ΑΥΤΟΝΟΜΗ ΠΛΟΗΓΗΣΗ ΡΟΜΠΟΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ ΜΕ ΤΕΧΝΙΚΕΣ ΕΝΙΣΧΥΤΙΚΗΣ ΜΑΘΗΣΗΣ

Η ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης

του τμήματος Πληροφορικής Εξεταστική Επιτροπή

από τον

Νικόλαο Τζιωρτζιώτη

ως μέρος των Υποχρεώσεων για τη λήψη του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ

ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Σεπτέμβριος 2010

ΑΦΙΕΡΩΣΗ

Στην οικογένειά μου

Ευχαριστιες

Αρχικά θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Κωνσταντίνο Μπλέκα για τις πολύτιμες συμβουλές, την υπομονή και την άψογη καθοδήγηση του σε όλη τη διάρκεια εκπόνησης της μεταπτυχιακής μου διατριβής.

Επίσης αισθάνομαι την ανάγκη να ευχαριστήσω τους γονείς μου που στέκονται δίπλα μου όλα αυτά τα χρόνια και με στηρίζουν σε κάθε μου προσπάθεια.

Τέλος, οφείλω ένα μεγάλο ευχαριστώ σε όλους τους δικούς μου ανθρώπους για τη κατανόηση και την υπομονή που επέδειξαν όλο αυτό το χρονικό διάστημα.

Περιεχομενα

Ευρετηριο Σχηματων

Ευρετηριο Αλγοριθμων

Περιληψη

Νικόλαος Τζιωρτζιώτης του Βασιλείου και της Αναστασίας. MSc, Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Σεπτέμβριος, 2010. Αυτόνομη Πλοήγηση Ρομποτικών Συστημάτων με Τεχνικές Ενισχυτικής Μάθησης.
Επιβλέπων: Κωνσταντίνος Μπλέκας.

Η εργασία πραγματεύεται την αυτόνομη πλοήγηση ρομποτικών συστημάτων με τεχνικές ενισχυτικής μάθησης. Η πλοήγηση είναι ένα από τα σημαντικότερα συστατικά ενός ρομποτικού συστήματος το οποίο προσπαθεί να κατευθύνει ένα ρομπότ με ασφάλεια μέσα σε ένα άγνωστο περιβάλλον. Στη παρούσα διατριβή μελετούμε τεχνικές ενισχυτικής μάθησης για την επίτευξη της πλοήγησης. Η ενισχυτική μάθηση είναι μια κλάση προβλημάτων μάθησης που ασχολούνται με την επίτευξη μακροπρόθεσμων στόχων σε άγνωστα, αβέβαια και δυναμικά περιβάλλοντα. Τέτοιες εργασίες συνήθως μοντελοποιούνται ως Μαρκοβιανές διαδικασίες απόφασης ή πιο γενικά ως μερικά παρατηρήσιμες Μαρκοβιανές διαδικασίες απόφασης. Στο πρώτο μέρος, μελετώνται οι τρεις θεμελιώδεις κλάσεις μεθόδων για την επίλυση του προβλήματος της ενισχυτικής μάθησης: ο δυναμικός προγραμματισμός, οι μέθοδοι Monte Carlo και η μάθηση χρονικών διαφορών. Όλες αυτές οι μέθοδοι επιλύουν πλήρως το πρόβλημα της ενισχυτικής μάθησης. Στη συνέχεια, παρουσιάζεται μια Μπεϋσιανή προσέγγιση εκτίμησης πολιτικής σε γενικούς χώρους καταστάσεων και ενεργειών, η οποία χρησιμοποιεί Γκαουσιανές διαδικασίες για τις συναρτήσεις αξίας. Στην παρούσα διατριβή η επέκταση που προτείνεται εισαγεί τη χρήση RVM στην παραπάνω Μπεϋσιανή προσέγγιση εκτίμησης πολιτικής. Η συγκεκριμένη επέκταση οδηγεί σε ακόμη αραιότερα μοντέλα εκτίμησης της συνάρτησης αξίας. Οι μέθοδοι που αναπτύχθηκαν εφαρμόζονται και αξιολογούνται σε δύο γνωστά πειραματικά περιβάλλοντα αυτόνομης πλοήγησης ρομποτικών συστημάτων: στο ανεστραμμένο εκκρεμές (Cart Pole) και το Mountain Car. Επιπλέον, μελετήθηκε το πρόβλημα της πλοήγησης ενός πραγματικού ρομπότ τύπου PeopleBot, το οποίο υπάρχει διαθέσιμο στο τμήμα. Η σημαντικότερη καινοτομία της παρούσας διατριβής είναι η υλοποίηση και η εφαρμογή των παραπάνω μεθόδων στο πρόβλημα πλοήγησης ενός πραγματικού αυτόνομου ρομποτικού συστήματος, δίνοντάς μας τη δυνατότητα να αξιολογήσουμε τις συγκεκριμένες μεθόδους σε συνθήκες πραγματικού περιβάλλοντος.

Extended Abstract in English

Tziortziotis, Nikolaos, V. MSc, Computer Science Department, University of Ioannina, Greece. September, 2010. Autonomous Mobile Robot Navigation using Reinforcement Learning.
Thesis Supervisor: Konstantinos Blekas.

Reinforcement learning is a class of problems frequently encountered by both biological and artificial agents, and concerned with achieving long-term goals in unfamiliar, uncertain and dynamic environments. Reinforcement learning is learning what to do (how to map situations to actions) so as to maximize a numerical reward signal. It is distinguished from other computational approaches by its emphasis on learning by the individual from direct interaction with its environment, without relying on exemplary supervision or complete models of the environment. The learner is not told which actions to take, as in most forms of Machine learning, but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the next situations and, through that, all subsequent rewards. These two characteristics, trial-and-error and delayed reward, are the two most important distinguishing features of reinforcement learning.

Reinforcement learning uses a formal framework defining the interaction between a learning agent and its environment in terms of states, actions, and rewards. The reinforcement learning agent and its environment interact over a sequence of discrete time steps. The specification of their interface defines a particular task: the actions are the choices made by the agent, the states are the basis for making the choices, and the reward are the basis for evaluating the choices. This framework is intended to be a simple way of representing essential features of the artificial intelligent problem. These features include a sense of cause and effect, a sense of uncertainty and nondeterminism, and the existence of explicit goals.

The concepts of value and value functions are the key features of the reinforcement learning methods. We take the position that value functions are essential for efficient search in the space of policies. A policy is a stochastic rule by which the agent selects actions as a function of states. Additionally, a policy's value functions assign to each state, or state-action pair, the expected return from that state, or state-action pair, given that the agent uses that policy. Their use of value functions distinguish reinforcement learning methods from evolutionary methods that search directly in policy space guided by scalar evaluations of entire policies. An important algorithmic component of many

Reinforcement learning methods is the estimation of state or state-action values of a fixed policy controlling a Markov decision Process (MDP), a task known as policy evaluation.

