων παραμέτρων.

Όπως και προηγουμένως, υποθέτεται ότι τα δεδομένα X παρατηρούνται και παράγονται από κάποια κατανομή. Το X καλείται *μη-πλήρη δεδομένα*. Υποθέτεται πως ένα πλήρες σύνολο δεδομένων $Z = (X, Y)$ υφίσταται και υποθέτεται επίσης (ή ορίζεται) η κοινή συνάρτηση πυκνότητας πιθανότητας:

$$p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)p(x|\Theta). \tag{3.3}$$

Ένα ζήτημα είναι η προέλευση της κοινής συνάρτησης πυκνότητας. Συχνά προκύπτει από την οριακή συνάρτηση πυκνότητας $p(x|\Theta)$ και την υπόθεση κρυφών μεταβλητών και εκτιμήσεων τιμής παραμέτρων (όπως στις πυκνότητες μίξης και στον Baum-Welch). Σε άλλες περιπτώσεις (π.χ., ελλειπόντων τιμών δεδομένων σε δείγματα κατανομής), απαιτείται η υπόθεση σχέσης εξάρτησης των ελλειπόντων και των παρατηρημένων τιμών.

Με την νέα αυτή συνάρτηση πυκνότητας, μπορούμε να ορίσουμε μια νέα συνάρτηση πιθανότητας, $L(\Theta|Z) = L(\Theta|X, Y) = p(X, Y|\Theta)$, που ονομάζεται πιθανότητα συνολικών δεδομένων (complete-data likelihood). Η συνάρτηση αυτή, βέβαια, είναι μια τυχαία μεταβλητή, καθώς η υπολειπόμενη πληροφορία Y είναι άγνωστη, τυχαία, και υποθετικά οριζόμενη από μία βασική κατανομή. Μπορεί έτσι να τεθεί $L(\Theta|X, Y) = h_{X,\Theta}(Y)$ για κάποια συνάρτηση $h_{X,\Theta}(\cdot)$ όπου τα X και Θ είναι σταθερές και το Y είναι μια τυχαία μεταβλητή. Η αρχική πιθανότητα $L(\Theta|X)$

αποκαλείται συνάρτηση πιθανότητας ημιτελών δεδομένων (incomplete-data likelihood function).

Ο αλγόριθμος ΕΜ βρίσκει πρώτα την αναμενόμενη τιμή της *log*-πιθανότητας συνολικών δεδομένων $\log p(X, Y|\Theta)$ πάνω στα άγνωστα δεδομένα Y δεδομένου των παρατηρούμενων X και των τρεχόντων εκτιμήσεων παραμέτρων. Ορίζεται λοιπόν:

$$Q(\Theta, \Theta^{i-1}) = E\{\log p(X, Y|\Theta) \mid X, \Theta^{i-1}\}, \tag{3.4}$$

όπου Θ^{i-1} είναι οι τρέχουσες εκτιμήσεις των παραμέτρων που χρησιμοποιούνται για την εκτίμηση της προσδοκίας, και Θ είναι οι νέες παράμετροι που βελτιστοποιούνται για την αύξηση του Q.

Τα βασικότερα σημεία της παραπάνω σχέσης είναι ότι οι X και Θ^{i-1} είναι σταθερές, Θ είναι μια μεταβλητή που ζητείται να ρυθμιστεί, και Y είναι μια τυχαία μεταβλητή οριζόμενη από την κατανομή $f(y|X, \Theta^{i-1})$. Το δεξί μέλος της (3.4) μπορεί να ξαναγραφτεί ως:

$$E\{\log p(X, Y|\Theta) \mid X, \Theta^{i-1}\} = \int_{y \in Y} \log p(X, y|\Theta)\, f(y|X, \Theta^{i-1})\, \mathrm{d}y. \tag{3.5}$$

$f(y|X, \Theta^{i-1})$ είναι η οριακή κατανομή των μη παρατηρούμενων δεδομένων και εξαρτάται τόσο από τα παρατηρούμενα δεδομένα X όσο και από τις τρέχουσες παραμέτρους, ενώ Y είναι ο χώρος των τιμών του y. Στην καλύτερη των περιπτώσεων, η οριακή αυτή κατανομή είναι μια απλή αναλυτική έκφραση των παραμέτρων Θ^{i-1} και ίσως των δεδομένων. Στην χειρότερη των περιπτώσεων, η πυκνότητα αυτή υπολογίζεται πολύ δύσκολα. Ορισμένες φορές, μάλιστα, η πυκνότητα που χρησιμοποιείται είναι η $f(y, X|\Theta^{i-1}) \cdot f(X|\Theta^{i-1})$, πράγμα που δεν επηρεάζει όμως τα υπόλοιπα βήματα, δεδομένου ότι ο όρος $f(X|\Theta^{i-1})$ δεν εξαρτάται από το Θ.

Ως μια αναλογία, έστω συνάρτηση $h(\cdot, \cdot)$ δύο μεταβλητών. Ας θεωρηθεί $h(\vartheta, Y)$ όπου ϑ μια σταθερά και Y μια τυχαία μεταβλητή οριζόμενη από κάποια κατανομή $f_Y(y)$. Τότε η $q(\vartheta) = E_Y[h(\vartheta, Y)] = \int_y h(\vartheta, y)\, f_Y(y)\, \mathrm{d}y$ είναι τώρα μια ντετερμινιστική συνάρτηση που μπορεί να μεγιστοποιηθεί.

Ο υπολογισμός της αναμενόμενης τιμής καλείται το βήμα προσδοκίας (**E-step**) του αλγορίθμου. Σημασία πρέπει να δωθεί στις παραμέτρους της συνάρτησης $Q(\Theta, \Theta^{i-1})$. Ο πρώτος όρος Θ αντιστοιχεί στις παραμέτρους που τελικά θα βελτιστοποιηθούν στην προσπάθεια μεγιστοποίησης της πιθανότητας. Ο δεύτερος όρος Θ^{i-1} αντιστοιχεί στις παραμέτρους που χρησιμοποιούνται για την εκτίμηση της προσδοκίας.

Το δεύτερο βήμα του αλγορίθμου, το βήμα μεγιστοποίησης (**M-step**), εκτελεί την μεγιστοποίηση της προσδοκίας που υπολογίστηκε στο πρώτο βήμα. Βρίσκεται δηλαδή, το:

$$\Theta^i = \arg\max_{\Theta} Q(\Theta, \Theta^{i-1}). \tag{3.6}$$

Τα δύο βήματα, E-step και M-step, επαναλαμβάνονται μέχρι σύγκλισης. Κάθε επανάληψη έχει αποδειχθεί πως αυξάνει την *log*-πιθανότητα και ο αλγόριθμος συγκλίνει εγγυημένα σε ένα τοπικό ελάχιστο της συνάρτησης πιθανότητας.

Μία παραλλαγή του M-step του αλγορίθμου είναι, αντί της μεγιστοποίησης του $Q(\Theta, \Theta^{i-1})$, να αναζητείται κάποιο Θ^i έτσι ώστε να ισχύει $Q(\Theta^i, \Theta^{i-1}) > Q(\Theta, \Theta^{i-1})$. Αυτή η μορφή του αλγορίθμου αποκαλείται γενικευμένος αλγόριθμος προσδοκίας-μεγιστοποίησης (Generalized EM, GEM) και συγκλίνει επίσης εγγυημένα.

Η πάνω παρουσίαση του αλγορίθμου ήταν στην γενική του μορφή, χωρίς να εξηγείται η υλοποίηση του. Οι λεπτομέρειες των βημάτων εξαρτώνται σε μεγάλο βαθμό από τη συγκεκριμένη εφαρμογή.

3.3 Εύρεση των Παραμέτρων Μικτών Πυκνοτήτων Μέγιστης Πιθανότητας

Το πρόβλημα υπολογισμού πυκνότητας μίξης είναι πιθανότατα μια από της πιο ευρέως χρησιμοποιούμενες εφαρμογές του αλγορίθμου ΕΜ. Εδώ υποθέτεται το παρακάτω πιθανοτικό μοντέλο:

$$p(x|\Theta) = \sum_{i=1}^{M} a_i p_i(x|\vartheta_i) \tag{3.7}$$

όπου οι παράμετροι είναι $\Theta = (a_1, \ldots, a_M, \vartheta_1, \ldots, \vartheta_M)$ έτσι ώστε $\sum_{i=1}^{M} a_i = 1$ και κάθε p_i είναι μια συνάρτηση πυκνότητας που παραμετροποιείται κατά ϑ_i. Με άλλα λόγια, υποθέτεται πως υπάρχουν M συνιστώσες πυκνότητας σε μίξη, με M συντελεστές μίξης a_i.

Η έκφραση της *log*-πιθανότητας των ημιτελών δεδομένων από το X είναι:

$$\log L(\Theta|X) = \log \prod_{i=1}^{N} p(x_i|\Theta) = \sum_{i=1}^{N} \log\Big(\sum_{j=1}^{M} a_j p_j(x_i|\vartheta_j)\Big) \tag{3.8}$$

που είναι δύσκολο να βελτιστοποιηθεί λόγω του λογαρίθμου του αθροίσματος που περιέχει. Αν θεωρηθεί το X ως ημιτελές, όμως, και θεωρηθεί η ύπαρξη μη παρατηρήσιμων δεδομένων $Y = \{y_i\}_{i=1}^{N}$ με τιμές που πληροφορούν πιο τμήμα της μικτής πυκνότητας 'παρήγαγε' κάθε δεδομένο, η έκφραση της πιθανότητας απλοποιείται σημαντικά. Υποθέτονται $y_i \in 1, \ldots, M$ για κάθε i, και $y_i = k$ αν το δείγμα i παράχθηκε από την συνιστώσα μίξης k. Έχοντας γνώση των τιμών του Y, η πιθανότητα γίνεται:

$$\log L(\Theta|X, Y) = logP(X, Y|\Theta) = \sum_{i=1}^{N} \log(P(x_i|y_i)P(y)) = \sum_{i=1}^{N} \log(a_{y_i} p_{y_i}(x_i|\vartheta_{y_i})) \tag{3.9}$$

που, με δεδομένη μορφή των συνιστωσών πυκνότητας, μπορεί να βελτιστοποιηθεί με μια ποικιλία τεχνικών.

Το πρόβλημα, βέβαια, είναι μη γνώση των τιμών του Y. Θεωρώντας όμως το Y ως τυχαίο διάνυσμα μπορεί να συνεχιστεί η ανάλυση.

Πρώτα θα πρέπει να παραχθεί μία έκφραση για την κατανομή των μη παρατηρηθέντων δεδομένων. Για αρχή χρησιμοποιείται μία εικασία των παραμέτρων της πυκνότητας μίξης,δηλαδή, υποθέτεται πως $\Theta^g = (a_1^g, \ldots, a_M^g, \vartheta_1^g, \ldots, \vartheta_M^g)$ είναι οι σωστές παραμέτροι για την πιθανότητα $L(\Theta^g|X, Y)$. Με δεδομένο το Θ^g, μπορεί εύκολα να υπολογιστεί το $p_j(x_i|\vartheta_j^g)$ για κάθε i και j. Επιπλέον, οι παράμετροι μίξης, a_j μπορούν να θεωρηθούν ως εκ των προτέρων πιθανότητες (prior) για κάθε συνιστώσα μίξης ή, αλλιώς, $a_j = p(\text{συνιστώσα} j)$. Έτσι, χρήση κανόνα του Bayes υπολογίζεται:

$$p(y_i|x_i, \Theta^g) = \frac{a_{y_i}^g p_{y_i}(x_i|\vartheta_{y_i}^g)}{p(x_i|\Theta^g)} = \frac{a_{y_i}^g p_{y_i}(x_i|\vartheta_{y_i}^g)}{\sum_{k=1}^{M} a_k^g p_k(x_i|\vartheta_k^g)} \tag{3.10}$$

και

$$p(y|X,\Theta^g) = \prod_{i=1}^{N} p(y_i|x_i,\Theta^g) \tag{3.11}$$

όπου $y = (y_1,\ldots,y_N)$ είναι μία περίσταση των μη παρατηρημένων δεδομένων τραβηγμένη ανεξάρτητα. Μια δεύτερη ματιά στη εξίσωση (3.5), δείχνει ότι στην περίπτωση αυτή αποκτήθηκε η επιθυμητή οριακή πυκνότητα χάρη στην υπόθεση της ύπαρξης κρυφών μεταβλητών και με μία εικασία για τις αρχικές παραμέτρους της κατανομής τους.

Στη περίπτωση αυτή, η εξίσωση (3.5) παίρνει την μορφή:

$$\begin{aligned}
Q(\Theta,\Theta^g) &= \sum_{y\in Y} \log L(\Theta|X,y)p(y|X,\Theta^g) \\
&= \sum_{y\in Y}\sum_{i=1}^{N} \log(a_{y_i}p_{y_i}(x_i|\partial_{y_i})) \prod_{j=1}^{N} p(y_j|x_j,\Theta^g) \\
&= \sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\sum_{i=1}^{N} \log(a_{y_i}p_{y_i}(x_i|\partial_{y_i})) \prod_{j=1}^{N} p(y_j|x_j,\Theta^g) \\
&= \sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M}\sum_{i=1}^{N}\sum_{\ell=1}^{M} \delta_{\ell,y_i}\log(a_\ell p_\ell(x_i|\partial_\ell)) \prod_{j=1}^{N} p(y_j|x_j,\Theta^g) \\
&= \sum_{l=1}^{M}\sum_{i=1}^{N} \log(a_\ell p_\ell(x_i|\partial_\ell)) \sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M} \delta_{\ell,y_i} \prod_{j=1}^{N} p(y_j|x_j,\Theta^g).
\end{aligned} \tag{3.12}$$

η παραπάνω μορφή μπορεί να απλοποιηθεί σημαντικά. Σημειώνεται πρώτα πως για $\ell \in 1,\ldots,M$, ισχύει:

$$\begin{aligned}
&\sum_{y_1=1}^{M}\sum_{y_2=1}^{M}\cdots\sum_{y_N=1}^{M} \delta_{\ell,y_i} \prod_{j=1}^{N} p(y_j|x_j,\Theta^g) \\
&= \left(\sum_{y_1=1}^{M}\cdots\sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M}\cdots\sum_{y_N=1}^{M}\prod_{j=1,j\neq i}^{N} p(y_j|x_j,\Theta^g)\right) p(\ell|x_i,\Theta^g) \\
&= \prod_{j=1,j\neq i}^{N}\left(\sum_{y_j=1}^{M} p(y_j|x_j,\Theta^g)\right) p(\ell|x_i,\Theta^g) = p(\ell|x_i,\Theta^g),
\end{aligned} \tag{3.13}$$

αφού $\sum_{i=1}^{M} p(i|x_j,\Theta^g) = 1$. Χρησιμοποιώντας την εξίσωση (3.13), η (3.12) μπορεί να γραφτεί ως:

$$\begin{aligned}
Q(\Theta,\Theta^g) &= \sum_{\ell=1}^{M}\sum_{i=1}^{N} \log(a_\ell p_\ell(x_i|\partial_\ell))p(\ell|x_i,\Theta^g) \\
&= \sum_{\ell=1}^{M}\sum_{i=1}^{N} \log a_\ell p(\ell|x_i,\Theta^g) + \sum_{\ell=1}^{M}\sum_{i=1}^{N} \log p_\ell(x_i|\partial_\ell)p(\ell|x_i,\Theta^g).
\end{aligned} \tag{3.14}$$

Για την μεγιστοποίηση της πάνω έκφρασης, αρκεί η μεγιστοποίηση των όρων που περιέχουν το a_ℓ και το ∂_ℓ ανεξάρτητα, μιας και δεν σχετίζονται.

Για την εύρεση της έκφρασης για όλα τα a_ℓ, εισάγεται ο πολλαπλασιαστής Lagrange λ με τον περιορισμό $\sum_\ell a_\ell = 1$, και επιλύεται η παρακάτω εξίσωση:

$$\frac{\partial}{\partial a_\ell}\Big[\sum_{\ell=1}^{M}\sum_{i=1}^{N}\log a_\ell p(\ell|x_i,\Theta^g) + \lambda\Big(\sum_\ell a_\ell - 1\Big)\Big] = 0, \tag{3.15}$$

ή

$$\sum_{i=1}^{N}\frac{1}{a_\ell}p(\ell|x_i,\Theta^g) + \lambda = 0. \tag{3.16}$$

Αθροίζοντας και τις δύο πλευρές πάνω στο ℓ, προκύπτει ότι $\lambda = -N$ με αποτέλεσμα:

$$a_\ell = \frac{1}{N}\sum_{i=1}^{N}p(\ell|x_i,\Theta^g). \tag{3.17}$$

Για κάποιες κατανομές, είναι δυνατών να βρεθεί μια αναλυτική έκφραση των ϑ_ℓ ως συναρτήσεις όλων των υπολοίπων ποσοτήτων. Για παράδειγμα, εάν υποτεθούν κατανομές συνιστωσών Gauss διάστασης d με μέση τιμή μ και πίνακα συνδιασποράς Σ, δηλαδή, $\vartheta = (\mu, \Sigma)$ τότε

$$p_\ell(x|\mu_\ell,\Sigma_\ell) = \frac{1}{(2\pi)^{d/2}|\Sigma_\ell|^{1/2}}e^{-\frac{1}{2}(x-\mu_\ell)^T\Sigma_\ell^{-1}(x-\mu_\ell)}. \tag{3.18}$$

Για την επαγωγή των εξισώσεων ενημέρωσης της κατανομής αυτής χρειάζονται κάποια αποτελέσματα από την άλγεβρα πινάκων.