In the first part of the thesis, we describe three fundamental classes of methods for solving the reinforcement learning problem: dynamic programming, Monte Carlo methods, and temporal-difference learning. All of these methods solve the full version of the problem, including delayed rewards. Each class of methods has its strengths and weaknesses. Dynamic programming methods are well developed mathematically, but require a complete model of the environment. Monte Carlo methods don't require a model and are conceptually simple, but are not suited for step-by-step incremental computation. Finally, temporal-difference methods require no model and are fully incremental, but are more complex to analyze. The methods also differ in several ways with respect to their efficiency and speed of convergence.

Moreover, we describe a Bayesian approach to policy evaluation in general state and action spaces, which uses a statistical generative model for value functions via Gaussian processes. The posterior distribution based on a GP-based statistical model provides us with a value-function estimate, as well as a measure of the variance of that estimate, opening the way to a range of possibilities. An efficient sequential kernel sparsification method allows us to derive efficient online algorithms for learning good approximations of the posterior moments. Evaluating state-action values we derive model-free algorithms based on Policy Iteration for improving policies, thus tackling the complete RL problem. Furthermore, in the current work we propose the use of RVM in the existent Bayesian approach to policy evaluation in general spaces. In this way more sparse models are generated.

Finally, we conducted a number of experiments based on the algorithms described previously. These algorithms are applied and evaluated in several experimental environments of autonomous mobile robots navigation systems, such as Mountain Car and Cart Pole. More specifically we applied the following methods: Sarsa, Sarsa(λ) and GPTD. The novelty of this thesis is the application of these methods in a real mobile robot system (PeopleBot). In this way, we had the opportunity to evaluate and analyze these methods to real world problems.

Κεφαλαιο 1

ΕΙΣΑΓΩΓΗ

1.1 Αυτόνομη Πλοήγηση Ρομποτικών Συστημάτων

Ένα ρομποτικό σύστημα (Σχήμα 1.1) είναι μια ευφυής μηχανή ικανή να λειτουργεί αυτόνομα σε ένα οποιοδήποτε περιβάλλον και η οποία είναι σε θέση να αντιλαμβάνεται (αντίληψη του περιβάλλοντος), να σκέφτεται (σχεδίαση) και να ενεργεί (κίνηση). Ωστόσο, τα ρομπότ είναι κατάλληλα εργαλεία για την διερεύνηση προβλημάτων τεχνητής νοημοσύνης όπως η αποφυγή εμποδίων (obstacle avoidance), σχεδίαση μονοπατιού (path planning), κλπ. Ένα από τα χαρακτηριστικά των ρομποτικών συστημάτων είναι η πολυπλοκότητα των περιβαλλόντων μέσα στα οποία κινούνται. Ως εκ τούτου, ένα από τα κρισιμότερα προβλήματα για τα ρομπότ είναι η σχεδίαση του μονοπατιού.

Σχήμα 1.1: Ρομποτικό Σύστημα

Η πλοήγηση (navigation) είναι ένα ζωτικής σημασίας συστατικό ενός αυτόνομου ρομποτικού συστήματος που προσπαθεί να το κατευθύνει μέσα σε κάποιο άγνωστο περιβάλλον ενώ αυτό κινείται. Στόχος των συστημάτων πλοήγησης είναι η οδήγηση του ρομπότ σε κάποιο προορισμό μέσα σε ένα γνωστό, άγνωστο, ή μερικά γνωστό περιβάλλον, χωρίς να

χαθεί ή να προσκρούσει σε κάποιο εμπόδιο. Πρακτικά, το ρομπότ δεν μπορεί να βρει άμεσα ένα μονοπάτι από κάποιο αρχικό σημείο πρός ένα προορισμό. Για το λόγο αυτό θα πρέπει να χρησιμοποιηθούν τεχνικές εύρεσης μονοπατιού που συνεπάγονται τη μετάβαση από μια αφετηρία σε κάποιο προορισμό ενώ ταυτόχρονα ελαχιστοποιούν κάποιο κόστος, όπως για παράδειγμα το χρόνο που δαπανάται για τη μετάβαση στο προορισμό του. Όταν ένα ρομπότ αρχίζει να κινείται ακολουθώντας ένα ήδη σχεδιασμένο μονοπάτι υπάρχει πιθανότητα να συναντήσει κάποιο εμπόδιο, τότε θα πρέπει να αποφύγει το συγκεκριμένο εμπόδιο και να σχεδιάσει ένα καινούριο μονοπάτι. Με το τρόπο αυτό επιτυγχάνεται η εργασία της πλοήγησης.

Η εργασία της πλοήγησης περιλαμβάνει συνήθως τη σχεδίαση μονοπατιού (path planning) και τη σχεδίαση πορείας (trajectory planning). Η σχεδίαση μονοπατιού είναι η εύρεση ενός μονοπατιού απαλλαγμένου από εμπόδια σε ένα περιβάλλον με εμπόδια και η βελτιστοποίηση του σύμφωνα με κάποια κριτήρια. Η σχεδίαση πορείας είναι ο προγραμματισμός των κινήσεων ενός ρομπότ κατά μήκος του σχεδιασμένου μονοπατιού. Αρκετές προσεγγίσεις έχουν προταθεί για το πρόβλημα σχεδίασης της κίνησης ενός ρομποτικού συστήματος μέσα σε ένα περιβάλλον. Ένας αλγόριθμος ονομάζεται εξομοιούμενος (off-line) εάν παράγει εκ των προτέρων ένα μονοπάτι για ένα ήδη γνωστό στατικό περιβάλλον. Αντίθετα ονομάζεται απευθείας (on-line) εάν έχει τη δυνατότητα εύρεσης ενός καινούριου μονοπατιού εξαιτίας κάποιας αλλαγής του συμβαίνει στο περιβάλλον του. Τα συστήματα που ελέγχουν τη πλοήγηση ενός ρομποτικού συστήματος παρακινούνται συνήθως από θεωρίες της ψυχολογίας οι οποίες εξηγούν με ποιό τρόπο τα έμβια όντα μαθαίνουν διάφορες συμπεριφορές.

1.2 Αντικείμενο Διατριβής

Στη παρούσα διατριβή μελετούμε μεθόδους της ενισχυτικής μάθησης για την αυτόνομη πλοήγηση ρομποτικών συστημάτων. Η Μηχανική Μάθηση στοχεύει στη δημιουργία ευφυών μηχανών οι οποίες αποκτούν και βελτιώνουν τις δεξιότητες τους μέσω της μάθησης και της προσαρμογής. Η έννοια της εκπαίδευσης εντοπίζεται κυρίως πάνω στην εύρεση κατάλληλων τιμών-παραμέτρων των μοντέλων-μηχανών, ώστε να πετύχουμε το καλύτερο ταίριασμα με το εκάστοτε πρόβλημα και το χώρο αναζήτησης. Τρείς είναι οι μορφές εκπαίδευσης που συναντάμε στα προβλήματα Μηχανικής Μάθησης: η μάθηση με επίβλεψη (supervised learning), η μάθηση χωρίς επίβλεψη (unsupervised learning) και η ενισχυτική μάθηση (Reinforcement Learning).

Η Ενισχυτική Μάθηση χρονολογείται από την εποχή που ο άνθρωπος άρχισε να αναπτύσσεται νοητικά και βρίσκει εφαρμογή στη στατιστική, τη ψυχολογία, τη νευροεπιστήμη και την πληροφορική. Είναι μια υπολογιστική προσέγγιση για την κατανόηση και την αυτοματοποίηση της κατευθυνόμενης μάθησης από κάποιο στόχο. Διαφέρει από άλλες υπολογιστικές προσεγγίσεις εξαιτίας της έμφασης που δίνει στη μάθηση μέσω της άμεσης αλληλεπίδρασης με το περιβάλλον, χωρίς να βασίζεται σε κάποια υποδειγματική εποπτεία ή στα πλήρη μοντέλα του περιβάλλοντος. Τα προβλήματα της ενισχυτικής μάθησης χαρακτηρίζονται από μια μακροπρόθεσμη αλληλεπίδραση ανάμεσα σε έναν μαθητευόμενο πράκτορα

(learning agent) και σε ένα δυναμικό, άγνωστο, αβέβαιο και πιθανώς εχθρικό περιβάλλον. Μαθηματικά αυτή η αλληλεπίδραση μοντελοποιείται ως μια Μαρκοβιανή Διαδικασία Απόφασης (ΜΔΑ).

Ο πράκτορας προσπαθεί να επιτύχει ένα στόχο παρά την αβεβαιότητα σχετικά με το περιβάλλον. Οι ενέργειες του πράκτορα επηρεάζουν τη μελλοντική κατάσταση του περιβάλλοντος, ως εκ τούτου επηρεάζουν και τις επιλογές και τις ευκαιρίες που θα παρουσιαστούν στον πράκτορα σε μελλοντικές στιγμές. Η σωστή επιλογή απαιτεί να λάβουμε υπόψη τις έμμεσες, μελλοντικές συνέπειες αυτών των ενεργειών. Την ίδια στιγμή, τα αποτελέσματα αυτών των ενεργειών δεν μπορεί να είναι πλήρως προβλέψιμα. Η επιλογή μιας ενέργειας από τον πράκτορα έχει ως αποτέλεσμα τη μετάβαση του σε μια νέα κατάσταση του περιβάλλοντος. Το περιβάλλον αποκρίνεται σε αυτές τιμωρώντας ή ανταμείβωντας τη συγκεκριμένη επιλογή, δίνοντας στον πράκτορα κάποια ανταμοιβή. Στόχος του πράκτορα είναι η μεγιστοποίηση της συνολικής απόκρισης του περιβάλλοντος. Σε κάθε χρονική στιγμή ο πράκτορας απεικονίζει τις καταστάσεις σε πιθανότητες επιλογής όλων των δυνατών ενεργειών. Αυτή η απεικόνιση ονομάζεται πολιτική του πράκτορα. Οι μέθοδοι ενισχυτικής μάθησης προσδιορίζουν τον τρόπο με τον οποίο ο πράκτορας αλλάζει τη πολιτική του ως αποτέλεσμα της εμπειρίας του. Συγκεκριμένα οι μέθοδοι ενισχυτική μάθησης βασίζονται στην εκτίμηση της συνάρτησης αξίας για την εύρεση της βέλτιστης πολιτικής. Η συνάρτηση αξίας προσδιορίζουν την αξία μιας κατάστασης (ή ενός ζεύγους κατάστασης) ενέργειας υπό μια πολιτική και είναι η αναμενόμενη απολαβή που λαμβάνει ο πράκτορας αν ξεκινήσει από μια κατάσταση και ακολουθήσει τη συγκεκριμένη πολιτική.

Μια από τις σημαντικότερες προκλήσεις που προκύπτουν στη ενισχυτική μάθηση σε αντίθεση με άλλα είδη μάθησης είναι η εξισσορρόπηση ανάμεσα στην εξερεύνηση (exploration) και στην εκμετάλλευση (exploitation). Για να πάρουμε αρκετά καλές ανταμοιβές πρέπει να προτιμήσουμε ενέργειες που δοκιμάστηκαν στο παρελθόν και βρέθηκε πως είναι αποτελεσματικές για την παραγωγή ανταμοιβών. Αλλά για να ανακαλύψουμε τέτοιες ενέργειες, θα πρέπει να δοκιμάσουμε ενέργειες που δεν έχουν επιλεχθεί προηγουμένως. Ο πράκτορας θα πρέπει να εκμεταλλευτεί την ήδη υπάρχουσα γνώση του για να πάρει καλές ανταμοιβές, αλλά θα πρέπει επίσης να εξερευνεί έτσι ώστε να μπορέσει να κάνει καλύτερες επιλογές ενεργειών στο μέλλον.

Τις τελευταίες δυο δεκαετίες έχει παρατηρηθεί μια έκρηξη στην ερευνητική περιοχή της ενισχυτικής μάθησης, με αποτέλεσμα ένα μεγάλο αριθμό νέων αλγορίθμων και αρχιτεκτονικών. Ωστόσο, παραμένουν αρκετά βασικά εμπόδια που δεν επιτρέπουν την ευρεία εφαρμογή της μεθοδολογίας της ενισχυτικής μάθησης σε προβλήματα του πραγματικού κόσμου. Τέτοια προβλήματα που αντιμετωπίζουμε στο πραγματικό κόσμο χαρακτηρίζονται από τα περισσότερα, αν όχι όλα, ακόλουθα χαρακτηριστικά:

- Τεράστιοι ή σχεδόν άπειροι χώροι καταστάσεων και/ή ενεργειών.
- Απαίτηση για απευθείας μάθηση (online learning). Η εξομοιούμενη μάθηση (offline learning) είναι επαρκής στη περίπτωση που το περιβάλλον είναι ολοκληρωτικά σταθερό, ωστόσο αυτή είναι μια αρκετά σπάνια περίπτωση.

- Τα δεδομένα εκπαίδευσης είναι ακριβά. Σε αντίθεση με τη μάθηση με επίβλεψη όπου υπάρχουν μεγάλα σύνολα δεδομένων διαθέσιμα και άμεσα χρησιμοποιούμενα, στην ενισχυτική μάθηση τα δεδομένα εκπαίδευσης παράγονται μέσω της αλληλεπίδρασης του πράκτορα με το δυναμικό σύστημα (πραγματικό ή προσωμοιούμενο) που προσπαθεί να ελέγξει.

- Μερική παρατηρησιμότητα. Σε αρκετά προβλήματα η κατάσταση του συστήματος δεν είναι πλήρως ή άμεσα μετρήσιμη από το χρήστη.