Το ίχνος ενός τετραγωνικού πίνακα $\mathrm{tr}(A)$ είναι ίσο με το άθροισμα των διαγωνίων στοιχείων του A. Το ίχνος μιας σταθεράς ισούται με την σταθερά. Επίσης, $\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B)$, και $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, πράγμα που υπονοεί ότι $\sum_i x_i^T A x_i = \mathrm{tr}(AB)$ όπου $B = \sum_i x_i X_i^T$. Επίσης με $|A|$ συμβολίζεται η ορίζουσα ενός πίνακα και $|A^{-1}| = 1/|A|$.

Παρακάτω θα χρειαστεί η παράγωγος της συνάρτησης ενός πίνακα $f(A)$ ως προς στοιχεία του πίνακα. Ορίζεται, έτσι, ως $\frac{\partial f(A)}{\partial A}$ ο πίνακας με στοιχείο διάταξης i,j το $[\frac{\partial f(A)}{\partial a_{i,j}}]$, όπου $a_{i,j}$ είναι το στοιχείο i,j του A. Ισχύει $\frac{\partial x^T A x}{\partial x} = (A+A^T)x$. Επίσης, μπορεί να δειχθεί ότι όταν ο A είναι συμμετρικός:

$$\frac{\partial|A|}{\partial a_{i,j}} = \begin{cases} A_{i,j} & \text{αν } i = j \\ 2A_{i,j} & \text{αν } i \neq j \end{cases} \tag{3.19}$$

όπου $A_{i,j}$ είναι ο i,j συμπαράγοντας του A. Δεδομένων των από πάνω, προκύπτει:

$$\frac{\partial\log|A|}{\partial A} = \begin{Bmatrix} A_{i,j}/|A| & \text{αν } i = j \\ 2A_{i,j}/|A| & \text{αν } i \neq j \end{Bmatrix} = 2A^{-1} - \mathrm{diag}(A^{-1}) \tag{3.20}$$

από τον ορισμό του αντιστρόφου πίνακα. Τελικά, μπορεί να δειχθεί ότι:

$$\frac{\partial\,\mathrm{tr}(AB)}{\partial A} = B + B^T - \mathrm{diag}(B). \tag{3.21}$$

Αντικαθιστώντας τον λογάριθμο της εξίσωσης (3.18), έχοντας αγνοήσει οποιοδήποτε σταθερό όρο, στο δεξί μέλος της εξίσωσης (3.14), προκύπτει:

$$\begin{aligned}&\sum_{\ell=1}^{M}\sum_{i=1}^{N}\log(p_\ell(x_i|\mu_\ell,\Sigma_\ell))p(\ell|x_i,\Theta^g)\\&=\sum_{\ell=1}^{M}\sum_{i=1}^{N}\Big(-\frac{1}{2}\log|\Sigma_\ell|-\frac{1}{2}(x_i-\mu_\ell)^T\Sigma_\ell^{-1}(x_i-\mu_\ell)\Big)p(\ell|x_i,\Theta^g).\end{aligned}\tag{3.22}$$

Η παράγωγος της (3.22) πάνω στο μ_ℓ εξισωμένη με το μηδέν, δίνει:

$$\sum_{i=1}^{N}\Sigma_\ell^{-1}(x_i-\mu_\ell)p(\ell|x_i,\Theta^g)=0,\tag{3.23}$$

που μπορεί να λυθεί εύκολα προς μ_ℓ δίνοντας:

$$\mu_\ell=\frac{\sum_{i=1}^{N}x_ip(\ell|x_i,\Theta^g)}{\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)}.\tag{3.24}$$

Για την εύρεση του Σ_ℓ, σημειώνεται ότι η (3.22) μπορεί να γραφτεί ως:

$$\begin{aligned}&\sum_{\ell=1}^{M}\Big[\frac{1}{2}\log(|\Sigma_\ell^{-1}|)\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)-\frac{1}{2}\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)\,\mathrm{tr}(\Sigma_\ell^{-1}(x_i-\mu_\ell)(x_i-\mu_\ell)^T)\Big]\\&=\sum_{\ell=1}^{M}\Big[\frac{1}{2}\log(|\Sigma_\ell^{-1}|)\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)-\frac{1}{2}\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)\,\mathrm{tr}(\Sigma_\ell^{-1}N_{\ell,i})\Big]\end{aligned}\tag{3.25}$$

όπου $N_{\ell,i}=(x_i-\mu_\ell)(x_i-\mu_\ell)^T$).

Παίρνοντας την παράγωγο πάνω στο Σ_ℓ^{-1}, προκύπτει:

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)(2\Sigma_\ell-\mathrm{diag}(\Sigma_\ell))-\frac{1}{2}\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)(2N_{\ell,i}-\mathrm{diag}(N_{\ell,i}))\\&=\frac{1}{2}\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)(2M_{\ell,i}-\mathrm{diag}(M_{\ell,i})\\&=2S-\mathrm{diag}(S),\end{aligned}\tag{3.26}$$

όπου $M_{\ell,i}=\Sigma_\ell-N_{\ell,i}$ και $S=\frac{1}{2}\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)M_{\ell,i}$. Η εξίσωση της παραγώγου με το μηδέν, $2S-\mathrm{diag}(S)=0$, δίνει $S=0$. Οπότε

$$\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)(\Sigma_\ell-N_{\ell,i})=0\tag{3.27}$$

ή

$$\Sigma_\ell=\frac{\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)N_{\ell,i}}{\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)}=\frac{\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)(x_i-\mu_\ell)(x_i-\mu_\ell)^T}{\sum_{i=1}^{N}p(\ell|x_i,\Theta^g)}.\tag{3.28}$$

Τελικά, οι εκτιμήσεις των νέων παραμέτρων βάση των παλιών είναι:

$$a_\ell^{new} = \frac{1}{N}\sum_{i=1}^{N} p(\ell|x_i, \Theta^g) \tag{3.29}$$

$$\mu_\ell^{new} = \frac{\sum_{i=1}^{N} x_i p(\ell|x_i, \Theta^g)}{\sum_{i=1}^{N} p(\ell|x_i, \Theta^g)} \tag{3.30}$$

$$\Sigma_\ell^{new} = \frac{\sum_{i=1}^{N} p(\ell|x_i, \Theta^g)(x_i - \mu_\ell^{new})(x_i - \mu_\ell^{new})^T}{\sum_{i=1}^{N} p(\ell|x_i, \Theta^g)}. \tag{3.31}$$

Σημειώνεται πως οι παραπάνω εξισώσεις εκτελούν τόσο το βήμα προσδοκίας, όσο και το βήμα μεγιστοποίησης ταυτόχρονα. Ο αλγόριθμος συνεχίζει χρησιμοποιώντας τις πιο πρόσφατες παραμέτρους ως εικασία για την επόμενη επανάληψη.

3.4 Μαθαίνοντας τις παραμέτρους ενός Κρυφού μοντέλου Markov

Ένα Κρυφό Μοντέλο Markov (Hidden Markov Model, HMM) είναι ένα πιθανοτικό μοντέλο της κοινής πιθανότητας μιας συλλογής τυχαίων μεταβλητών $\{O_1, \ldots, O_T, Q_1, \ldots, Q_T\}$. Οι μεταβλητές O_t είναι είτε συνεχείς είτε διακριτές παρατηρήσεις και οι μεταβλητές Q_t είναι 'κρυφές' και διακριτές. Σε ένα HMM, υπάρχουν δύο υποθέσεις υπό όρους ανεξαρτησίας των μεταβλητών που μαζί καθιστούν το πρόβλημα ρεαλιστικά επιλύσιμο. Οι υποθέσεις αυτές είναι:

1. Η κρυφή μεταβλητή t, δοθέντος της $t-1$ κρυφής μεταβλητής, είναι ανεξάρτητη των προηγούμενων μεταβλητών, ή

$$P(Q_t|Q_{t-1}, O_{t-1}, \ldots, Q_1, O_1) = P(Q_t|Q_{t-1}). \tag{3.32}$$

2. Η παρατήρηση t, δοθέντος της t κρυφής μεταβλητής, είναι ανεξάρτητη των υπολοίπων μεταβλητών, ή

$$\begin{aligned} &P(O_t|Q_T, O_T, Q_{T-1}, O_{T-1}, \ldots, Q_{t+1}, O_{t+1}, Q_t, Q_{t-1}, O_{t-1}, \ldots, Q_1, O_1) \\ &= P(O_t|Q_t). \end{aligned} \tag{3.33}$$

Στην παράγραφο αυτή θα παραχθεί ο αλγόριθμος ΕΜ για την εύρεση των εκτιμήσεων μέγιστης πιθανότητας των παραμέτρων ενός κρυφού μοντέλου Markov δεδομένου ενός συνόλου διανυσμάτων παρατηρήσεων. Ο αλγόριθμος αυτός είναι επίσης γνωστός ως αλγόριθμος Baum-Welch.

Έστω Q_t μια διακριτή τυχαία μεταβλητή με N πιθανές τιμές $\{1, \ldots, N\}$. Επιπλέον υποθέτεται πως η 'κρυφή' αλυσίδα Markov όπως ορίζεται από την $P(Q_t|Q_{t-1})$ είναι χρονικά ομογενής (δλδ, ανεξάρτητη του χρόνου t). Για αυτό είναι δυνατή η αντιπροσώπευση του $P(Q_t|Q_{t-1})$ ως έναν χρονικά ανεξάρτητο πίνακα στοχαστικών μεταβάσεων $A = \{a_{i,j}\} = p(Q_t = j|Q_{t-1} = i)$. Η ειδική περίπτωση του χρόνου $t = 1$ περιγράφεται από την αρχική κατανομή κατάστασης, $\pi_i = p(Q_1 = i)$. Θα λέγεται ότι είμαστε στην κατάσταση j το χρόνο t αν $Q_t = j$. Μια συγκεκριμένη αλληλουχία

καταστάσεων περιγράφεται από το $q = (q_1, \ldots, q_T)$ όπου $q_t \in \{1, \ldots, N\}$ είναι η κατάσταση το χρόνο t.

Μια συγκεκριμένη αλληλουχία παρατηρήσεων O περιγράφεται ως $O = (O_1 = o_1, \ldots, O_T = o_T)$. Η πιθανότητα ενός συγκεκριμένου διανύσματος παρατηρήσεων στο χρόνο t για κατάσταση j περιγράφεται από το $b_j(o_t) = p(O_t = o_t|Q_t = j)$. Η πλήρης συλλογή παραμέτρων για όλες τις κατανομές παρατηρήσεων συμβολίζεται $B = \{b_j(\cdot)\}$.

Υπάρχουν δύο μορφές κατανομών εξόδου που θα θεωρηθούν. Η πρώτη υποθέτει διακριτές παρατηρήσεις όπου μια παρατήρηση είναι μία εκ των L δυνατών συμβόλων παρατήρησης $o_t \in V = \{v_1, \ldots, v_L\}$. Στη περίπτωση αυτή, αν $o_t = v_k$, τότε $b_j(o_t) = p(O_t = v_k|q_t = j)$. Η δεύτερη μορφή κατανομής πιθανοτήτων που υποθέτεται είναι μία μίξη M πολυμεταβλητών κατανομών Gauss για κάθε κατάσταση όπου $b_j(o_t) = \sum_{\ell=1}^{M} c_{j\ell} N(o_t|\mu_{j\ell}, \Sigma_{j\ell}) = \sum_{\ell=1}^{M} c_{j\ell} b_{j\ell}(o_t)$.

Το πλήρες σύνολο των παραμέτρων HMM για ένα δεδομένο μοντέλο δίνεται από το $\lambda = (A, B, \pi)$. Υπάρχουν τρία βασικά προβλήματα σχετιζόμενα με τα HMMs:

1. Εύρεση της $p(O|\lambda)$ για κάποιο $O = (o_1, \ldots, o_T)$. Χρησιμοποιείται η διαδικασία προς τα μπρος, ή προς τα πίσω, μιας και είναι πολύ πιο αποτελεσματική από τον απευθείας υπολογισμό.

2. Δεδομένου κάποιων O και λ, η εύρεση της καλύτερης ακολουθίας $q = (q_1, \ldots, q_T)$ που εξηγεί το O. Ο αλγόριθμος **Viterbi** επιλύει αυτό το πρόβλημα.

3. Εύρεση του $\lambda^* = \arg\max_{\lambda} p(O|\lambda)$. Ο αλγόριθμος **Baum-Welch** (που επίσης αποκαλείται αλγόριθμος μπρος-πίσω ή EM για HMMs) επιλύει αυτό το πρόβλημα.

Τα προβλήματα αυτά θα αναλυθούν στις παρακάτω ενότητες.

3.4.1 Αποτελεσματικός Υπολογισμός Επιθυμητών Παραμέτρων

Ένα από τα πλεονεκτήματα των HMM είναι ότι σχετικά αποτελεσματικοί αλγόριθμοι μπορούν να παραχθούν για τα τρία προβλήματα που παραπάνω αναφέρθηκαν. Πριν την επαγωγή του αλγορίθμου EM από την απευθείας χρήση της συνάρτησης Q, θα αναλυθούν οι αποτελεσματικές αυτές διαδικασίες.

Πρώτα θα αναλυθεί η διαδικασία προς τα μπρος. Ορίζεται

$$a_i(t) = p(O_1 = o_1, \ldots, O_t = o_t, Q_t = i|\lambda) \tag{3.34}$$

που είναι η πιθανότητα παρατήρησης της μερικής ακολουθίας $o_1, \ldots, o_t$ και κατάληξης στην κατάσταση i τον χρόνο t. Το $a_i(t)$ μπορεί να οριστεί επαναληπτικά ως:

1. $a_i(1) = \pi_i b_i(o_1)$ (3.35)

2. $a_j(t+1) = [\sum_{i=1}^{N} a_i(t)a_{ij}]b_j(o_{t+1})$ (3.36)

3. $p(O|\lambda) = \sum_{i=1}^{N} a_i(T)$ (3.37)

Η προς τα πίσω διαδικασία είναι παρόμοια. Ορίζεται

$$\beta_i(t) = p(O_{t+1} = o_{t+1}, \ldots, O_T = o_T | Q_t = i, \lambda) \tag{3.38}$$

που είναι η πιθανότητα ολοκλήρωσης της ακολουθίας με $o_1, \ldots, o_t$ δεδομένου ότι ξεκίνησε στην κατάσταση i τον χρόνο t. Το $\beta_i(t)$ μπορεί να οριστεί επαναληπτικά ως:

1. $\beta_i(T) = 1$ (3.39)

2. $\beta_i(t) = \sum_{j=1}^{N} a_{ij}b_j(o_{t+1})\beta_j(t+1)$ (3.40)

3. $p(O|\lambda) = \sum_{i=1}^{N} \beta_i(1)\pi_i b_i(o_1)$ (3.41)

Ορίζεται

$$\gamma_i(t) = p(Q_t = i|O, \lambda), \tag{3.42}$$

που είναι η πιθανότητα της κατάστασης i τον χρόνο t για την ακολουθία O. Σημειώνεται ότι:

$$p(Q_t = i|O, \lambda) = \frac{p(O, Q_t = i|\lambda)}{P(O|\lambda} = \frac{p(O, Q_t = i|\lambda)}{\sum_{j=1}^{N} p(O, Q_t = j|\lambda)}. \tag{3.43}$$

Επίσης, λόγω της υπό συνθήκη ανεξαρτησίας Markov ισχύει

$$a_i(t)\beta_i(t) = p(o_1, \ldots, o_t, Q_t = i|\lambda)p(o_{t+1}, \ldots, o_T|Q_t = i, \lambda) = p(O, Q_t = i|\lambda), \tag{3.44}$$

και μπορεί να οριστεί το $\gamma_i(t)$ έτσι σε όρους $a_i(t)$ και $\beta_i(t)$ ως

$$\gamma_i(t) = \frac{a_i(t)\beta_i(t)}{\sum_{j=1}^{N} a_j(t)\beta_j(t)}. \tag{3.45}$$