Στη παρούσα διατριβή αρχικά θα ασχοληθούμε με τις τρείς θεμελιώδεις κλάσεις μεθόδων για την επίλυση του προβλήματος της ενισχυτικής μάθησης: τον Δυναμικό Προγραμματισμό (Dynamic Programming), τις μεθόδους Monte Carlo και τη μάθηση Χρονικών Διαφορών (Temporal Difference Learning). Με τον όρο Δυναμικός προγραμματισμός αναφερόμαστε σε μια συλλογή αλγορίθμων που απαιτούν το πλήρες μοντέλο του περιβάλλοντος για τον υπολογισμό βέλτιστων πολιτικών. Συγκεκριμένα μελετούμε δυο από τους πιο γνωστούς αλγορίθμους Δυναμικού Προγραμματισμού, τον αλγόριθμο επανάληψης ως προς την πολιτική (policy iteration) και τον αλγόριθμο επανάληψης ως προς την αξία (value iteration). Ο αλγόριθμος επανάληψης ως προς την αξία σε σχέση με τον αλγόριθμο επανάληψης ως προς τη πολιτική αποφεύγει το βήμα της εκτίμησης πολιτικής το οποίο εισάγει μεγάλο υπολογιστικό κόστος. Αντίθετα με τις μεθόδους Δυναμικού Προγραμματισμού οι μέθοδοι Monte Carlo δεν απαιτούν τη γνώση του μοντέλου του περιβάλλοντος και βασίζονται σε συλλεγόμενη εμπειρία. Η μάθηση Χρονικών Διαφορών αποτελεί συνδυασμό των Monte Carlo μεθόδων και του Δυναμικού Προγραμματισμού. Ένας από τους πιο γνωστούς αλγορίθμους ΧΔ είναι ο αλγόριθμος Sarsa ο οποίος είναι ένας αλγόριθμος εντός πολιτικής (on-policy), δηλαδή ένας αλγόριθμος στον οποίο η πολιτική η οποία αξιολογείται είναι και αυτή που χρησιμοποιούμε για να λάβουμε αποφάσεις. Επίσης θα ασχοληθούμε με ένα από τους βασικότερους μηχανισμούς της ενισχυτικής μάθησης, τα ίχνη επιλεξιμότητας. Τα ίχνη επιλεξιμότητας είναι έναν μηχανισμός καταγραφής γεγονότων όπως η επίσκεψη μιας κατάστασης ή η επιλογή μιας ενέργειας και χρησιμοποιούνται για να αυξήσουν την αποδοτικότητα των μεθόδων χρονικών διαφορών. Στη συνέχεια θα ασχοληθούμε με μια Μπεϋσιανή προσέγγιση εκτίμησης πολιτικής σε γενικούς χώρους καταστάσεων και ενεργειών, η οποία βασίζεται σε στατιστικά γεννητικά μοντέλα μέσω Γκαουσιανών διαδικασιών για τις συναρτήσεις αξίας. Η προσέγγιση αυτή μας παρέχει την εκτίμηση της συνάρτησης αξίας καθώς επίσης και τη διακύμανση αυτής της εκτίμησης. Στη παρούσα διατριβή προτείνεται μια επέκταση της Μπεϋσιανής προσέγγισης η οποία εισάγει την χρήση RVM για την απόκτηση αραιότερων μοντέλων. Αυτό μας δίνει τη δυνατότητα να παράγουμε αποδοτικότερους και αραιότερους απευθείας αλγόριθμους για τη μάθηση καλών προσεγγίσεων της συνάρτησης αξίας. Τέλος, πολλές από τις παραπάνω μεθόδους εφαρμόζονται και αξιολογούνται σε διάφορα πειραματικά περιβάλλοντα αυτόνομης πλοήγησης ρομποτικών συστημάτων. Η σημαντικότερη καινοτομία της παρούσας διατριβής είναι η εφαρμογή και η αξιολόγηση μεθόδων ενισχυτικής μάθησης σε ένα πραγματικό ρομποτικό σύστημα τύπου Pioneer PeopleBot. Με το συγκεκριμένο τρόπο έχουμε τη δυνατότητα να μελετήσουμε τη συμπεριφορά των μεθόδων ενισχυτικής μάθησης σε πρόβλημα του πραγματικού κόσμου.

1.3 Διάρθρωση της Διατριβής

Η παρούσα διατριβή αποτελείται από έξι κεφάλαια. Στο Κεφάλαιο 2 παρουσιάζονται κάποιες εισαγωγικές έννοιες της Ενισχυτικής Μάθησης. Στο Κεφάλαιο 3 περιγράφονται οι βασικές μέθοδοι επίλυσης προβλημάτων Ενισχυτικής Μάθησης. Συγκεκριμένα παρουσιάζονται οι κλάσεις μεθόδων Δυναμικού Προγραμματισμού, Χρονικών Διαφορών και Monte Carlo καθώς επίσης και ο μηχανισμός των ιχνών επιλεξιμότητας. Στο Κεφάλαιο 4 παρουσιάζεται μια Μπεϋσιανή προσέγγιση εκτίμησης πολιτικής σε γενικούς χώρους καταστάσεων και ενεργειών. Αρχικά περιγράφεται μια Μπεϋσιανή προσέγγιση για την εκτίμηση της συνάρτησης αξίας καθώς επίσης και πως αυτή μπορεί να επεκταθεί σε εργασίες με επεισόδια. Επίσης παρουσιάζεται και μια πιο αραιή εκδοχή της συγκεκριμένης μεθόδου. Στη συνέχεια περιγράφονται αλγόριθμοι για τη βελτίωση της πολιτικής βασιζόμενοι στη συγκεκριμένη μέθοδο. Στο κεφάλαιο 5 προτείνεται μια επέκταση της Μπεϋσιανής προσέγγισης για την εκτίμηση της συνάρτησης αξίας. Συγκεκριμένα, συνδυάζουμε τη παραπάνω μέθοδο με το RVM (Relevance Vector Machine) για τη απόκτηση αραιότερων προσεγγίσεων. Τέλος, στο Κεφάλαιο 6 υλοποιούνται και αξιολογούνται κάποιες από τις κυριότερες μεθόδους επίλυσης προβλημάτων ενισχυτικής μάθησης που περιγράφονται στη παρούσα διατριβή σε διάφορα πειραματικά περιβάλλοντα αυτόνομης πλοήγησης ρομποτικών συστημάτων.

Κεφαλαιο 2

ΕΝΙΣΧΥΤΙΚΗ ΜΑΘΗΣΗ

2.1 Εισαγωγή

Η ενισχυτική μάθηση (RL) αντιμετωπίζει το πρόβλημα της εκπαίδευσης μιας βέλτιστης συμπεριφοράς ενός πράκτορα μέσω της αλληλεπίδρασης του με το περιβάλλον. Η ενισχυτική μάθηση παρουσιάζει ομοιότητες με θεωρίες της ψυχολογίας οι οποίες περιγράφουν τον τρόπο με τον οποίο τα έμβια όντα μαθαίνουν να συμπεριφέρονται. Για παράδειγμα όταν ένα βρέφος παίζει, κουνάει τα χέρια του ή κοιτάζει έχει μια άμεση επικοινωνία με το περιβάλλον του. Αναπτύσσοντας αυτή την επικοινωνία παράγεται ένα πλήθος πληροφοριών σχετικά με τις συνέπειες των ενεργειών και με το πως θα πρέπει να ενεργήσουμε έτσι ώστε να πετύχουμε τους στόχους μας. Στόχος του πράκτορα είναι να μεγιστοποιήσει το σήμα ανταμοιβής που λαμβάνει από το περιβάλλον ως αποτέλεσμα των ενεργειών του. Ο πράκτορας σε αντίθεση με το τι συμβαίνει με την εκπαίδευση με δάσκαλο (επιβλεπόμενη μάθηση-supervised learning), δεν γνωρίζει εκ των προτέρων τις ενέργειες που πρέπει να εκτελέσει, αλλά θα πρέπει να ανακαλύψει μόνος του τις ενέργειες που του αποφέρουν τις μεγαλύτερες ανταμοιβές.

Στο παρόν κεφάλαιο παρουσιάζονται βασικές έννοιες της ενισχυτικής μάθησης οι οποίες χρησιμοποιούνται στα επόμενα κεφάλαια. Στην ενότητα 2.2 περιγράφονται τα βασικά στοιχεία της ενισχυτικής μάθησης. Στην ενότητα 2.3 παρουσιάζεται το πλαίσιο της ενισχυτικής

μάθησης. Στην ενότητα 2.4 περιγράφεται ο τρόπος μοντελοποίησης των προβλημάτων της ενισχυτικής μάθησης. Στην ενότητα 2.5 περιγράφονται οι συναρτήσεις αξίας οι οποίες παίζουν κύριο ρόλο για την εκτίμηση και την εύρεση πολιτικών. Τέλος, στην ενότητα 2.6 παρουσιάζεται το θέμα της αντιστάθμισης μεταξύ της εξερεύνησης και της εκμετάλλευσης

2.2 Βασικά Στοιχεία της Ενισχυτικής Μάθησης

Εκτός από τον πράκτορα και το περιβάλλον, μπορούμε να διακρίνουμε τέσσερα ακόμη βασικά στοιχεία ενός συστήματος RL: την πολιτική, τη συνάρτηση ανταμοιβής, τη συνάρτηση αξίας και το μοντέλο του περιβάλλοντος.

Η πολιτική (policy) προσδιορίζει τον τρόπο με τον οποίο ο πράκτορας ενεργεί σε μια δεδομένη στιγμή. Με απλά λόγια, είναι μια απεικόνιση των καταστάσεων του περιβάλλοντος σε ενέργειες, που μπορούν να επιλεχθούν όταν βρισκόμαστε σε αυτές τις καταστάσεις. Η πολιτική είναι ο πυρήνας του πράκτορα RL υπό την έννοια ότι μπορεί να καθορίσει την συμπεριφορά του. Γενικά, οι πολιτικές θα μπορούσαν να είναι στοχαστικές.

Η συνάρτηση ανταμοιβής (reward function) ορίζει το στόχο σε ένα πρόβλημα RL. Απεικονίζει κάθε κατάσταση (ή ζευγάρι κατάστασης-ενέργειας) του περιβάλλοντος σε έναν αριθμό, την ανταμοιβή, ο οποίος εκφράζει κατά πόσο είναι επιθυμητή η συγκεκριμένη κατάσταση. Ο μοναδικός στόχος ενός πράκτορα είναι να μεγιστοποιήσει τη συνολική ανταμοιβή που θα λάβει μακροπρόθεσμα. Η συνάρτηση ανταμοιβής ορίζει τι είναι καλό και τι κακό για έναν πράκτορα. Είναι σημαντικό να σημειωθεί, πως η συνάρτηση ανταμοιβής θα πρέπει να μη μεταβάλλεται από τον πράκτορα.

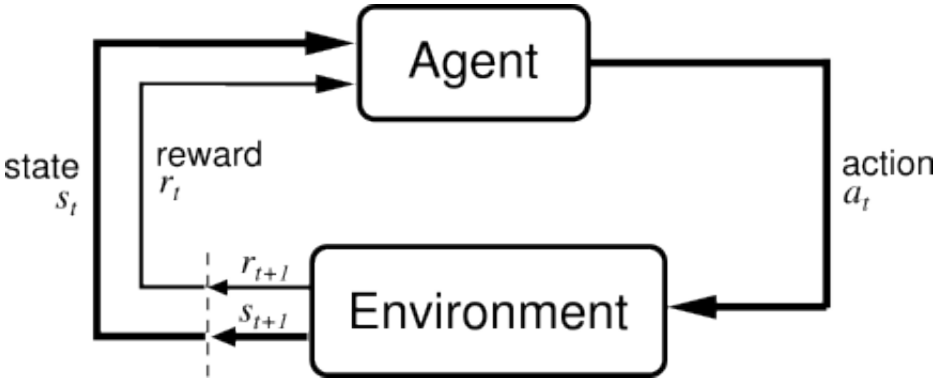
Η συνάρτηση αξίας (value function) είναι η συνολική ανταμοιβή που αναμένεται να λάβει ένας πράκτορας στο μέλλον, ξεκινώντας από τη συγκεκριμένη κατάσταση. Ενώ η ανταμοιβή καθορίζει την άμεση, πραγματική καταλληλότητα των καταστάσεων του περιβάλλοντος, η αξία δείχνει την μακροπρόθεσμη καταλληλότητα των καταστάσεων αφού λαμβάνει υπόψη τις καταστάσεις που είναι πιθανόν να ακολουθήσουν, και τις διαθέσιμες ανταμοιβές των καταστάσεων αυτών. Για παράδειγμα, μια κατάσταση μπορεί να αποφέρει μια μικρή άμεση ανταμοιβή αλλά ταυτόχρονα να έχει μεγάλη αξία επειδή μόνιμα ακολουθείται από καταστάσεις που αποφέρουν μεγάλες ανταμοιβές (ή μπορεί να συμβεί το ανάποδο). Αντίθετα με τις ανταμοιβές που δίνονται απευθείας από το περιβάλλον, οι αξίες των καταστάσεων πρέπει να εκτιμηθούν και να επανεκτιμηθούν από τις ακολουθίες των παρατηρήσεων που κάνει ένας πράκτορας σε όλη την διάρκεια της ζωής του.

Το μοντέλο (model) του περιβάλλοντος δοθέντος της τρέχουσας κατάστασης και μίας ενέργειας, προβλέπει τη διάδοχη κατάσταση και την ανταμοιβή. Ουσιαστικά μιμείται την συμπεριφορά του περιβάλλοντος. Συνήθως, ο πράκτορας δεν έχει καμία γνώση του μοντέλου του περιβάλλοντος.

2.3 Το πλαίσιο της Ενισχυτικής Μάθησης

Το πρόβλημα της ενισχυτικής μάθησης σχετίζεται με το πως ένας πράκτορας μπορεί να μάθει μια συμπεριφορά μέσω της αλληλεπίδρασης του με το περιβάλλον για την επίτευξη ενός στόχου. Ο πράκτορας αλληλεπιδρά συνεχώς με το περιβάλλον, επιλέγει ενέργειες και το περιβάλλον αποκρίνεται σε αυτές ανταμείβοντας ή τιμωρώντας την τρέχουσα επιλογή. Στόχος είναι η μεγιστοποίηση της συνολικής απόκρισης του περιβάλλοντος.