Ορίζεται επίσης

$$\xi_{ij}(t) = p(Q_t = i, Q_{t+1} = j|O, \lambda) \tag{3.46}$$

που είναι η πιθανότητα κατάστασης i στο χρόνο t και κατάστασης j στο χρόνο $t+1$. Αυτό μπορεί επίσης να προεκταθεί στο:

$$\xi_{ij}(t) = \frac{p(Q_t = i, Q_{t+1} = j, O|\lambda)}{p(O|\lambda)} = \frac{a_i(t)a_{ij}b_j(o_{t+1})\beta_j(t+1)}{\sum_{i=1}^{N}\sum_{j=1}^{N} a_i(t)a_{ij}b_j(o_{t+1})\beta_j(t+1)} \tag{3.47}$$

ή στο:

$$\xi_{ij}(t) = \frac{p(Q_t = i|O)p(o_{t+1} \ldots o_T, Q_{t+1} = j|Q_t = i, \lambda)}{p(o_{t+1} \ldots o_T|Q_t = i, \lambda)} = \frac{\gamma_i(t)a_{ij}b_j(o_{t+1})\beta_j(t+1)}{\beta_i(t)}. \tag{3.48}$$

Αθροίζωντας τις ποσότητες αυτές πάνω στο χρόνο προκύπτουν κάποιες χρήσιμες εκφράσεις. Έτσι,

$$\sum_{t=1}^{T} \gamma_i(t) \tag{3.49}$$

είναι ο αναμενόμενος αριθμός επισκέψεων της κατάστασης i και, συνεπώς, ο αναμενόμενος αριθμός μεταβάσεων από την κατάσταση i για το O. Παρόμοια,

$$\sum_{t=1}^{T-1} \xi_{ij}(t) \tag{3.50}$$

είναι ο αναμενόμενος αριθμός μεταβάσεων από την κατάσταση i στην κατάσταση j για το O. Αυτά προκύπτουν από το γεγονός ότι

$$\sum_t \gamma_i(t) = \sum_t E\{I_t(i)\} = E\Big\{\sum_t I_t(i)\Big\} \tag{3.51}$$

και

$$\sum_t \xi_{ij}(t) = \sum_t E\{I_t(i,j)\} = E\Big\{\sum_t I_t(i,j)\Big\} \tag{3.52}$$

όπου $I_t(i)$ είναι μια δυαδική τυχαία μεταβλητή δείκτης με τιμή 1 όταν είμαστε στην κατάσταση i τον χρόνο t, και $I_t(i,j)$ είναι μια δυαδική τυχαία μεταβλητή δείκτης με τιμή 1 όταν γίνεται μετάβαση από την κατάσταση i στην κατάσταση j μετά το χρόνο t.

Ζητείται ο σχηματισμός ενός αλγορίθμου ΕΜ για την εκτίμηση νέων παραμέτρων για το HMM από τις παλιές παραμέτρους και τα δεδομένα. Διαισθητικά, αυτό μπορεί να γίνει με τη χρήση σχετικών συχνοτήτων. Μπορούν να οριστούν, δηλαδή, οι κανόνες ανανέωσης όπως παρακάτω:

Η ποσότητα

$$\tilde{\pi}_i = \gamma_i(1) \tag{3.53}$$

είναι η αναμενόμενη σχετική συχνότητα που καταβάλλεται στην κατάσταση i την χρονική στιγμή 1.

Η ποσότητα

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_{ij}(t)} \tag{3.54}$$

είναι ο αναμενόμενος αριθμός μεταβάσεων από την κατάσταση i στη κατάσταση j προς με τον αναμενόμενο συνολικό αριθμό μεταβάσεων από την κατάσταση i.

Και, για διακριτές κατανομές, η ποσότητα

$$\tilde{b}_i(k) = \frac{\sum_{t=1}^{T} \delta_{o_t,v_k} \gamma_i(t)}{\sum_{t=1}^{T} \gamma_i(t)} \tag{3.55}$$

είναι ο αναμενόμενος αριθμός των φορών που οι παρατηρήσεις εξόδου θα είναι ίσες με v_k στην κατάσταση i προς τον αναμενόμενο συνολικό αριθμό φορών στην κατάσταση i.

Για μίξεις Gauss, ορίζεται η πιθανότητα η συνιστώσα ℓ της i μίξης να παρήγαγε την παρατήρηση o_t ως

$$\gamma_{i\ell}(t) = \gamma_i(t)\frac{c_{i\ell} b_{i\ell}(o_t)}{b_i(o_t)} = p(Q_t = i, X_{it} = \ell | O, \lambda) \tag{3.56}$$

Όπου X_{it} είναι μια τυχαία μεταβλητή με τιμή την συνιστώσα μίξης στο χρόνο t για την κατάσταση i.

Οι εξισώσεις ανανέωσης για την περίπτωση αυτή είναι:

$$c_{i\ell} = \frac{\sum_{t=1}^{T} \gamma_{i\ell}(t)}{\sum_{t=1}^{T} \gamma_i(t)} \tag{3.57}$$

$$\mu_{i\ell} = \frac{\sum_{t=1}^{T} \gamma_{i\ell}(t) o_t}{\sum_{t=1}^{T} \gamma_{i\ell}(t)} \tag{3.58}$$

$$\Sigma_{i\ell} = \frac{\sum_{t=1}^{T} \gamma_{i\ell}(t)(o_t - \mu_{i\ell})(o_t - \mu_{i\ell})^T}{\sum_{t=1}^{T} \gamma_{i\ell}(t)} \tag{3.59}$$

Όταν υπάρχουν E ακολουθίες παρατηρήσεων με την e να έχει μήκος T_e, οι εξισώσει ανανέωσης γίνονται:

$$\pi_i = \frac{\sum_{e=1}^{E} \gamma_i^e(1)}{E} \tag{3.60}$$

$$c_{i\ell} = \frac{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{i\ell}^e(t)}{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_i^e(t)} \tag{3.61}$$

$$\mu_{i\ell} = \frac{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{i\ell}^e(t) o_t^e}{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{i\ell}^e(t)} \tag{3.62}$$

$$\Sigma_{i\ell} = \frac{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{i\ell}^e(t)(o_t^e \dot{\mu}_{i\ell})(o_t^e \dot{\mu}_{i\ell})^T}{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{i\ell}^e(t)} \tag{3.63}$$

και

$$a_{ij} = \frac{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \xi_{ij}^e(t)}{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_i^e(t)} \tag{3.64}$$

Οι εξισώσεις αυτές είναι πράγματι ο αλγόριθμος EM ή (Baum-Welch) για την εκτίμηση παραμέτρων HMM. Η επόμενη παράγραφος θα καταλήξει στους ίδιους τύπους χρησιμοποιώντας τους πιο τυπικούς για τον EM συμβολισμούς.

3.4.2 Εξισώσεις Ενημέρωσης με χρήση της συνάρτησης Q

Έστω $O = (o_1, \ldots, o_T)$ τα παρατηρούμενα δεδομένα ενώ η ακολουθία καταστάσεων $q = (q1, \ldots, q_T)$ θεωρήται κρυφή ή μη παρατηρήσιμη. Η συνάρτητη πιθανότητας των ημιτελών δεδομένων δίνεται από το $P(O, \lambda)$ ενώ η συνάρτηση πιθανότητας πλήρη δεδομένων είναι $P(O, q|\lambda)$. Η συνάρτηση Q είναι έτσι:

$$Q(\lambda, \lambda') = \sum_{q \in Q} \log P(O, q|\lambda) P(O, q|\lambda') \tag{3.65}$$

όπου λ' είναι οι αρχικές μας εκτιμήσεις των παραμέτρων και Q είναι ο χώρος όλων των ακολουθιών καταστάσεων μήκους T.

Δοθέντης συγκεκριμένης ακολουθίας καταστάσεων q, ο τύπος του $P(O, q|\lambda')$ είναι

$$P(O, q|\lambda) = \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(o_t) \tag{3.66}$$

Σημειώνεται πως εδώ υποθέτεται ότι η αρχική κατανομή ξεκινά στο $t = 0$ αντί για $t = 1$ γαι ευκολότερη σήμανση. Επίσης, στο υπόλοιπο κείμενο, οι τονισμένες παράμετροι θεωρούνται οι αρχικές, ενώ οι μη τονισμένες παράμετροι είναι που βελτιστοποιούνται.

Η συνάρτηση Q γίνεται:

$$Q(\lambda, \lambda') = \sum_{q \in Q} \log \pi_{q_0} P(O, q|\lambda') + \sum_{q \in Q} \Big(\sum_{t=1}^{T} \log a_{q_{t-1} q_t} \Big) p(O, q|\lambda') + \sum_{q \in Q} \Big(\sum_{t+1}^{T} \log b_{q_t}(o_t) \Big) P(O, q|\lambda'). \tag{3.67}$$

Δεδομένου ότι οι παράμετροι προς βελτιστοποίηση είναι χωρισμένοι ανεξάρτητα δύναται η βελτιστοποίηση κάθε όρου χωριστά.

Ο πρώτος όρος της εξίσωσης (3.67) γίνεται

$$\sum_{q \in Q} \log \pi_{q_0} P(O, q|\lambda') = \sum_{i=1}^{N} \log \pi_i p(O, q_0 = i|\lambda') \tag{3.68}$$

αφού η επιλογή όλων των $q \in Q$ αποτελεί απλά μια επαναλαμβανόμενη επιλογή των τιμών το q_0, οπότε το δεξί μέλος είναι απλά η οριακή έκφραση για χρόνο $t = 0$. Προσθέτοντας τον πολλαπλασιαστή Lagrange, με το περιορισμό $\sum_i \pi_i = 1$, και εξισώνοντας την παράγωγο με το μηδέν, προκύπτει:

$$\frac{\partial}{\partial \pi_i} \Big(\sum_{i=1}^{N} \log \pi_i p(O, q_0 = i|\lambda') + \gamma (\sum_{i=1}^{N} \pi_i - 1) \Big) = 0 \tag{3.69}$$

Παίρνοντας την παράγωγο, προσθέτοντας πάνω στο i για το γ, και επιλύοντας για π_i προκύπτει:

$$\pi_i = \frac{P(O, q_0 = i|\lambda')}{P(O|\lambda')} \tag{3.70}$$

Ο δεύτερος όρος της (3.67) γίνεται:

$$\sum_{q \in Q} \Big(\sum_{t=1}^{T} \log a_{q_{t-1} q_t} \Big) p(O, q|\lambda') = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \log a_{ij} P(O, q_{t-1} = i, q_t = j|\lambda') \tag{3.71}$$

επειδή για τον όρο αυτό, σε κάθε χρόνο t εντοπίζονται όλες οι μεταβάσεις από το i στο j και σταθμίζονται από την αντίστοιχη πιθανότητα. Το δεξί μέλος είναι απλά το άθροισμα της οριακής κοινής πιθανότητας για χρόνους $t-1$ και t. Παρομοίως, μπορεί να χρησιμοποιηθεί ένας πολλαπλασιαστής Lagrange με τον περιορισμό $\sum_{j=1}^{N} a_{ij} = 1$ για να προκύψει:

$$a_{ij} = \frac{\sum_{t=1}^{T} P(O, q_{t-1} = i, q_t = j|\lambda')}{\sum_{t=1}^{T} P(O, q_{t-1} = i|\lambda')}. \tag{3.72}$$

Ο τρίτος όρος της εξίσωσης (3.67) γίνεται:

$$\sum_{q \in Q} \Big(\sum_{t+1}^{T} \log b_{q_t}(o_t) \Big) P(O, q|\lambda') = \sum_{i=1}^{N} \sum_{t=1}^{T} \log b_i(o_t) p(O, q_t = i|\lambda') \tag{3.73}$$

επειδή για τον όρο αυτό, σε κάθε χρόνο t, εντοπίζονται οι εκροές από όλες τις καταστάσεις και σταθμίζονται από την αντίστοιχη τους πιθανότητα. ο δεξί μέλος είναι απλά το άθροισμα των οριακών πιθανοτήτων για χρόνο t.

Για διακριτές κατανομές γίνεται, και πάλι, να χρησιμοποιηθεί ένας πολλαπλασιαστής Lagrange αλλά αυτή τη φορά με τον περιορισμό $\sum_{j=1}^{L} b_i(j) = 1$. Μόνο οι παρατηρήσεις που είναι ίσες με v_k συνεισφέρουν στην τιμή της πιθανότητας k, οπότε:

$$b_i(k) = \frac{\sum_{t=1}^{T} P(O, q_t = i|\lambda')\delta_{o_t,v_k}}{\sum_{t=1}^{T} P(O, q_t = i|\lambda')} \tag{3.74}$$

Για μίξεις Gauss, η μορφή της συνάρτησης Q είναι λίγο διαφορετική, δηλαδή, οι κρυφές μεταβλητές πρέπει να περιλαμβάνουν όχι μόνο την ακολουθία κρυφών καταστάσεων, αλλά και μία μεταβλητή που να δείχνει το στοιχείο μίξης για κάθε κατάσταση την κάθε στιγμή. Έτσι, η συνάρτηση Q μπορεί να γραφτεί:

$$Q(\lambda, \lambda') = \sum_{q \in Q} \sum_{m \in (M)} \log P(O, q, m|\lambda) P(O, q, m|\lambda') \tag{3.75}$$

όπου m είναι το διάνυσμα $m = \{m_{q_1 1}, m_{q_2 2}, \ldots, m_{q_T T}\}$ που δείχνει το στοιχείο μίξης για κάθε κατάσταση στον κάθε χρόνο. Εάν επεκταθεί αυτή η σχέση όπως η (3.67), ο πρώτος και δεύτερος όρος θα παραμείνουν αμετάβλητοι, αφού οι παράμετροι είναι ανεξάρτητοι του m και έτσι εξαφανίζονται πάνω στο άθροισμα. Ο τρίτος όρος της (3.67) γίνεται:

$$\begin{aligned} &\sum_{q \in Q} \sum_{m \in \mathcal{M}} \Big(\sum_{t=1}^{T} \log b_{q_t}(o_t, m_{q_t t}) \Big) P(O, q, m|\lambda') = \\ &\sum_{i=1}^{N} \sum_{\ell=1}^{M} \sum_{t=1}^{T} \log(c_{i\ell} b_{i\ell}(o_t)) p(O, q_t = i, m_{q_t t} = \ell|\lambda') \end{aligned} \tag{3.76}$$

Η εξίσωση αυτή είναι σχεδόν ίδια με την (3.14), με εξαίρεση τον επιπλέον όρο αθροίσματος πάνω στις κρυφές μεταβλητές κατάστασης. Μπορεί να βελτιστοποιηθεί ακριβώς ανάλογα με την παράγραφο (3.3), δίνοντας:

$$c_{il} = \frac{\sum_{t=1}^{T} P(q_t = i, m_{q_t t} = l|O, \lambda')}{\sum_{t=1}^{T} \sum_{\ell=1}^{M} P(q_t = i, m_{q_t t} = \ell|O, \lambda')} \tag{3.77}$$

$$\mu_{il} = \frac{\sum_{t=1}^{T} o_t P(q_t = i, m_{q_t t} = l|O, \lambda')}{\sum_{t=1}^{T} P(q_t = i, m_{q_t t} = l|O, \lambda')} \tag{3.78}$$

και

$$\Sigma_{il} = \frac{\sum_{t=1}^{T} (o_t - \mu_{il})(o_t - \mu_{il})^T P(q_t = i, m_{q_t t} = l|O, \lambda')}{\sum_{t=1}^{T} P(q_t = i, m_{q_t t} = l|O, \lambda')} \tag{3.79}$$

Αυτές είναι οι ίδιες εξισώσεις ανανέωσης που βρέθηκαν στην προηγούμενη παράγραφο (3.4.1).

Οι εξισώσεις ανανέωσης για κρυφά μοντέλα Markov με πολλαπλές ακολουθίες παρατηρήσεων μπορούν να αναχθούν παρόμοια (Rabiner and Juang (1993)).