Ο πράκτορας και το περιβάλλον αλληλεπιδρούν σε κάθε χρονική στιγμή, t, με τον πράκτορας να λαμβάνει μια αναπαράσταση της κατάστασης του περιβάλλοντος, $s_t \in \mathcal{S}$, όπου $\mathcal{S}$ είναι το σύνολο όλων των δυνατών καταστάσεων. Με βάση την τρέχουσα κατάσταση επιλέγει μία ενέργεια, $a_t \in \mathcal{A}(s_t)$, όπου $\mathcal{A}(s_t)$ είναι το σύνολο των διαθέσιμων ενεργειών στην κατάσταση s_t. Την επόμενη χρονική στιγμή $(t+1)$, ο πράκτορας λαμβάνει από το περιβάλλον μια ανταμοιβή, $r_{t+1} \in \Re$, ως συνέπεια της ενέργειας του και μεταβαίνει σε μία καινούρια κατάσταση, s_{t+1}. Στο Σχήμα 2.1 φαίνεται η αλληλεπίδραση μεταξύ του πράκτορα και του περιβάλλοντος.



Σχήμα 2.1: Το πλαίσιο της ενισχυτικης μάθησης

Σε κάθε χρονική στιγμή, ο πράκτορας απεικονίζει τις καταστάσεις σε πιθανότητες επιλογής όλων των δυνατών ενεργειών. Αυτή η απεικόνιση ονομάζεται πολιτική (policy) του πράκτορα και συμβολίζεται ως π_t, όπου $\pi_t(s,a)$ είναι η πιθανότητα ο πράκτορας την χρονική στιγμή t να επιλέξει την ενέργεια $a_t = a$ αν βρίσκεται στην κατάσταση $s_t = s$. Οι μέθοδοι ενισχυτικής μάθησης προσδιορίζουν τον τρόπο με τον οποίο ο πράκτορας αλλάζει την πολιτική του ως αποτέλεσμα της εμπειρίας του. Ο στόχος του πράκτορα είναι να μεγιστοποιήσει τη συνολική ανταμοιβή που λαμβάνει μακροπρόθεσμα.

Το συγκεκριμένο πλαίσιο είναι αφηρημένο και ευέλικτο, έτσι μπορεί να εφαρμοστεί σε πολλά διαφορετικά προβλήματα και με πολλούς διαφορετικούς τρόπους. Για παράδειγμα, οι χρονικές στιγμές δεν χρειάζεται να αναφέρονται σε πραγματικό χρόνο αλλά μπορεί να αναφέρονται σε διαδοχικά στάδια λήψης απόφασης και δράσης.

Αυτό το οποίο προτείνει το πλαίσιο της ενισχυτικής μάθησης είναι ότι οποιοδήποτε πρόβλημα μάθησης συμπεριφοράς, οδηγούμενο από κάποιο συγκεκριμένο στόχο, μπορεί να αναχθεί σε τρία σήματα τα οποία ανταλλάσσονται μεταξύ του πράκτορα και του περιβάλλοντος ένα σήμα για να αναπαρασταθούν οι επιλογές του πράκτορα (ενέργειες), ένα

σήμα για να αναπαρασταθεί η πληροφορία στην οποία βασίζονται οι επιλογές των ενεργειών του πράκτορα (καταστάσεις), και ένα σήμα για να προσδιοριστεί ο στόχος του πράκτορα (ανταμοιβή).

2.4 Μοντελοποίηση Προβλημάτων Ενισχυτικής Μάθησης

Στη γενική περίπτωση του προβλήματος ενισχυτικής μάθησης, οι ενέργειες που εκτελεί ο πράκτορας δεν καθορίζουν μόνο την άμεση ανταμοιβή που λαμβάνει, αλλά επίσης και την επόμενη κατάσταση του περιβάλλοντος. Τέτοια περιβάλλον μπορούν να εκληφθούν ως δίκτυα, όπου ο πράκτορας πρέπει να λαμβάνει υπόψη τόσο την επόμενη κατάσταση όσο και την άμεση ανταμοιβή, για να αποφασίσει ποιά ενέργεια πρέπει να επιλέξει. Για το λόγο αυτό, το περιβάλλον των προβλημάτων ενισχυτικής μάθησης μοντελοποιείται ως Μαρκοβιανή Διαδικασία Απόφασης (Markov Decision Process). Μία Μαρκοβιανή Διαδικασία Απόφασης περιγράφεται από μία τετράδα $\langle S, A, T, R\rangle$ όπου,

- S το πεπερασμένο σύνολο όλων των καταστάσεων,
- A το πεπερασμένο σύνολο των ενεργειών,
- $T : S \times A \mapsto Pr(S)$ είναι η συνάρτηση μετάβασης (transition function), όπου $Pr(S)$ είναι μία κατανομή πιθανοτήτων στό σύνολο καταστάσεων S, η οποία δεδομένης μίας κατάστασης και μίας ενέργειας μας επιστρέφει τις πιθανότητες μετάβασης σε κάθε πιθανή επόμενη κατάσταση. Γράφουμε $\mathcal{T}^a_{SS'}$ για την πιθανότητα μετάβασης από την κατάσταση s στην κατάσταση s' εκτελώντας την ενέργεια a,
- $R : S \times A \times S \mapsto \Re$ είναι η συνάρτηση ανταμοιβής, η οποία καθορίζει την επόμενη αναμενόμενη ανταμοιβή ως συνάρτηση της τρέχουσας κατάστασης και ενέργειας, καθώς και της επόμενης κατάστασης. Γράφουμε $\mathcal{R}^a_{SS'}$ για την αναμενόμενη ανταμοιβή που θα παρούμε, αν στην κατάσταση s επιλέξουμε την ενέργεια a και μεταβούμε στην κατάσταση s'.

Για να είναι το μοντέλο Μαρκοβιανό θα πρέπει να ισχύει η ιδιότητα Markov, η οποία ορίζει ότι η απόκριση του περιβάλλοντος τη χρονική στιγμή $t+1$ εξαρτάται μόνο από την αναπαράσταση της κατάστασης και της ενέργειας τη χρονική στιγμή t και όχι από όλες τις προηγούμενες χρονικές στιγμές, δηλαδή

$$Pr\{s_{t+1} = s', r_{t+1} = r \mid s_t, a_t\} = Pr\{s_{t+1} = s', r_{t+1} = r | s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_0, a_0\}. \tag{2.1}$$

2.5 Συναρτήσεις Αξίας

Όπως προαναφέρθηκε, στόχος του πράκτορα είναι να μεγιστοποιήσει την συνολική ανταμοιβή πού λαμβάνει μακροπρόθεσμα. Γενικά, ο πράκτορας θέλει να μεγιστοποιήσει την

αναμενόμενη απολαβή (expected return), όπου η απολαβή, R_t, ορίζεται ως μία ειδική συνάρτηση της ακολουθίας των ανταμοιβών. Αν η σειρά τών ανταμοιβών πού λαμβάνονται μετά τη χρονική στιγμή t συμβολίζεται ως $r_{t+1}, r_{t+2}, r_{t+3}, \ldots$, ο πράκτορας στην απλούστερη περίπτωση θέλει να μεγιστοποιήσει το παρακάτω άθροισμα των ανταμοιβών:

$$D_t = r_{t+1} + r_{t+2} + r_{t+3} + \cdots + r_T, \tag{2.2}$$

όπου T είναι το τελικό χρονικό βήμα. Αυτή η προσέγγιση έχει νόημα σε εφαρμογές που υπάρχει η έννοια του τελικού χρονικού βήματος, δηλαδή σε εργασιές (tasks) που εκτελούνται σε επεισόδια. Η ακολουθία των ενεργειών που εκτελούνται από κάποιον πράκτορα για να φθάσει από μια αρχική κατάσταση σε μια τελική είναι ένα επεισόδιο (episode). Κάθε επεισόδιο τερματίζει σε μια ειδική κατάσταση που ονομάζεται τερματική κατάσταση (terminal state), όπου όλες οι ενέργειες οδηγούν στην ίδια κατάσταση λαμβάνοντας μηδενική ανταμοιβή. Στη συνέχεια, ο πράκτορας μεταβαίνει στην αρχική κατάσταση ή σε κάποια από τις αρχικές καταστάσεις με την ίδια πιθανότητα και ξεκινά ένα καινούργιο επεισόδιο. Σε προβλήματα όπου ο πράκτορας παίρνει κάποια ανταμοιβή μόνο όταν φθάνει στην τελική κατάσταση, είναι βολικό να θεωρηθεί η τελική κατάσταση ως στόχος. Μερικές φορές είναι αναγκαίο να διακρίνουμε το σύνολο όλων των μη τερματικών καταστάσεων, συμβολίζοντας το ως S, από το σύνολο όλων των καταστάσεων, συμβολίζοντας αυτό ως S^+.

Αντίθετα, σε πολλές περιπτώσεις η αλληλεπίδραση του πράκτορα με το περιβάλλον δεν διακόπτεται σε αναγνωρίσιμα επεισόδια, αλλά συνεχίζεται χωρίς κάποιο περιορισμό επ'άπειρον. Ονομάζουμαι αυτή την εργασία, συνεχόμενη (continuing task). Ο ορισμός της απολαβής 2.2 είναι προβληματικός για συνεχόμενες εργασίες διότι το τελικό χρονικό βήμα θα είναι $T = \infty$, και η απολαβή, που είναι αυτό πού προσπαθούμε να μεγιστοποιήσουμε, μπορεί εύκολα να γίνει ίση με το άπειρο από μόνη της. Για τον λόγο αυτό, συνήθως, χρησιμοποιούμε ένα ορισμό για την απολαβή που είναι ελαφρός πιο περίπλοκος εννοιολογικά αλλά απλούστερος μαθηματικά. Η επιπλέον ίδεα πού εισάγουμε είναι αυτή της έκπτωσης (discount). Σύμφωνα με αυτή τη προσέγγιση, ο πράκτορας προσπαθεί να επιλέξει ενέργειες έτσι ώστε το άθροισμα των εκπτώμενων ανταμοιβών πού λαμβάνει να μεγιστοποιείται. Συγκεκριμένα, επιλέγει την ενέργεια a_t ετσί ώστε να μεγιστοποιήσει την αναμενόμενη εκπτώμενη απολαβή (discount return):

$$D_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \tag{2.3}$$

όπου $0 \leq \gamma \leq 1$ είναι ο ρυθμός έκπτωσης (discount rate).

Ο ρυθμός έκπτωσης καθορίζει την παρούσα αξία των μελλοντικών ανταμοιβών: μία ανταμοιβή πού λαμβάνεται k χρονικά βήματα στο μέλλον αξίζει μόνο γ^{k-1} φορές σε σχέση με την αξία πού θα έχει αν ληφθεί άμεσα. Εάν $\gamma < 1$, το άπειρο άθροισμα έχει μία πεπερασμένη αξία όσο η ακολουθία των ανταμοιβών r_k είναι οριοθετημένη. Εάν $\gamma = 0$, ο πράκτορας ενδιαφέρεται μόνο να μεγιστοποιήσει τις άμεσες ανταμοιβές: στόχος του είναι να μάθει πώς θα επιλέξει την ενέργεια a_t έτσι ώστε να μεγιστοποιήσει μόνο το r_{t+1}. Καθώς το γ προσεγγίζει το 1, οι μελλοντικές ανταμοιβές λαμβάνονται περισσότερο υπόψη με αποτέλεσμα ο πράκτορας να γίνεται περισσότερο προνοητικός.

Η αξία μίας κατάστασης s υπό μία πολιτική π, συμβολίζεται ως $V^\pi(s)$ και είναι η αναμενόμενη (μέση) απολαβή αν ο πράκτορας ξεκινήσει από τη κατάσταση s και ακολουθήσει τη πολιτική π. Για Μαρκοβιανές Διαδικασίες Απόφασης (ΜΔΑ), μπορούμε να ορίσουμε τη συνάρτηση $V^\pi(s)$ ως:

$$V^\pi(s) = E_\pi\{D_t \mid s_t = s\} = E_\pi\Big\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \mid s_t = s\Big\}, \tag{2.4}$$

όπου το $E_\pi\{\}$ υποδηλώνει την μέση τιμή δοθέντος ότι ο πράκτορας ακολουθεί τη πολιτική π. Η αξία της τερματικής κατάστασης είναι πάντα μηδέν. Η συνάρτηση $V^\pi(s)$ ονομάζεται, συνάρτηση αξίας κατάστασης (state-value function).

Παρόμοια, ορίζουμε την $Q^\pi(s,a)$ για τη λήψη μίας ενέργειας a στη κατάσταση s υπό μια πολιτική π ως την αναμενόμενη απολαβή ξεκινώντας από την κατάσταση s, επιλέγοντας την ενέργεια a, και ακολουθώντας στη συνέχεια τη πολιτική π:

$$Q^\pi(s,a) = E_\pi\{D_t \mid s_t = s, a_t = a\} = E_\pi\Big\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \mid s_t = s, a_t = a\Big\}. \tag{2.5}$$

Ονομάζουμε τη συνάρτηση $Q^\pi(s,a)$ ως συνάρτηση αξίας κατάστασης-ενέργειας (action-value function).