3.5 Εύρεση Βέλτιστης Τροχιάς Καταστάσεων

Στην παράγραφο αυτή εξετάζεται το πρόβλημα εύρεσης της βέλτιστης τροχιάς καταστάσεων. Σε αντίθεση με τα προηγούμενα προβλήματα που είχαν, γενικά, μία αναλυτική λύση, υπάρχουν αρκετοί τρόποι επίλυσης του προβλήματος αυτού. Το πρόβλημα εντοπίζεται στην έννοια της βέλτιστης τροχιάς, υπάρχουν ,δηλαδή, αρκετά κριτήρια βελτιστότητας. Έτσι, π.χ., ένα κριτήριο βελτιστότητας μπορεί να είναι η επιλογή των καταστάσεων q_t που να είναι ατομικά πιο πιθανά σε μια χρονική στιγμή t. Το κριτήριο αυτό μεγιστοποιεί τον αναμενόμενο αριθμό σωστών ατομικών καταστάσεων. Για την εφαρμογή αυτής της λύσης, μπορεί να οριστεί η μεταβλητή της εκ των υστέρων πιθανότητας

$$\gamma_t(i) = P(q_t = i|O, \lambda), \tag{3.80}$$

που είναι η πιθανότητα της κατάστασης i στο χρόνο t, δεδομένης της ακολουθίας παρατηρήσεων O, και του μοντέλου λ. Το $\gamma_t(i)$ μπορεί να εκφραστεί σε πολλές μορφές, μεταξύ των οποίων η

$$\gamma_t(i) = P(q_t = i|O, \lambda) = \frac{P(O, q_t = i|\lambda)}{P(O|\lambda)} = \frac{P(O, q_t = i|\lambda)}{\sum_{j=1}^{N} P(O, q_t = j|\lambda)}. \tag{3.81}$$

Αφού το $P(O, q_t = i|\lambda)$ ισούται με $a_t(i)\beta_t(i)$, το $\gamma_t(i)$ μπορεί να γραφτεί ως

$$\gamma_t(i) = \frac{a_t(i)\beta_t(i)}{\sum_{j=1}^{N} a_t(j)\beta_t(j)}, \tag{3.82}$$

όπου το $a_t(i)$ είναι υπεύθυνο για την ακολουθία παρατηρήσεων ως το χρόνο t, ενώ το $\beta_t(i)$ είναι υπεύθυνο για την υπόλοιπη ακολουθία από τον χρόνο $t + 1$ ως τον T με δεδομένη κατάσταση $q_t = i$ στο χρόνο t.

Χρησιμοποιώντας το $\gamma_t(i)$, μπορεί να γίνει επίλυση προς την πιο πιθανή κατάσταση q_t^* στο χρόνο t

$$q_t^*(i) = \arg\max_i \gamma_t(i), \qquad 1 \leq t \leq T. \tag{3.83}$$

Αν και η (3.83) μεγιστοποιεί τον αναμενόμενο αριθμό σωστών καταστάσεων, μπορεί να υπάρχουν προβλήματα με την τελική ακολουθία. Για παράδειγμα, όταν στο HMM υπάρχουν μεταβάσεις μηδενικής πιθανότητας ($a_{ij} = 0$, για κάποια i, j), η 'βέλτιστη' τροχιά μπορεί να μην είναι εφικτή. Αυτό γιατί η λύση της εξίσωσης (3.83) απλά εντοπίζει την πιο πιθανή κατάσταση σε κάθε χρονική στιγμή, χωρίς να ασχολείται με την πιθανότητα εμφάνισης μιας αλληλουχίας καταστάσεων.

Μια πιθανή λύση σε αυτό το πρόβλημα είναι η αλλαγή του κριτηρίου βελτιστότητας. Για παράδειγμα, θα μπορούσε να γίνει επίλυση προς μία ακολουθία καταστάσεων που να μεγιστοποιεί τον αναμενόμενο αριθμό σωστών ζεύγων μετάβασης (q_t, q_{t+1}), ή τριπλέτας καταστάσεων (q_t, q_{t+1}, q_{t+2}), κ.τ.λ.. Αν και τα κριτήρια αυτά μπορεί να είναι επαρκή για κάποιες εφαρμογές, το πιο συχνά χρησιμοποιημένο κριτήριο είναι η εύρεση της *μοναδικής* καλύτερης ακολουθίας καταστάσεων (τροχιά). Τη μεγιστοποίηση δηλαδή του $P(q|O, \lambda)$, που είναι ισοδύναμη της μεγιστοποίησης του $P(q, O|\lambda)$. Μία έγκυρη τεχνική για την εύρεση της μοναδικής βέλτιστης ακολουθίας καταστάσεων, βασίζεται στον δυναμικό προγραμματισμό, και είναι ο αλγόριθμος Viterbi.

3.5.1 Αλγόριθμος Viterbi

Για την εύρεση της καλύτερης τροχιάς, $q = (q_1 q_2 \dots q_T)$, δεδομένης ακολουθίας παρατηρήσεων $O = (o_1 o_2 \dots o_T)$, θα πρέπει να οριστεί η ποσότητα

$$\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} P(q_1 q_2 \dots q_{t-1}, q_t = i, o_1 o_2 \dots o_t | \lambda), \tag{3.84}$$

όπου $\delta_t(i)$ είναι η μεγαλύτερη πιθανότητα κατά μήκος μιας τροχιάς, το χρόνο t, που είναι υπεύθυνη για τις πρώτες t παρατηρήσεις και τελειώνει στην κατάσταση i. Μέσω επαγωγής προκύπτει

$$\delta_{t+1}(j) = \max_j \delta_t(i) a_{ij} \cdot b_j(o_{t+1}). \tag{3.85}$$

Για την επαγωγή της ακολουθίας καταστάσεων, πρέπει να αποθηκεύεται το στοιχείο που μεγιστοποίησε την (3.85), για κάθε t και j. Έστω ότι η αποθήκευση αυτή γίνεται στο $\psi_t(j)$. Η ολοκληρωμένη διαδικασία για την εύρεση της βέλτιστης τροχιάς μπορεί να τεθεί ως ακολούθως:

1. Αρχικοποίηση

$$\delta_1(i) = \pi_i b_i(o_1), \qquad 1 \le i \le N \tag{3.86}$$

$$\psi_1(i) = 0. \tag{3.87}$$

2. Επαναλήψεις

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(o_t), \qquad \begin{matrix} 2 \le t \le T \\ 1 \le i \le N \end{matrix} \tag{3.88}$$

$$\psi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], \qquad \begin{matrix} 2 \le t \le T \\ 1 \le i \le N \end{matrix} \tag{3.89}$$

3. Τερματισμός

$$P^* = \max_{1 \le i \le N} [\delta_T(i)] \tag{3.90}$$

$$q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)]. \tag{3.91}$$

4. Επανάκτηση ακολουθίας καταστάσεων (τροχιάς)

$$q^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, \dots, 1. \tag{3.92}$$

Πρέπει να σημειωθεί ότι ο αλγόριθμος Viterbi είναι παρόμοιος (εκτός του βήματος επανάκτησης) με εφαρμογές του προς τα μπρος υπολογισμού των εξισώσεων (3.35)-(3.37). Η σημαντικότερη διαφορά είναι στην μεγιστοποίηση πάνω στις προηγούμενες καταστάσεις (3.88), που χρησιμοποιείται στη θέση της διαδικασίας αθροίσματος στην (3.36). Είναι επίσης ξεκάθαρο ότι μία δικτυωτή δομή εφαρμόζει αποτελεσματικά τους υπολογισμούς της διαδικασίας Viterbi.

Κεφάλαιο 4

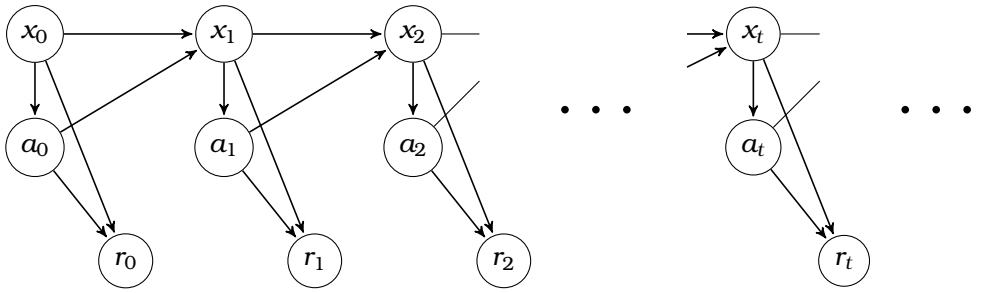
Επίλυση διαδικασιών απόφασης Markov μέσω πιθανοτικού συμπερασμού

4.1 Εισαγωγή

Το 2006, οι Toussaint και Storkey παρουσίασαν την δημοσίευση τους 'Probabilistic Inference for Solving Markov Decision Processes' (Toussaint and Storkey (2006)), βασικό κομμάτι της οποίας ήταν ένας αλγόριθμος βελτιστοποίησης διαδικασιών απόφασης Markov (Markov Decision Process, MDP). Στην δουλειά τους αυτή, κατάφεραν να αποδείξουν σύγκλιση προς το βέλτιστο. Το βασικότερο σημείο της προσέγγισης τους ήταν η παρουσίαση μιας μίξης MDP πεπερασμένου χρόνου ως ένα μοντέλο ισότιμο του αρχικού, αορίστου χρόνου MDP. Αυτό τους επέτρεψε να συνθέσουν μια πιθανότητα, ανάλογη της μελλοντικά αναμενόμενης απόδοσης. Παράλληλα, η εξαγωγή συμπερασμάτων από το μοντέλο μίξης μπορεί να επιτευχθεί με μία συγχρονισμένη διάδοση μηνυμάτων μπρος-πίσω, χωρίς να απαιτείται ο εκ των προτέρων ορισμός του χρόνου τέλους. Παρακάτω θα παρουσιαστεί η δομή αυτή, θα δειχθεί η αναλογία της με την μεγιστοποίηση των αναμενόμενων μελλοντικών αμοιβών και θα περιγραφεί ο αλγόριθμος μεγιστοποίησης αναμονής ΕΜ για τον υπολογισμό βέλτιστων πολιτικών.

4.2 Διαδικασίες Απόφασης Markov (MDPs)

Το σχήμα (4.1) παρουσιάζει το δυναμικό δίκτυο Bayes (Dynamic Bayesian Network, DBN) για ένα MDP αορίστου χρόνου (time-unlimited MDP), το οποίο ορίζεται από την πιθανότητα μεταβάσεων κατάστασης $P(x_{t+1}|a_t, x_t)$, την πιθανότητα επιλογής δράσεων $P(a_t|x_t; \pi)$, και την πιθανότητα αμοιβής $P(r_t|a_t, x_t)$. Οι τυχαίες μεταβλητές κατάστασης, x και δράσης, a, μπορούν να είναι διακριτές ή συνεχείς ενώ η μεταβλητή αμοιβών θεωρείται, χωρίς απώλεια της γενικότητας, ότι είναι δυαδική, $r_t \in \{0, 1\}$. Η υπόθεση αυτή είναι επαρκής για την συσχέτιση αυθαίρετων προσδοκιών αμοιβής $P(r_t|a_t, x_t)$ στο διάστημα $[0, 1]$ με καταστάσεις και δράσεις, πράγμα που είναι, παρ' όλη την αλλαγή κλίμακας, η γενική περίπτωση στα



Σχήμα 4.1: Δυναμικό Δίκτυο Bayes (DBN) για διαδικασία αποφάσεων Markov (MDP). Με x συμβολίζονται οι μεταβλητές κατάστασης, a οι δράσεις και r οι αμοιβές.

σενάρια ενισχυτικής μάθησης. Στο κεφάλαιο αυτό θεωρείται πως οι πιθανότητες μετάβασης και αμοιβής είναι γνωστές εκ των προτέρων.

Η πιθανότητα $P(a_t|x_t;\pi)$ παραμετροποιείται άμεσα από μία άγνωστη πολιτική π έτσι ώστε $P(a_t = a|x_t = i;\pi) = \pi_{ai}$, οι αριθμοί $\pi_{ai} \in [0,1]$ κανονικοιποιούνται βάση του a. Το πρόβλημα είναι η *επίλυση του MDP*, η εύρεση, δηλαδή, μιας πολιτικής π του γραφικού μοντέλου στο Σχήμα (4.1) που να μεγιστοποιεί την προσδοκώμενη μελλοντική απόδοση,

$$V^{\pi}(i) = E\Big\{\sum_{t=0}^{\infty} \gamma^t r_t | x_0 = i;\pi\Big\},$$

όπου $\gamma \in [0,1]$ είναι ο συντελεστής μείωσης. Η κλασική προσέγγιση επίλυσης των MDP στηρίζεται στην εξίσωση του Bellman. Τα παραπάνω αναλύονται στο κεφάλαιο 2, και συγκεκριμένα στην παράγραφο 2.4.

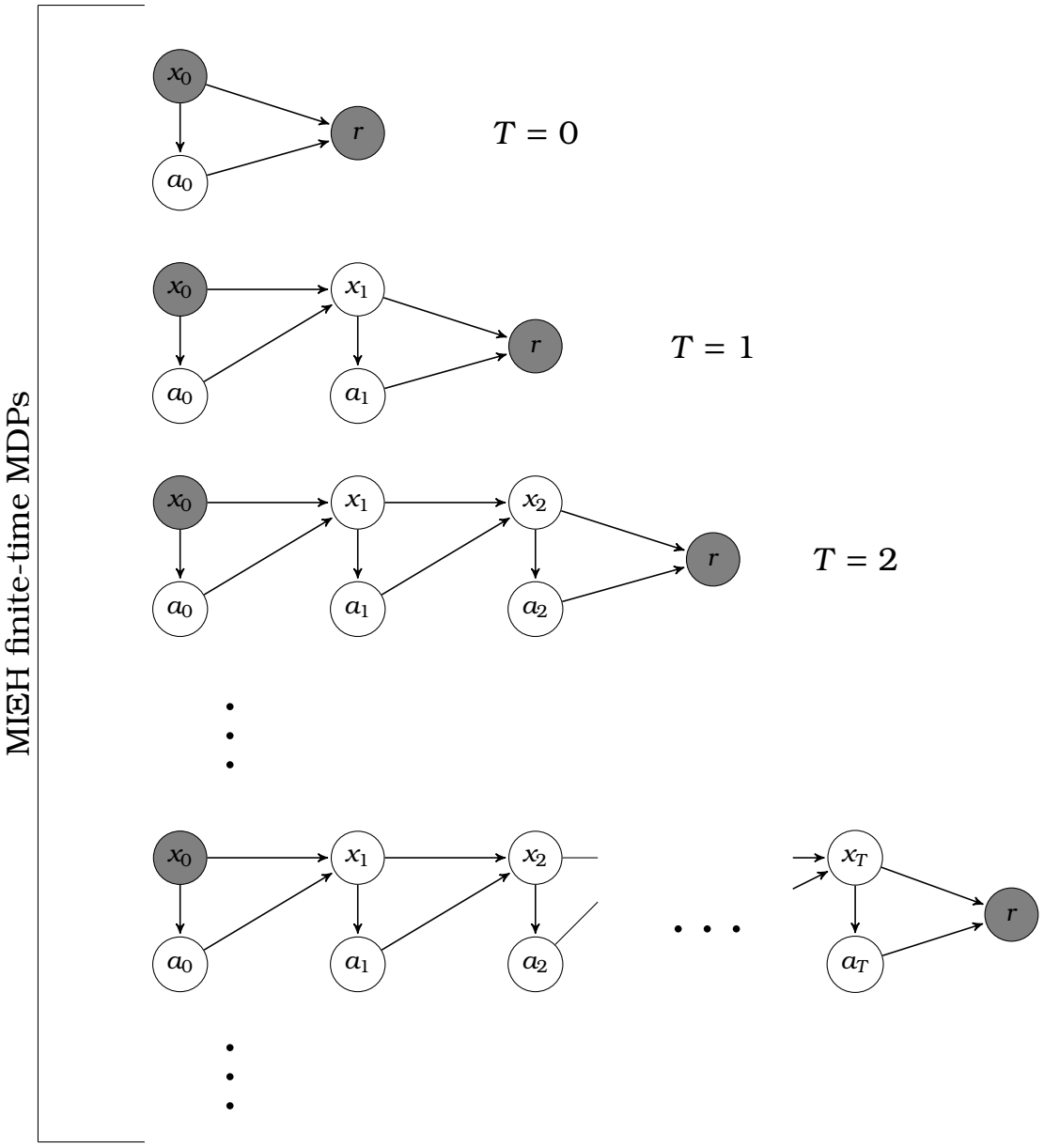
4.3 Μίξη MDPs και πιθανοτήτων

Το πρόβλημα επίλυσης ενός MDP μετατρέπεται σε ένα πρόβλημα βελτιστοποίησης των παραμέτρων ενός γραφικού μοντέλου με πολλές κρυφές μεταβλητές, που δεν είναι άλλες από το σύνολο των μελλοντικών καταστάσεων και δράσεων. Οι μόνες παρατηρήσιμες είναι η τρέχουσα κατάσταση και οι αμοιβές. Για να επιτευχθεί ακριβής αντιστοιχία μεταξύ της επίλυσης του MDP και της τυπικής μεγιστοποίησης πιθανότητας θα πρέπει να ορισθεί μια πιθανότητα ανάλογη της αναμενόμενης μελλοντικής απόδοσης. Ο ευκολότερος τρόπος επίτευξης αυτού είναι με τον ορισμό μιας μίξης πεπερασμένου χρόνου (finite-state) MDPs. Ως πεπερασμένου χρόνου νοείται εδώ ένα MDP με περιορισμένο χρόνο T, που αποδίδει μια μεταβλητή αμοιβής μόνο στο τελευταίο χρονικό βήμα όπως παρουσιάζεται για διάφορα T στο 4.2. Η συνολική κοινή πιθανότητα για ένα πεπερασμένου χρόνου MDP είναι

$$P(r, x_{0:T}, a_{0:T}|T;\pi) = P(r|a_T, x_T)P(a_0|x_0;\pi)P(x_0) \cdot \prod_{t=1}^{T} P(a_t|x_t;\pi)P(x_t|a_{t-1}, x_{t-1}). \tag{4.1}$$

Η πλήρη μίξη των πεπερασμένου χρόνου MDPs δίνεται τότε από την κονή

$$P(r, x_{0:T}, a_{0:T}, T;\pi) = P(r, x_{0:T}, a_{0:T}|T;\pi)P(T). \tag{4.2}$$



Σχήμα 4.2: Μίξη διαδικασιών απόφασης Markov πεπερασμένου χρόνου (finite-time MDPs). Με x συμβολίζονται οι μεταβλητές κατάστασης, a οι δράσεις και r οι αμοιβές.

Εδώ, το $P(T)$ είναι μία εκ των προτέρων πιθανότητα (prior) πάνω στον συνολικό χρόνο, το οποίο και επιλέγουμε ανάλογο της μείωσης, $P(T) = \gamma^T(1-\gamma)$. Πρέπει να τονισθεί πως κάθε MDP πεπερασμένου χρόνου μοιράζεται τις ίδιες πιθανότητες μετάβασης και παραμετροποιείται από την ίδια πολιτική π.

Τώρα θεωρούνται οι αμοιβές r ως παρατηρήσεις, δεσμεύεται πάνω στην αρχική κατάσταση, και ορίζεται η πιθανότητα ενός πεπερασμένου χρόνου MDP,

$$L_T^\pi(i) = P(r = 1 | x_0 = i, T; \pi) = E\{r | x_0 = i, T; \pi\}, \tag{4.3}$$

και για την πλήρη μίξη των MDP,

$$L^{\pi}(i) = P(r = 1|x_0 = i; \pi) = \sum_T P(T)E\{r|x_0 = i, T; \pi\}. \quad (4.4)$$

Η πιθανότητα αυτή, είναι για το μειωμένο prior του χρόνου $P(T) = \gamma^T(1-\gamma)$, ανάλογη της αναμενόμενης μελλοντικής απόδοσης,

$$L^{\pi}(i) = (1-\gamma)V^{\pi}(i). \quad (4.5)$$

Εδώ σημειώνεται πως ο όρος $E\{r|x_0 = i, T; \pi\}$ είναι ακριβώς ο ίδιος με τους όρους $E\{r_t|x_0 = i; \pi\}$ για $t = T$. Αυτό γιατί παίρνουμε την προσδοκία λαμβάνοντας υπόψη ένα πλήρες πιθανοτικό πέρασμα του MDP προς τα εμπρός, από το χρόνο 0 στο χρόνο T με δεδομένη πολιτική π. Όλα τα MDPs μοιράζονται τις ίδιες πιθανότητες μετάβασης.

Αποδείξαμε έτσι ότι

Θεώρημα 4.1 *Η μεγιστοποίηση της πιθανότητας (4.4) στη μίξη πεπερασμένου χρόνου MDPs (4.2) αντιστοιχεί στην επίλυση του αρχικού MDP.*

Πέραν της εξακρίβωσης της πλήρης αντιστοιχίας μεταξύ της μεγιστοποίησης πιθανότητας και του αναμενόμενου μελλοντικού οφέλους, η θεώρηση της μίξης μοντέλων πεπερασμένου χρόνου έχει και ένα δεύτερο πλεονέκτημα: η ιδιότητα πεπερασμένου χρόνου κάθε στοιχείου της μίξης κάνει το E-step στο πλήρες μοντέλο μίξης απλό και αποτελεσματικό, όπως αναλύεται στην παρακάτω ενότητα.

4.4 Ένας αλγόριθμος EM για τον υπολογισμό της βέλτιστης πολιτικής

Η διατύπωση της αντικειμενικής συνάρτησης σε όρους πιθανότητας επιτρέπει την εφαρμογή του EM για την εύρεση των βέλτιστων παραμέτρων, της πολιτικής π δηλαδή, του μοντέλου. Οι μεταβλητές δράσης και κατάστασης, με εξαίρεση της αρχικής x_0, είναι κρυφές μεταβλητές. Το E-step υπολογίζει, για δεδομένη π, εκ των υστέρων πιθανότητες (posteriors) πάνω σε αλληλουχίες κατάστασης-δράσης αλλά και στο T, δεσμευμένα πάνω στα $x_0 = A$ και $r = 1$. Το M-step τότε προσαρμόζει τις παραμέτρους μοντέλου π για την βελτιστοποίηση της αναμενόμενης πιθανότητας. Αν και το E-step στο Μαρκοβιανό αυτό μοντέλο είναι εννοιολογικά ξεκάθαρο, η ειδική δομή των πεπερασμένου χρόνου MDP επιτρέπει ορισμένες απλοποιήσεις και δίνει τη δυνατότητα αποφυγής ξεχωριστών περασμάτων συμπερασμού σε κάθε ένα από αυτά.

4.4.1 E-step: ταυτόχρονη διάδοση μηνυμάτων σε όλα τα MDP.

Λόγω της υπόθεσης σταθερών πιθανοτήτων μετάβασης, δύναται η χρήση απλούστερης γραφής $p(j|a,i) = P(x_{t+1} = j|a_t = a, x_t = i)$ και $p(j|i;\pi) = P(x_{t+1} = j|x_t = i;\pi) =$

$\sum_a p(j|a,i)\pi_{ai}$. Επιπλέον, για την έναρξη της προς τα πίσω διάδοσης μηνυμάτων, ορίζουμε

$$\begin{aligned}\hat{\beta} &= P(r=1|x_T=i;\pi)\\ &= \sum_a P(r=1|a_T=a, x_T=i)\pi_{ai}. \end{aligned} \tag{4.6}$$

Στο E-step, θεωρείται μια δεδομένη, σταθερή πολιτική π και όλες οι υπολογιζόμενες ποσότητες εξαρτώνται από το π, ακόμη και όπου αυτό δεν σημειώνεται. Για ένα MDP πεπερασμένου χρόνου T η τυπική προς τα εμπρός και προς τα πίσω διάδοση μηνυμάτων υπολογίζεται

$$\begin{aligned} a_0(i) &= \delta_{i=A}, & a_t(i) &= P(x_t=i|x_0=A;\pi)\\ & & &= \sum_j p(i|j;\pi)a_{t-1}(j), \end{aligned} \tag{4.7}$$

$$\begin{aligned} \tilde{\beta}_T(i) &= \hat{\beta}(i), & \tilde{\beta}_t(i) &= P(r=1|x_t=i;\pi)\\ & & &= \sum_j p(j|i;\pi)\tilde{\beta}_{t+1}(j). \end{aligned} \tag{4.8}$$

Τα παραπάνω δείχνουν πως οι ποσότητες a δεν εξαρτώνται από το T με κανένα τρόπο. Είνια έγκυρες δηλαδή για οποιοδήποτε ορισμένο χρόνο T. Το ίδιο δεν ισχύει όμως για τις ποσότητες $\tilde{\beta}$ όπως πάνω ορίστηκαν. Ορίζοντας τα β με δείκτες πίσω στο χρόνο (χρήση εναπομείναντα χρόνου τ), έχουμε

$$\begin{aligned}\beta_0(i) = \hat{\beta}(i), \beta_T(i) &= P(r=1|x_{T-\tau}=i;\pi)\\ &= \sum_j p(j|i;\pi)\beta_{\tau-1}(j). \end{aligned} \tag{4.9}$$

Ορισμένες κατά αυτό το τρόπο, οι ποσότητες β δεν εξαρτώνται από το T. Για ένα συγκεκριμένο MDP ορισμένου χρόνου T, η υπόθεση $\tau = T - t$ επιτρέπει τη συλλογή των ποσοτήτων $\tilde{\beta}$ με δείκτες προς τα εμπρός.

Λόγω των παραπάνω καθίσταται δυνατή η παράλληλη διανομή μηνυμάτων a και β μέσω της ταυτόχρονης αύξησης των t και τ, καθώς και η συλλογή των a και β για όλα τα MDP ορισμένου χρόνου T. Παρ' όλη την εισαγωγή μιας μίξης MDP απαιτείται μόλις ένα πέρασμα μηνυμάτων μπρος-πίσω. Η ασυνήθιστη αυτή, για τυπικά Κρυφά Μοντέλα Markov (Hidden Markov Models, HMM), διαδικασία είναι δυνατή λόγω της εξάρτησης του από την πρώτη ($x_0 = A$) και την τελευταία κατάσταση ($r = 1$). Πέραν της υλοποίησης της μείωσης με το prior του χρόνου, αυτό αποτελεί το κύριο λόγο θεώρησης της μίξης πεπερασμένου χρόνου MDPs (σχήμα 4.2) αντί ενός χρονικά αόριστου MDP με δυνατότητα εκπομπής αμοιβής σε κάθε χρόνο (σχήμα 4.1).

Κατά την διάδοση των μηνυμάτων a, β μπορούμε να υπολογίσουμε τις εκ των υστέρων πιθανότητες κατάστασης, οι οποίες εξαρτώνται από το συνολικό χρόνο T. Ορίζεται

$$\begin{aligned}\gamma_{t\tau}(i) &= P(x_t=i|x_0=A, r=1, T=t+\tau;\pi)\\ &= \frac{1}{Z(t,\tau)}\beta_t(i)a_t(i), \end{aligned} \tag{4.10}$$

υπό την κανονικοποίηση

$$Z(t,\tau) = \sum_i a_t(i)\beta_\tau(i) = \frac{P(x_0 = A, r = 1|T = t+\tau;\pi)}{P(x_0 = A)} = Z(t+\tau). \tag{4.11}$$

4.4.2 Εκ των υστέρων πιθανότητες χρόνου και αμοιβής.

Η σταθερά κανονικοποίησης Z εξαρτάται μόνο από το άθροισμα $t+\tau$, ή $Z(t+\tau) = Z(t,\tau)$. Επιπλέον, το $Z(t+\tau)$ σχετίζεται με την πιθανότητα $P(x_0 = A, r = 1|T = t+\tau;\pi)$, την πιθανότητα, δηλαδή, η αρχική κατάσταση να είναι A και η τελική κατάσταση να οδηγεί σε αμοιβή, εάν υποτεθεί MDP συγκεκριμένου μήκους T. Η χρήση του κανόνα του Bayes οδηγεί στην εκ των υστέρων πιθανότητα πάνω στο T (σε συμπτυγμένη γραφή),

$$P(T|x_0, r;\pi) = \frac{P(x_0, r|T;\pi)}{P(x_0, r;\pi)}P(T) = \frac{Z(T)P(T)}{P(r|x_0;\pi)}, \tag{4.12}$$

και της εκ των υστέρων πιθανότητας αμοιβής

$$P(r|x_0;\pi) = \sum_T P(T)Z(T). \tag{4.13}$$

4.4.3 Εκ των υστέρων πιθανότητες δράσεων και καταστάσεων.

Παρακάτω παράγονται οι εκ των υστέρων πιθανότητες δράσεων και καταστάσεων που παρουσιάζουν σημασία για το υπόλοιπο κείμενο. Για απλότητα, ορίζεται

$$\begin{aligned} q_{t\,\tau}(a,i) &= P(r = 1|a_t = a, x_t = i, T = t+\tau;\pi) \\ &= \begin{cases} \sum_j p(j|i,a)\beta_{\tau-1}(j) & \text{αν } \tau > 1 \\ P(r = 1|a_T = a, x_T = i) & \text{αν } \tau = 0 \end{cases}. \end{aligned} \tag{4.14}$$

Η πάνω ποσότητα είναι ανεξάρτητη των A και t λογω εξάρτησης από το x_t και του ότι η ιστορία προ χρόνου t καθίσταται μη σχετική στην αλυσίδα Markov. Θα χρησιμοποιείται η απλούστερη γραφή $q_\tau(a,i) = q_{t\,\tau}(a,i)$. Πολλαπλασιάζοντας με το prior του χρόνου και απαλείφοντας τον ολικό χρόνο προκύπτει η *εξαρτώμενη από τη δράση πιθανότητα*

$$P(r = 1|a_t = a, x_t = i;\pi) = \frac{1}{C}\sum_{\tau=0}^{\infty} P(T = t+\tau)p_\tau(a,i), \tag{4.15}$$

όπου $C = \sum_{\tau'} P(T = t+\tau')$, το οποίο για μειωμένο prior του χρόνου είναι ανεξάρτητο του t αφού $P(t+\tau) = \gamma^t P(\tau)$, το οποίο απορροφάται κατά την κανονικοποίηση. Στη συνέχεια, χρήση κανόνα του Bayes, προκύπτει η *η εκ των υστέρων πιθανότητα δράσης*

$$P(a_t = a|x_t = i, r = 1;\pi) = \frac{\pi_{ai}}{C'}\sum_{\tau=0}^{\infty} P(T = t+\tau)p_\tau(a,i), \tag{4.16}$$

όπου το $C' = P(r = 1|x_t = i;\pi)\sum_{\tau'} P(T = t+\tau')$ μπορεί να υπολογιστεί από την κανονικοποίηση, ενώ είναι επίσης ανεξάρτητο του t. Τέλος, η εκ των υστέρων πιθανότητα επίσκεψης μιας κατάστασης i είναι

$$P(i \in x_{0:T}|x_0, r = 1;\pi) = \sum_T \frac{P(T)Z(T)}{P(r = 1|x_0;\pi)}\Big[1 - \prod_{t=0}^{T}[1 - \gamma_{t,T-t}(i)]\Big]. \tag{4.17}$$

4.4.4 M-step: ενημέρωση πολιτικής.

Το τυπικό M-step ενός EM αλγορίθμου μεγιστοποιεί την προσδοκώμενη πλήρη λογαριθμική πιθανότητα $\mathcal{Q}(\pi^*, \pi) = \sum_T \sum_{x_{0:T}, a_{0:T}} P(x_{0:T}, a_{0:T}, T|r = 1; \pi) \log P(r = 1, x_{0:T}, a_{0:T}, T; \pi^*)$ πάνω στη νέα παράμετρο π^*, όπου οι προσδοκίες πάνω στις λανθάνουσες μεταβλητές $(T, x_{0:T}, a_{0:T})$ έχουν παρθεί από την εκ των υστέρων πιθανότητα, δεδομένων των παλιών παραμέτρων π. Στην εξεταζόμενη περίπτωση, σε ισχυρή αναλογία με την τυπική περίπτωση HMM, αυτό αναθέτει τις νέες παραμέτρους στην εκ των υστέρων πιθανότητα δράσης όπως δίνεται στην (4.16),

$$\pi^*_{ai} = P(a_t = a|x_t = i, r = 1; \pi). \tag{4.18}$$

Όμως, λόγω της δομής του MDP, η πιθανότητα μπορεί να γραφτεί ως

$$P(r = 1|x_0 = i; \pi^*) = \sum_{a_j} P(r = 1|a_t = a, x_t = j; \pi^*)\pi^*_{aj} \cdot P(x_t = j|x_0 = i; \pi^*). \tag{4.19}$$

Η μεγιστοποίηση της έκφρασης αυτής πάνω στο π^*_{aj} μπορεί να γίνει ξεχωριστά για κάθε ϑ και είναι έτσι ανεξάρτητη του $P(x_t = j|x_0 = i; \pi^*)$ οπότε και ανεξάρτητη του t επειδή η $P(r = 1|a_t = a, x_t = j; \pi^*)$ είναι επίσης (βλέπε εξίσωση 4.15). Χρησιμοποιώντας το E-step μπορεί να προσεγγιστεί ο πρώτος όρος και να μεγιστοποιηθεί το $\sum_a P(r = 1|a_t = a, x_t = j; \pi)\pi^*_{aj}$ μέσω

$$\pi^*_{ai} = \delta_{a=a^*(i)}, \qquad a^*(i) = \arg\max_a P(r = 1|a_t = a, x_t = i; \pi) \tag{4.20}$$

όπου το $a^*(i)$ μεγιστοποιεί την από την δράση εξαρτώμενη πιθανότητα (4.15). Δεδομένων των σχέσεων μεταξύ των εξισώσεων (4.15) και (4.16), η βασική διαφορά όσον αφορά το τυπικό M-step είναι ότι η ενημέρωση αυτή είναι πολύ πιο 'greedy' και συγκλίνει ταχύτερα για MDP.

4.5 Συνάφεια με το Policy Iteration

Οι ποσότητες β, όπως υπολογίζονται από τον αλγόριθμο διάδοσης προς τα πίσω, συναποτελούν στην πραγματικότητα τη συνάρτηση αξίας για ένα μοναδικό MDP πεπερασμένου χρόνου T. Πιο συγκεκριμένα, συγκρίνοντας την (4.9) με την πιθανότητα αμοιβής (4.3) για το MDP πεπερασμένου χρόνου T και τον ορισμό της συνάρτησης αξίας, έχουμε $\beta_\tau(i) \propto (V^\pi(i)$ του MDP χρόνου $T = \tau)$. Αντίστοιχα, η πλήρη συνάρτηση αξίας είναι η μίξη των β, $V^\pi(i) = \frac{1}{1-\gamma} \sum_T P(T)\beta_T(i)$, όταν έχει επιλεχθεί μειωμένο prior χρόνου. Επιπλέον, οι ποσότητες $q_\tau(a, i)$ όπως ορίζονται στην (4.14) σχετίζονται εξίσου με την συνάρτηση $\mathcal{Q}$, υπό την έννοια ότι $\mathcal{Q}^\pi(a, i) = \frac{1}{1-\gamma} \sum_T P(T)q_T(a, i)$ για κάποιο μειωμένο prior χρόνου. Το τελευταίο, μάλιστα, είναι και η δεσμευμένη πάνω στην δράση πιθανότητα (4.15) ενώ ισχύει επίσης ότι το 'posterior της δράσης (4.16) είναι ανάλογο του $\pi_{ai}\mathcal{Q}^\pi(a, i)$. Έτσι, *το E-step εκτελεί μία αξιολόγηση πολιτικής* που παραδίδει την κλασική συνάρτηση αξίας αλλά και επιπλέον τις εκ των υστέρων πιθανότητες του χρόνου, της κατάστασης και της δράσης, που δεν έχουν παραδοσιακό ανάλογο. Δεδομένης αυτής της σχέσης με την αξιολόγηση πολιτικής, το M-step εκτελεί μία ενημέρωση πολιτικής που (για το 'greedy' M-step (4.20)) είναι μια συνήθης ενημέρωση πολιτικής

στο Policy Iteration (Sutton and Barto (1998)), μεγιστοποιώντας την συνάρτηση Q πάνω στην δράση a σε κατάσταση i. Έτσι, ο αλγόριθμος EM με τη χρήση ακριβού συμπερασμού και αναπαράστασης πίστης είναι λειτουργικά ταυτόσημος του Policy Iteration (και από άποψη σύγκλισης) αλλά υπολογίζει τις απαραίτητες ποσότητες με διαφορετικό τρόπο. Όμως, με τη χρήση συμπερασματολογίας κατά προσέγγιση ο EM είναι ένας ποιοτικά διαφορετικός αλγόριθμος. Κατά την πρακτική υλοποίηση, η γνώση των εκ των υστέρων πιθανοτήτων για τον χρόνο, την κατάσταση και τη δράση μπορεί να χρησιμοποιηθεί για την αποκοπή αναίτιων υπολογισμών.

4.6 Ένας αλγόριθμος Viterbi-EM για τον υπολογισμό της βέλτιστης πολιτικής

Στην παράγραφο αυτή θα αναλυθεί ο νέος αλγόριθμος Viterbi-EM, που αποτελεί και τη βασική προσφορά αυτής της εργασίας. Η διατύπωση της αντικειμενικής συνάρτησης σε όρους πιθανότητας δίνει την δυνατότητα χρήσης του αλγορίθμου Viterbi (3.5.1) για την εύρεση της τροχιάς μέγιστης πιθανότητας $\arg\max_{\xi} P(\xi;\pi)$. Δεδομένου ότι η βελτιστοποίηση των παραμέτρων του προβλήματος έχει αναχθεί στην μεγιστοποίηση της πιθανότητας αμοιβής $L^{\pi}(i)$ (4.4), εκτιμάται ότι η χρήση της τροχιάς Viterbi ως οδηγό θα έχει ως αποτέλεσμα ικανοποιητικές προσεγγίσεις της βέλτιστης πολιτικής, με σημαντικά μειωμένο υπολογιστικό φόρτο. Ωστόσο, δεν παραβλέπεται ότι μια τέτοια προσέγγιση συνοδεύεται από τον κίνδυνο παραμονής σε τοπικό μέγιστο, κάτι το οποίο θα μελετηθεί, μεταξύ άλλων, στο επόμενο κεφάλαιο.

Ο Viterbi-EM αποτελείται από δύο επαναλαμβανόμενες διαδικασίες, όπως και ο αρχικός EM: το *E-step* και το *M-step*.

4.6.1 E-step: Εύρεση τροχιάς μέγιστης πιθανότητας MDP.

Στο βήμα αυτό υπολογίζεται η τροχιά μέγιστης πιθανότητας

$$\xi_{max} = \arg\max_{\xi} P(\xi;\pi), \tag{4.21}$$

ή

$$\xi_{max} = \arg\max_{T,x_{0:T}} P(x_0) \prod_{t=1}^{T} [\gamma\ P(x_t|x_{t-1};\pi)\ P(R|x_T)], \tag{4.22}$$

Η (4.22) απαιτεί λίγη ανάλυση. Ζητείται ο συνδυασμός τελικού χρόνου T, αρχικής κατάστασης x_0, τελικής κατάστασης x_T, και $T-1$ ενδιάμεσων καταστάσεων της τροχιάς, που να μεγιστοποιούν το γινόμενο $P(x_0)\prod_{t=1}^{T} P(x_t|x_{t-1};\pi)P(R|x_T)$. Αυτό αποτελείται από την πιθανότητα αρχικής κατάστασης $P(x_0)$, την πιθανότητα αμοιβής στην τελευταία κατάσταση $P(R|x_T)$, και τις πιθανότητες ενδιάμεσων μεταβάσεων δεδομένης της τρέχουσας πολιτικής $P(x_t|x_{t-1};\pi)$ για $t = 1$ ως $T-1$. Σημαντική παράμετρος στην πιο πάνω εξίσωση είναι και ο συντελεστής μείωσης γ που, όταν είναι μικρότερος της μονάδας, ευνοεί τροχιές με μικρότερο βάθος χρόνου.

Η μεγιστοποίηση, όπως ορίζεται στην (4.22), μπορεί να εκτελεστεί με μία παραλλαγή του αλγορίθμου Viterbi, έναν αλγόριθμο δυναμικού προγραμματισμού. Η μορφή του είναι 'από την αρχή προς το τέλος', που σημαίνει ότι σε κάθε χρονικό βήμα, t, εξετάζεται, για κάθε τωρινή κατάσταση, η προηγούμενη κατάσταση χρόνου $t-1$, που μεγιστοποιεί την ως τώρα συσσωρευμένη πιθανότητα. Η τωρινή κατάσταση αποθηκεύει την ως τώρα συλλεγμένη πιθανότητα και την προηγούμενη της κατάσταση. Σημειώνεται ότι, αφού μιλάμε για γινόμενα πιθανοτήτων, η αξίες των τροχιών θα μειώνονται συνεχώς.

Αυτό που κάνει ιδιαίτερο το δεδομένο πρόβλημα είναι ο ελεύθερος τελικός χρόνος T. Κάτι που λύνεται εύκολα με την παρατήρηση ότι, η πιθανότητα μιας ως τώρα ανεπτυγμένης υποτροχιάς εξαρτάται από το αν θα επιλεχθεί η τωρινή κατάσταση ως τελική ή θα συνεχιστεί η ανάπτυξη. Με άλλα λόγια, κάθε κατάσταση εκπέμπει ένα σήμα πιθανότητας $P(R|x_T)$, μόνο αν αποτελεί το τέλος της τροχιάς.

Σε ένα πρόβλημα με αυθαίρετες τιμές $P(R|x_T)$, θα πρέπει, σε κάθε βήμα του δυναμικού και για κάθε έγκυρη κατάσταση του προηγούμενου χρονικού βήματος, να συγκρίνεται η πιθανότητα της τροχιάς για αμοιβή στην κατάσταση αυτή, ξεχωριστά με την πιθανότητα συνέχισης $P(x_t|x_{t-1};\pi)$ κάθε αλγοριθμικά έγκυρης υποτροχιάς. Έτσι, αν στο χρονικό βήμα t, υπάρχουν δύο μεταβάσεις σε καταστάσεις i,j από μία κατάσταση k της χρονικής στιγμής $t-1$, θα πρέπει να συγκριθεί η πιθανότητα της τροχιάς που τερματίζει στην k, χωριστά με την ως τώρα συλλεγόμενη πιθανότητα των τροχιών που συνεχίζουν με τις μεταβάσεις i και j. Όποια μετάβαση οδηγεί σε τροχιά μικρότερης πιθανότητας, εξαιρείται από την συνέχεια του δυναμικού προγραμματισμού. Αν η τροχιά που τερματίζει στο k την στιγμή t έχει μικρότερη πιθανότητα από κάποια από τις συνεχιζόμενες από αυτήν τροχιές, θα πρέπει να αποθηκευτεί και είτε να συγκρίνεται κατ' επανάληψη με τις προερχόμενες από αυτήν τροχιές σε κάθε χρονικό βήμα, είτε να συγκριθεί με το σύνολο των τελικά εναπομείναντων τροχιών.

Σε περίπτωση που το πρόβλημα ορίζει δυαδικές τιμές $P(R|x_T)$, με τιμή 1 για τα σημεία τερματισμού και 0 αλλού, η παραπάνω διαδικασία είναι περιττή και αρκεί το σταμάτημα της ανάπτυξης του δυναμικού στις τελικές καταστάσεις. Και στις δύο παραπάνω περιπτώσεις επιλέγεται να μην αναπτύσσονται περαιτέρω οι τροχιές που ορίζουν κλειστούς βρόγχους.

Στο τέλος εκτέλεσης του δυναμικού προγραμματισμού θα υπάρχουν αρκετές αποθηκευμένες τροχιές με διαφορετικές πιθανότητες. Αυτή με την μεγαλύτερη πιθανότητα θα είναι και η ξ_{max}. Η ανάπτυξη του δυναμικού μοιάζει με δέντρο με διαφορετικά μήκη κλαδιών.

Ορίζεται $V(x)$ η τρέχουσα αξία της κατάστασης x και $V(x')$ η αξία μιας κατάστασης x', όπως υπολογίστηκε στο προηγούμενο E-step. Μπορούν να εκτελεστούν δύο διαδικασίες:

1. Η απόδοση μιας σχετικής αξίας $V(x)$ στις καταστάσεις που ανήκουν στην τροχιά μέγιστης πιθανότητας ξ_{max} εκτελώντας ένα βήμα 'back-up' μόνο στις καταστάσεις $x \in \xi_{max}$, ξεκινώντας από το τέλος της τροχιάς x_T και κατευθυνόμενοι προς την αρχική κατάσταση x_0.

$$V(x) = R(x;\pi) + \gamma \sum_{x'} P(x,x';\pi)V(x'), \qquad x \in \xi_{max}, \quad x' \in X \tag{4.23}$$

2. Μία πιο ελαστική εκτίμηση της αξίας, όπου τον υπολογισμό της (4.23) διαδέχεται ένα πλήρες βήμα Value Iteration:

$$V(x) = R(x;\pi) + \gamma \sum_{x'} P(x,x';\pi)V(x'), \qquad x \in X \tag{4.24}$$

Στην εκκίνηση του αλγορίθμου, τα $V(x)$ αρχικοποιούνται στο μηδέν. Η διαδικασία 1, δίνει προτεραιότητα ενημέρωσης της συνάρτησης αξίας $V(x)$, στις καταστάσεις της τροχιάς μέγιστης πιθανότητας ξ_{max}, και είναι το κομμάτι του $E-step$ που αξιοποιεί την τροχιά μέγιστης πιθανότητας ως οδηγό για την χάραξη της πολιτικής. Στην διαδικασία 2, στα $V(x')$ είναι συνυπολογισμένη η ενημέρωση κατά την διαδικασία 1. Η ενημέρωση της αξίας του συνόλου των καταστάσεων στη διαδικασία 2, δίνει θεωρητικά την δυνατότητα εύρεσης κοντινών στη τροχιά ξ_{max} λύσεων για την αποφυγή τοπικών βέλτιστων. Κατά την επανάληψη του επόμενου $E-step$, και μετά την εύρεση της νέας ξ_{max}, οι προηγούμενες υπολογισμένες τιμές $V(x)$ χρησιμοποιούνται ως $V(x')$ στις σχέσεις (4.23) και (4.24).

Τα x' στην (4.23) προέρχονται γενικά από το σύνολο του χώρου καταστάσεων X. Ωστόσο, χωρίς την εφαρμογή της (4.24), και εφόσον τα $V(x)$ έχουν αρχικοποιηθεί στο μηδέν, μπορεί, χωρίς καμία απώλεια πληροφορίας, να γίνει η απλοποίηση $x' \in \xi_{max}$

Τα παραπάνω στηρίζονται εν μέρη στη θεωρία σύγκλισης ασύγχρονου δυναμικού προγραμματισμού όπως αναλύθηκε στο (Bertsekas and Tsitsiklis (1989)). Παρόλα αυτά, στην περίπτωση μας δεν αποδεικνύεται σύγκλιση στο βέλτιστο αφού δεν εγγυάται η εξέταση του συνόλου του χώρου των καταστάσεων.

4.6.2 M-step: ενημέρωση πολιτικής.

Το M-step του αλγόριθμου Viterbi-EM είναι ανάλογο αυτού για τον EM. Τον ρόλο της ποσότητας β, ωστόσο, έρχεται να καλύψει η εκτίμηση V, όπως υπολογίστηκε στο E-step. Έτσι, για κάθε x:

$$\pi_{ax}^{new} = \pi_{ax}^{old}\Big[P(R|a,x) + \gamma \sum_{x'} V(x')P(x'|a,x)\Big] \tag{4.25}$$

Όπως και στον EM, μπορούμε και εδώ, εκμεταλλευόμενοι τη γνώση ότι η πολιτική του MDP είναι ντετερμινιστική, να εκτελέσουμε αντί του (4.25), το παρακάτω βήμα μεγιστοποίησης για κάθε x:

$$\pi_{ax}^{new} = \delta(a,a^*(x)), \qquad a^*(x) = \arg\max_a \Big[P(R|a,s) + \gamma \sum_{x'} V(x')P(x'|a,x)\Big] \tag{4.26}$$

4.7 Εξελίξεις και συμπεράσματα

Το μεγαλύτερο θετικό που προκύπτει από την σύνδεση του στατιστικού συμπερασμού με την κλασική θεωρία επίλυσης διαδικασιών απόφασης Markov είναι οι νέες επιλογές. Πλέον, το σύνολο των τεχνικών συμπερασμού είναι εφαρμόσιμο, ένα γεγονός ιδιαίτερα σημαντικό για πιο περίπλοκες δομές αναπαράστασης καταστάσεων όπως συνεχή ή ιεραρχικά δυναμικά δίκτυα Bayes.

Η νέα αυτή οπτική στην ενισχυτική μάθηση έχει αρχίσει να κινεί το ενδιαφέρον των ερευνητών (Vlassis and Toussaint (2009), Barber and Furmston (2009)).

Κεφάλαιο 5

Πειραματικά Αποτελέσματα

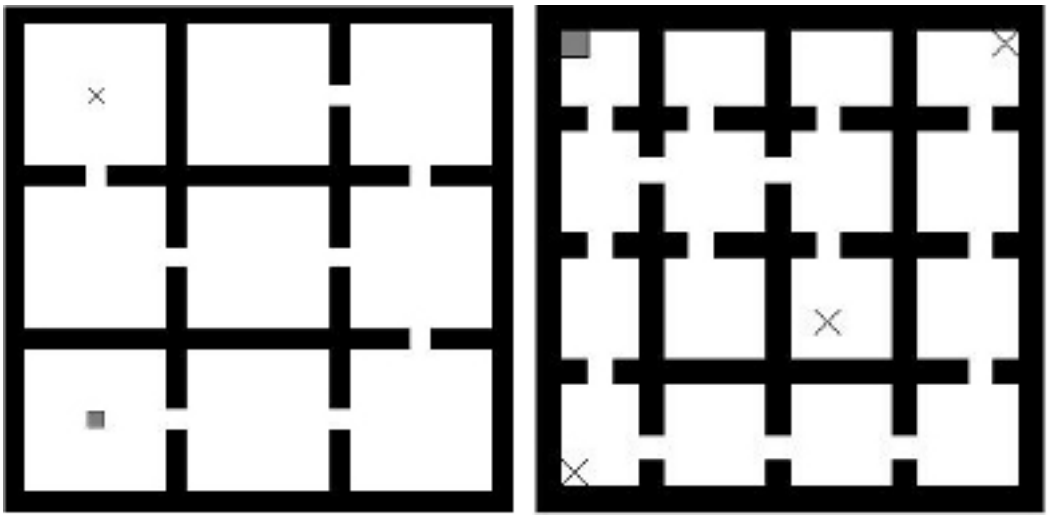
Στο κεφάλαιο αυτό, θα παρουσιαστούν τα αποτελέσματα από την εφαρμογή του αλγόριθμου Viterbi-EM σε τρία προβλήματα MDP, και θα συγκριθούν με αυτά των αλγόριθμων Policy και Value Iteration αλλά και, κυρίως, με αυτά του αρχικού αλγόριθμου EM. Πρώτα, θα δοθούν κάποιες γενικές πληροφορίες πάνω στη μοντελοποίηση των παρακάτω MDP.

5.1 Γενικά

Τα τρία προβλήματα που θα μελετηθούν έχουν την μορφή λαβυρίνθου, με τον πράκτορα να έχει δυνατότητες δράσης προς τις τέσσερις βασικές κατευθύνσεις, καθώς και μία δράση αναμονής στην ίδια κατάσταση. Όλες οι δράσεις χαρακτηρίζονται από υψηλά επίπεδα θορύβου με μια πιθανότητα 20% τυχαίας μετάβασης προς κάποια λανθασμένη κατάσταση, συμπεριλαμβανομένης της τρέχουσας. Οι χάρτες έχουν συντεταγμένες λευκού ή μαύρου χρώματος, με τις πρώτες να αντιπροσωπεύουν τις καταστάσεις του προβλήματος. Οι μαύρες συντεταγμένες έχουν ειδικό ρόλο, αφού μια μετάβαση προς αυτές οδηγεί σε μία κατάσταση παγίδα, εκτός των ορίων του χάρτη, από όπου δεν δύναται μετάβαση προς οποιαδήποτε άλλη κατάσταση. Κάτι τέτοιο οδηγεί σε αδυναμία εκπλήρωσης του στόχου του πράκτορα, που είναι η μετάβαση προς μια τελική κατάσταση (συμβολίζεται με 'X' στο χάρτη). Το σημείο εκκίνησης του κάθε λαβυρίνθου είναι μοναδικό και με πιθανότητα 1 (συμβολίζεται με ένα γκρίζο τετράγωνο).

Όλες οι δράσεις πλην αυτών από τις τελικές καταστάσεις αποδίδουν αμοιβή $r = 0$. Οποιαδήποτε δράση από τελική κατάσταση αποδίδει αμοιβή $r = 1$, μεταφέροντας στην συνέχεια τον πράκτορα στην κατάσταση παγίδα και προκαλώντας τον τερματισμό της διαδικασίας. Λόγου της μορφής αυτής του μοντέλου, η αξία της αρχικής θέσης, όπως υπολογίζεται από τους αλγόριθμους Policy και Value Iteration, θα είναι ίση με την τελική πιθανότητα αμοιβής, όπως θα υπολογιστεί για τους EM και Viterbi-EM, εφόσον ο πράκτορας δεν μείνει σε κάποιο τοπικό βέλτιστο.

Ο συντελεστής μείωσης επιλέχθηκε $\gamma = 0.99$, ενισχύοντας την αξία συντομότερων διαδρομών, ενώ η πολιτική αρχικοποιείται ομοιόμορφα σε όλες τις καταστάσεις.



(α') Χάρτης A (β') Χάρτης B

Σχήμα 5.1: Οι χάρτες A και B, όπως χρησιμοποιήθηκαν στα πειράματα.

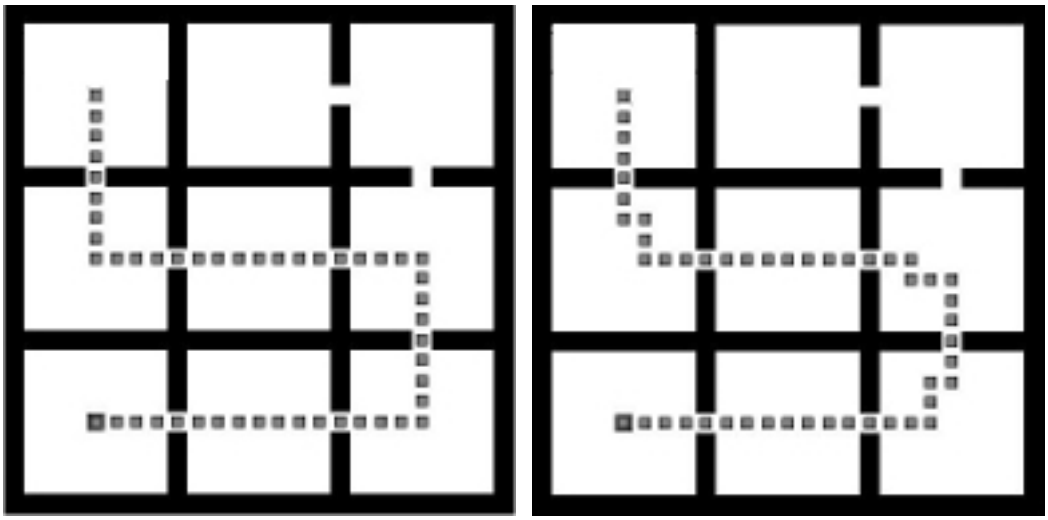
Στα συγκριτικά διαγράμματα (5.4, 5.6, 5.9) παρουσιάζεται η πιθανότητα εκπλήρωσης του στόχου $P(R;\pi)$ απέναντι στο υπολογιστικό κόστος, όπως υπολογίζεται από τον αριθμό εκτιμήσεων του περιβάλλοντος $P(x'|a,x)$ που απαιτούνται κατά το σχεδιασμό. Διαγράφονται οι καμπύλες για τους EM και Viterbi-EM, καθώς και αυτές για Policy και Value Iteration, για την καλύτερη εκτίμηση των αποτελεσμάτων.

5.2 Χάρτης A

Ο πρώτος χάρτης (5.1 α) έχει μια σχετικά απλή μορφή, με την λύση να δείχνει εξ' αρχής ξεκάθαρη. Παρόλο αυτό, η διαδρομή μπορεί να εμφανίσει αρκετές παραλλαγές, καθώς ο πράκτορας θα πρέπει να εξισορροπήσει την ανάγκη του να μικρύνει την απόσταση από το στόχο, με το κόστος του να κινείται κοντά στους τοίχους.

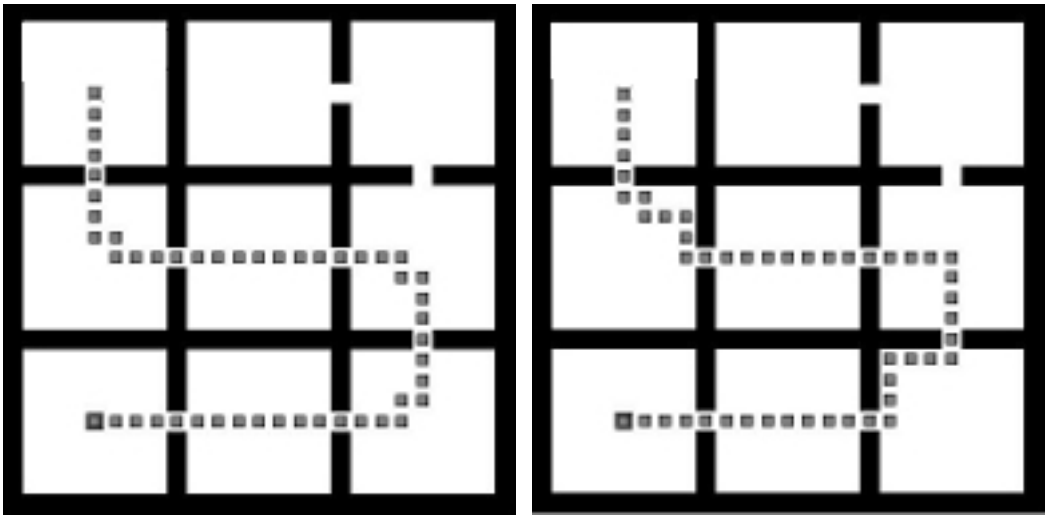
Ο EM αλλάζει δύο μέγιστες τροχιές (5.2 α, 5.2 β) ώσπου να φτάσει στην τρίτη (5.3 α), για να συγκλίνει σε πιθανότητα αμοιβής 0.22. Ο Viterbi-EM αντιθέτως παραμένει πιστός στην πρώτη υπολογισμένη μέγιστη τροχιά (5.3 β), καταλήγοντας σε μια χειρότερη λύση, 0.11.

Στο διάγραμμα (5.4) παρουσιάζεται η σύγκλιση των διαφόρων μεθόδων. Είναι εμφανές ότι η σύγκλιση του EM αργεί δραματικά σε σχέση με τους υπόλοιπους αλγόριθμους. Ο Viterbi-EM όμως, αν και ταχύτατα, συγκλίνει σε μία υποβέλτιστη λύση του προβλήματος. Εκτιμάται πως μετά την άμεση επιλογή μιας μέγιστης τροχιάς, ο Viterbi-EM δυσκολεύεται να ξεφύγει από τοπικά μέγιστα.



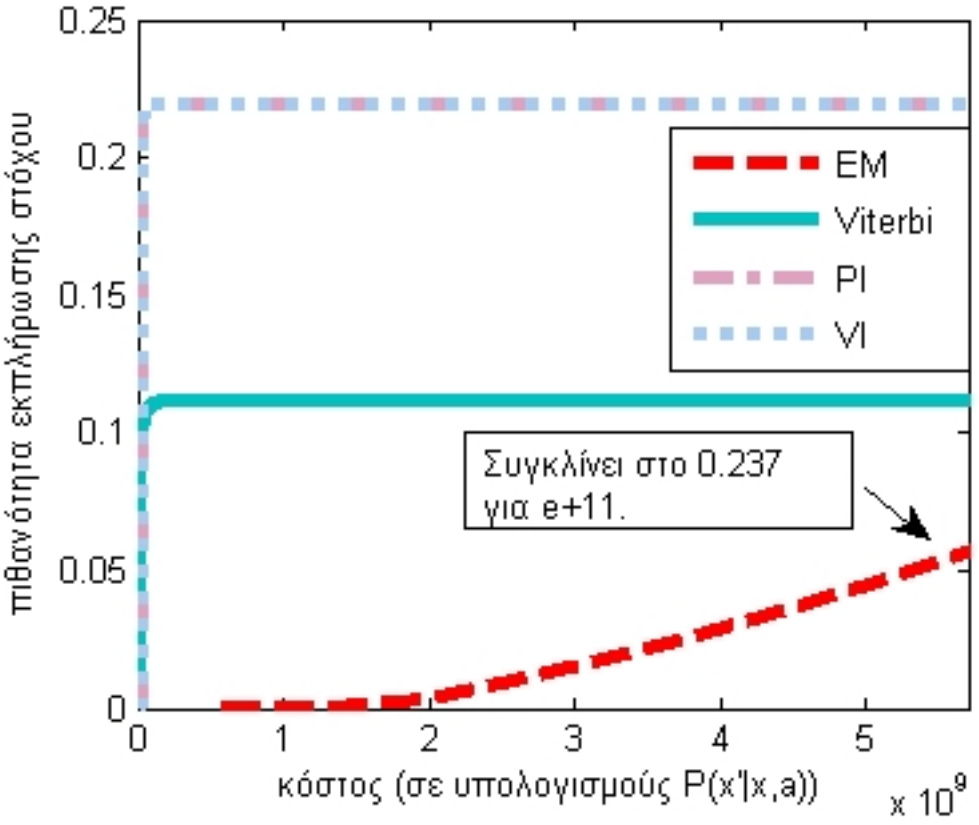
(α') EM 1 - $P(R;\pi) = 0.0576$ (β') EM 2 - $P(R;\pi) = 0.1758$

Σχήμα 5.2: Οι δύο πρώτες μέγιστες τροχιές όπως υπολογίστηκαν σταδιακά από τον EM για τον Χάρτη *A*.



(α') EM 3 - $P(R;\pi) = 0.2178$ (β') Viterbi-EM - $P(R;\pi) = 0.1118$

Σχήμα 5.3: Η τρίτη μέγιστη τροχιά (α) όπως υπολογίστηκε από τον EM για τον Χάρτη *A*, και η υπολογιζόμενη, μέγιστη τροχιά από τον Viterbi-EM, για τον ίδιο χάρτη.



Σχήμα 5.4: Αποτελέσματα εκτέλεσης αλγορίθμων στον Χάρτη *A*.

5.3 Χάρτης *B*

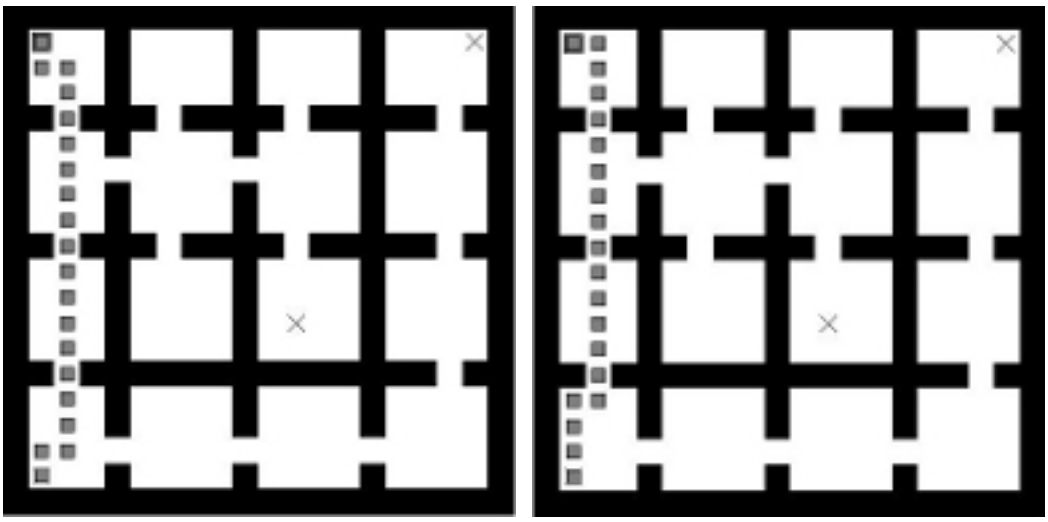
Στο Χάρτη *B* (5.1 β) έχουν τοποθετηθεί τρία σημεία τερματισμού, σε διαφορετικές αποστάσεις από την αρχική κατάσταση, με παρόμοια δυσκολία της κάθε διαδρομής. Ο σωστός υπολογισμός της πολιτικής, θα πρέπει να οδηγήσει σε μια τροχιά προς το κάτω αριστερά τέρμα.

Πράγματι, τόσο ο EM (5.5 α), όσο και ο Viterbi-EM (5.5 β), συγκλίνουν προς την βέλτιστη τροχιά. Ο Viterbi-EM δείχνει πάλι μια προσπάθεια να κινηθεί κατά μήκος του τοίχου, με αποτέλεσμα χειρότερη πιθανότητα αμοιβής 0.31, έναντι 0.36 του EM.

Στο διάγραμμα (5.6) φαίνεται πως η ταχύτητα σύγκλισης του Viterbi-EM έχει το τίμημα της στην τελική λύση.

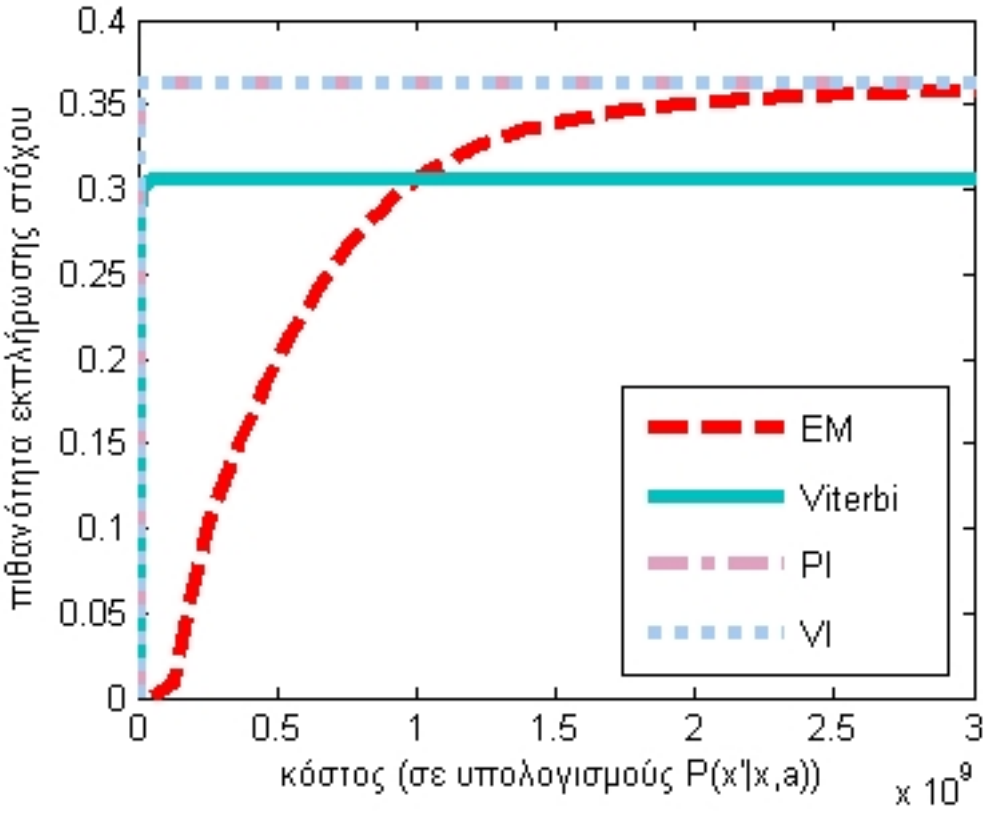
5.4 Χάρτης *Γ*

Ο τρίτος χάρτης (5.7) έχει σχεδιαστεί για την εξέταση της ευαισθησίας του αλγόριθμου Viterbi-EM σε τοπικά μέγιστα. Παρατηρούνται δύο γενικές πορείες προς το τέρμα. Η πιο χαμηλή, αν και συντομότερη, διασχίζει έναν στενό διάδρομο, με σημαντική πιθανότητα πρόωρου τερματισμού (10% ανά βήμα). Η ψηλότερη,

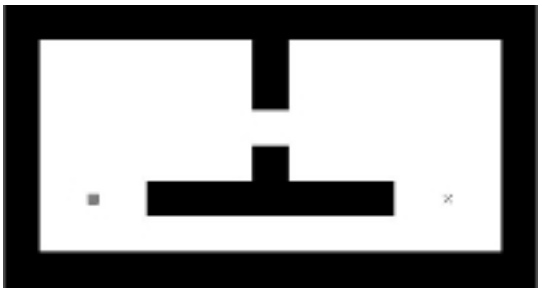


(α') EM - $P(R;\pi) = 0.36$ (β') Viterbi-EM - $P(R;\pi) = 0.31$

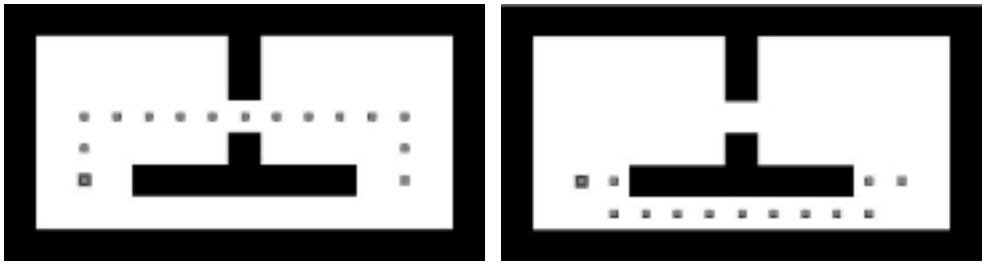
Σχήμα 5.5: Οι μέγιστες τροχιές όπως υπολογίστηκαν από τους EM και Viterbi-EM για τον Χάρτη *B*.



Σχήμα 5.6: Αποτελέσματα εκτέλεσης αλγορίθμων στον Χάρτη *B*.



Σχήμα 5.7: Ο χάρτης *Γ*, όπως χρησιμοποιήθηκε στα πειράματα.



(α') EM - $P(R;\pi) = 0.69$ (β') Viterbi-EM - $P(R;\pi) = 0.28$

Σχήμα 5.8: Οι μέγιστες τροχιές όπως υπολογίστηκαν από τους EM και Viterbi-EM για τον Χάρτη *Γ*.

αντιθέτως, παρουσιάζει μόλις ένα στενό σημείο, και τα δύο επιπλέον βήματα που της αναλογούν δεν αναμένεται να αποτρέψουν την επιλογή της.

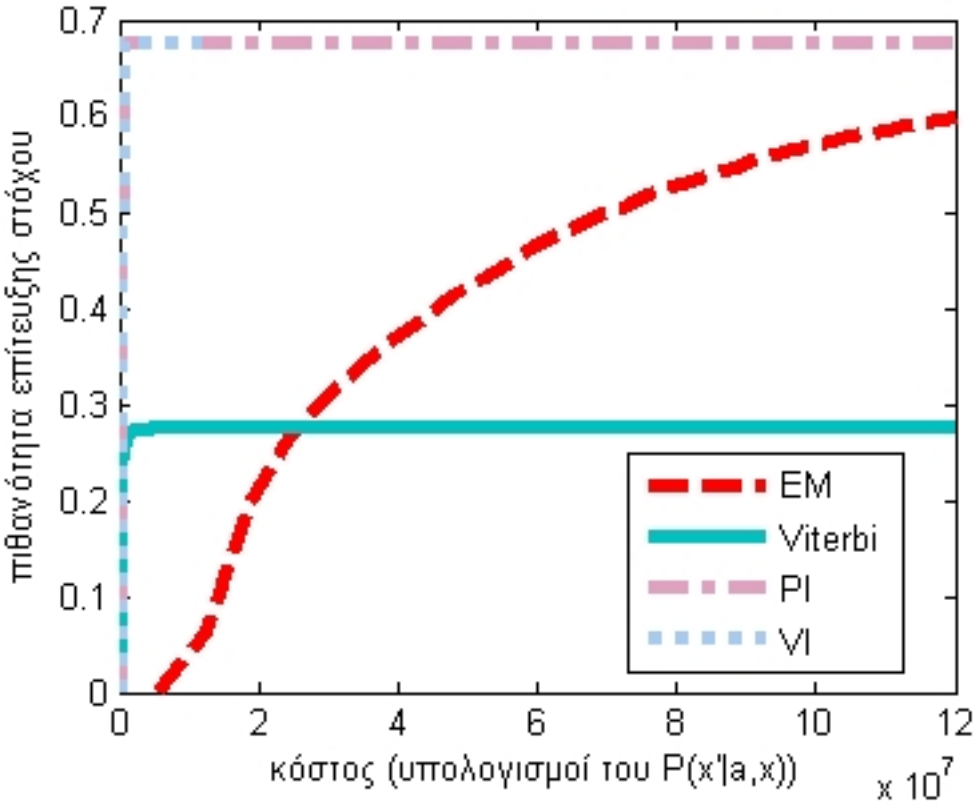
Παρατηρείται, ωστόσο, πως ο αλγόριθμος Viterbi-EM αποτυγχάνει στην εύρεση της βέλτιστης τροχιάς, κατευθυνόμενος από κάτω (5.8 β). Σε απόδειξη αυτού, ο EM κατευθυνόμενος προς τα πάνω (5.8 α), κατορθώνει σχεδόν τριπλάσια πιθανότητα αμοιβής, 0.69, έναντι 0.28 του Viterbi-EM. Αυτό που συνέβη, επιβεβαιώνοντας την αντίστοιχη εκτίμηση για τον Χάρτη *Α* (5.1 α), είναι ότι, ξεκινώντας από ομοιόμορφη κατανομή, η αρχική τροχιά μέγιστης πιθανότητας κατευθύνεται από κάτω. Ωθώντας στην συνέχεια την πολιτική των γειτονικών καταστάσεων προς την τροχιά αυτή, δεν δίνεται η δυνατότητα εύρεσης της πάνω τροχιάς. Σε παρόμοιες περιπτώσεις, η προϋπάρχουσα γνώση για το πρόβλημα θα μπορούσε να ωθήσει τον Viterbi-EM στην επιλογή της σωστής κατεύθυνσης.

Τα παραπάνω δίνονται παραστατικά στο σχήμα (5.9)

5.5 Συμπεράσματα

Εξαιρουμένου του τρίτου χάρτη, και οι τέσσερις αλγόριθμοι κατέληξαν σε παρόμοιες λύσεις, με την λύση του Viterbi-EM να υστερεί σταθερά σε πιθανότητα αμοιβής. Η μικρή αυτή απόκλιση προκύπτει από την μη ενημέρωση της πολιτικής, για καταστάσεις μακριά από την τροχιά μέγιστης πιθανότητας και την αδιαφορία που επιδεικνύει για τις καταστάσεις παγίδες.

Αξίζει να σημειωθεί η σημαντική δυνατότητα επιτάχυνσης των αλγόριθμων Viterbi και Viterbi-EM μέσω της αξιοποίησης του μοντέλου του προβλήματος



Σχήμα 5.9: Αποτελέσματα εκτέλεσης αλγορίθμων στον Χάρτη *Γ*.

ή και μεθόδων αποκοπής υπολογισμών. Ειδικά για τον EM έχει υλοποιηθεί αντίστοιχη μέθοδος στα Toussaint and Storkey (2006) και Toussaint (Storkey and Harmeling). Προτιμήθηκε ωστόσο ένα λιγότερο βεβιασμένο συγκριτικό, με την καθαρή μορφή των αλγόριθμων. Η φαινομενική υπεροχή των Policy Iteration, Value Iteration είναι έτσι μικρής σημασίας και τα γραφήματα τους παρατίθενται για λόγους πληρότητας και διευκόλυνσης της διαίσθησης.

Κεφάλαιο 6

Συμπεράσματα

Στην εργασία αυτή, παρουσιάστηκε για πρώτη φορά η υλοποίηση του αλγορίθμου Viterbi-EM για την βελτιστοποίηση ελεγκτών MDP με την χρήση του αλγόριθμου εύρεσης της πιο πιθανής τροχιάς καταστάσεων, Viterbi. Ο αλγόριθμος αυτός, προήλθε από μία προσπάθεια αξιοποίησης των δυνατοτήτων που εμφανίστηκαν έπειτα από την παρουσίαση του Toussaint and Storkey (2006). Στο κείμενο αυτό δόθηκε για πρώτη φορά μία αποδεδειγμένη ισοδυναμία της μεγιστοποίησης μιας πιθανότητας και της βελτιστοποίησης μιας διαδικασίας αποφάσεων Markov, στοιχείο πάνω στο οποίο στηρίχθηκε η υλοποίηση του Viterbi-EM.

Τα πειραματικά δεδομένα αποδεικνύουν την εγκυρότητα της υλοποίησης, παρουσιάζοντας όμως ταυτόχρονα την τάση του Viterbi-EM να συγκλίνει σε τοπικά βέλτιστα. Το θετικότερο στοιχείο που αποκομίστηκε είναι η μεγάλη ταχύτητα με την οποία ο αλγόριθμος συγκλίνει, κάτι το οποίο τον καθιστά άξιο περισσότερης μελέτης.

Προτεινόμενα θέματα στην κατεύθυνση της εργασίας είναι:

1. Η εξέταση τεχνικών αποφυγής τοπικών μεγίστων για τον Viterbi-EM.
2. Ενσωμάτωση τεχνικών επιτάχυνσης αλγορίθμων και αξιοποίηση του μοντέλου σε προβλήματα πρακτικής σημασίας.
3. Εκμετάλλευση του μικρού υπολογιστικού φόρτου που παρουσιάζει ο Viterbi-EM σε On-line εφαρμογές.
4. Υλοποίηση αλγορίθμου Viterbi-EM για την βελτιστοποίηση ελεγκτών POMDP, και model-free προβλημάτων.

Βιβλιογραφία

Baird, L., Moore, A. (1998). Gradient Descent for General Reinforcement Learning. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*,MIT Press, Cambridge, MA, 1999.

Barber, D. and Furmston, T. (2009). Solving deterministic policy (PO)MPDs using Expectation-Maximisation and Antifreeze. In *European Conference on Machine Learning (LEMIR workshop)*, pp. 50-64, 2009.

Barto, A. G. and Mahadevan, S. (2003). Recent Advances in Hierarchical Reinforcement Learning. In *Discrete Event Dynamic Systems 13*, 4 (2003), 341-379.

Bellman, R. E. (1957a) *Dynamic Programming*. Princeton University Press, Princeton, NJ.

Bellman, R. E. (1957b) A Markov decision process. In *Journal of Mathematical Mech.*, 6:679-684.

Bertsekas, D. P. and Tsitsiklis, J. N. (1989) Convergence rate and termination of asynchronous iterative algorithms. In *Proc. of the 3rd international conference on Supercomputing* (Crete, Greece, June 05 - 09, 1989). ICS '89. ACM, New York, NY, 461-470.

Cassandra, A., Littman, M. L. and Zhang, N. L. (1997). Incremental Pruning: A Simple, Fast, Exact Method for Partially Observable Markov Decision Processes. In *Proc. of the Thirteenth Conference on Uncertainty in Artificial Intelligence*,Morgan Kaufmann Publishers, 54–61 (1997).

Claus, C. and Boutilier, C. (1998). The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems. In *Proc. Proceedings of the Fifteenth National Conference on Artificial Intelligence*, AAAI Press, 746–752 (1998)

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. In *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, No. 1., pp. 1-38, 1977.

Fu, K. S. (1970). Learning control systems–Review and outlook. In *IEEE Transactions on Automatic Control*,pages 210-221.

Hansen, E. A. (1998). Solving POMDPs by Searching in Policy Space. In *Proc. of the 14th International Conference on Uncertainty in Artificial Intelligence (1998.*

Howard, R. (1960). *Dynamic Programming and Markov Processes.* MIT Press, Cambridge, MA.

Hsu, D., Lee, W. S. and Rong, N. (2007). What Makes Some POMDP Problems Easy to Approximate? In John C. Platt, Daphne Koller, Yoram Singer and Sam T. Roweis, ed., 'NIPS' , MIT Press.

Littman, M. L. (1996). *Algorithms for Sequential Decision Making.* Ph.D., Brown University, May 1996.

Mendel, J. M. (1966). Applications of artificial intelligence techniques to a spacecraft control problem. Technical Report NASA CR-755, National Aeronautics and Space Administration.

Mendel, J. M. and McLaren, R. W. (1970). In Mendel, J. M. and Fu, K. S., editors, *Adaptive, Learning and Pattern Recognition Systems: Theory and Applications*, pages 287-318. Academic Press, New York.

Minsky, M. L. (1961). Steps toward artificial intelligence. In *Proceedings of the Institute of Radio Engineers*, 49:8-30. Reprinted in E. A. Feigenbaum and J. Feldman, editors, *Computers and Thought.* McGraw- Hill, New York, 406-450, 1963.

Murphy, K. P. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning.*

Rabiner, L. and Juang, B. (1993) *Fundamentals of speech recognition.* Prentice-Hall, Inc.

Schmidhuber, J. 1997. A Computer Scientist's View of Life, the Universe, and Everything. In *Foundations of Computer Science: Potential - theory - Cognition, To Wilfried Brauer on the Occasion of His Sixtieth Birthday.* C. Freksa, M. Jantzen, and R. Valk, Eds. Lecture Notes In Computer Science, vol. 1337. Springer-Verlag, London, 201-208.

Spaan, M. T. J. and Vlassis, N. (2005). Perseus: Randomized point-based value iteration for POMDPs. In *Journal of Artificial Intelligence Research*, vol.24, 195-220 (2005).

Sutton, R. S. and Barto, A. G. (1998). Reinforcement Learning: An Introduction. In *IEEE Transactions on Neural Networks*, 9(5): 1054-1054 (1998)

Toussaint, M. and Storkey, A. (2006). Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *Proc. of the 23rd international Conference on Machine Learning (ICML 2006)*. vol. 148. ACM, New York, NY, 945-952

Toussaint, M., Storkey, A. and Harmeling, S. (2010). Expectation-Maximization methods for solving (PO)MDPs and optimal control problems. In Silvia Chiappa, David Barber (Eds.): *Inference and Learning in Dynamic Models*, Cambridge University Press, print expected in 2010.

Vlassis, N. and Toussaint, M. (2009). Model-Free Reinforcement Learning as Mixture Learning. In *Proc. of the 25nd International Conference on Machine Learning (ICML)*, 2009.

Waltz, M. D. and Fu, K. S. (1965). A heuristic approach to reinforcement learning control systems. In *IEEE Transactions on Automatic Control*, 10:390-398.