Οι συναρτήσεις αξίας $V^\pi(s)$ και $Q^\pi(s,a)$ μπορούν να υπολογιστούν εμπειρικά. Για παράδειγμα, αν ένας πράκτορας ακολουθεί μία πολιτική π και διατηρεί έναν μέσο όρο, για κάθε κατάσταση που συναντά, των πραγματικών απολαβών που αντιστοιχούν σε αυτή τη κατάσταση, τότε ο μέσος όρος θα συγκλίνει στην αξία της κατάστασης, $V^\pi(s)$, καθώς ο αριθμός των επισκέψεων στην κατάσταση αυτή τείνει στο άπειρο. Εάν επιπλέον διατηρούνται μέσοι όροι για κάθε ενέργεια που εκτελείται σε κάθε κατάσταση, τότε αυτοί οι μέσοι όροι θα συγκλίνουν στις αξίες κατάστασης-ενέργειας, $Q^\pi(s,a)$. Ονομάζουμε τις παραπάνω μεθόδους, Monte Carlo μεθόδους. Εάν το πλήθος των καταστάσεων είναι μεγάλο, δεν είναι πρακτικό να κρατάμε μέσους όρους για κάθε κατάσταση (ή ζεύγος κατάστασης-ενέργειας) χωριστά. Αντίθετα, ο πράκτορας θα πρέπει να διατηρεί τις $V^\pi(s)$ και $Q^\pi(s,a)$ ως παραμετροποιημένες συναρτήσεις και να προσαρμόζει τις παραμέτρους έτσι ώστε να ταιριάζουν καλύτερα στις παρατηρούμενες απολαβές.

Μια θεμελιώδης ιδιότητα των συναρτήσεων αξίας είναι ότι ικανοποιούν συγκεκριμένες επαναληπτικές σχέσεις. Για μία οποιαδήποτε πολιτική π και κατάσταση s, διατηρείται η ακόλουθη σχέση ανάμεσα στην αξία της s και της αξίας της πιθανής διάδοχης κατάστασης:

$$\begin{aligned}
V^\pi(s) &= E_\pi\Big\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \mid s_t = s\Big\} \\
&= E_\pi\Big\{r_{t+1} + \gamma\sum_{k=0}^{\infty}\gamma^k r_{t+k+2} \mid s_t = s\Big\} \\
&= \sum_a \pi(s,a)\sum_{s'}\mathcal{T}^a_{SS'}\Big[\mathcal{R}^a_{SS'} + \gamma E_\pi\Big\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+2} \mid s_{t+1} = s'\Big\}\Big] \\
&= \sum_a \pi(s,a)\sum_{s'}\mathcal{T}^a_{SS'}[\mathcal{R}^a_{SS'} + \gamma V^\pi(s')].
\end{aligned} \tag{2.6}$$

Η εξίσωση 2.6 είναι η εξίσωση Bellman για την συνάρτηση αξίας κατάστασης V^{π}(s), η οποία δηλώνει ότι η αξία της κατάστασης πρέπει να ισούται με την (μειωμένη) αξία της αναμενόμενης επόμενης κατάστασης, συν τη προσδοκόμενη ανταμοιβή.

Η λύση ενός προβλήματος ενισχυτικής μάθησης σημαίνει, την εύρεση μίας πολιτικής που επιτυγχάνει τη λήψη καλύτερων μακροπρόθεσμων ανταμοιβών. Για πεπερασμένες ΜΔΑ, μπορούμε να ορίσουμε τη βέλτιστη πολιτική με τον ακόλουθο τρόπο. Μια πολιτική π είναι καλύτερη ή ίση από μια πολιτική π', εάν η μέση απολαβή της, είναι μεγαλύτερη ή ίση σε σχέση με αυτή της π' για όλες τις καταστάσεις. Δηλαδή, $\pi \geq \pi'$ αν και μόνο αν $V^{\pi}(s) \geq V^{\pi'}(s) \quad , \forall s \in \mathcal{S}$. Υπάρχει τουλάχιστον μία πολιτική η οποία είναι καλύτερη ή ίση, σε σχέση με τις υπόλοιπες πολιτικές. Αυτή είναι η βέλτιστη πολιτική (optimal policy). Αν και μπορεί να υπάρχουν περισσότερες από μία, συμβολίζουμε όλες τις βέλτιστες πολιτικές ως π^*. Μοιράζονται την ίδια συνάρτηση αξίας κατάστασης, που ονομάζεται βέλτιστη συνάρτηση αξίας κατάστασης (optimal state-value function), V^*, και ορίζεται ως εξής:

$$V^*(s) = \max_{\pi} V^{\pi}(s) \qquad \text{για κάθε } s \in \mathcal{S}. \tag{2.7}$$

Οι βέλτιστες πολιτικές επίσης μοιράζονται την ίδια συνάρτηση αξίας κατάστασης-ενέργείας, Q^*, η οποία ορίζεται ως εξής:

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \qquad \text{για κάθε } s \in \mathcal{S} \text{ και } a \in \mathcal{A}(s). \tag{2.8}$$

Για κάθε ζεύγος κατάστασης-ενέργειας (s,a), η συνάρτηση αυτή δίνει την αναμενόμενη απολαβή που λαμβάνουμε αν στην κατάσταση s επιλέξουμε την ενέργεια a και στη συνέχεια ακολουθήσουμε την βέλτιστη πολιτική. Έτσι, μπορούμε να γράψουμε την Q^* σε σχέση με την V^* ως εξής:

$$Q^*(s,a) = E\{r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, a_t = a\}. \tag{2.9}$$

Οι εξισώσεις Bellman για τις V^* και Q^*, ονομάζονται βέλτιστες εξισώσεις Bellman. Η βέλτιστη εξίσωση Bellman για την V^* υπολογίζεται ως εξής:

$$\begin{aligned}
V^*(s) &= \max_{a \in \mathcal{A}(s)} Q^{\pi^*}(s,a) \\
&= \max_a E_{\pi^*}\{D_t \mid s_t = s, a_t = a\} \\
&= \max_a E_{\pi^*}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a \right\} \\
&= \max_a E_{\pi^*}\left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s, a_t = a \right\} \\
&= \max_a E\{r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, a_t = a\} \\
&= \max_a \sum_{s'} \mathcal{T}^a_{SS'}[\mathcal{R}^a_{SS'} + \gamma V^*(s')].
\end{aligned} \tag{2.10}$$

Οι δύο τελευταίες εξισώσεις είναι δύο μορφές της βέλτιστης εξίσωσης Bellman για την V^*. Η βέλτιστη εξίσωση Bellman για την Q^* είναι η εξής:

$$Q^*(s,a) = E\{r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a^{'}) \mid s_t = s, a_t = a\}$$
$$= \sum_{s'} \mathcal{T}^a_{SS'} \Big[\mathcal{R}^a_{SS'} + \gamma \max_{a'} Q^*(s', a') \Big]. \tag{2.11}$$

Για πεπερασμένες ΜΔΑ, η βέλτιστη συνάρτηση Bellman 2.10 έχει μία μοναδική λύση ανεξάρτητα από την πολιτική. Η βέλτιστη συνάρτηση Bellman είναι ένα σύστημα εξισώσεων, μια για κάθε κατάσταση. Έτσι αν υπάρχουν N καταστάσεις, τότε υπάρχουν N εξισώσεις με N αγνώστους. Αν το μοντέλο του περιβάλλοντος είναι γνωστό ($\mathcal{R}^a_{SS'}$ και $\mathcal{T}^a_{SS'}$) , τότε κάποιος μπορεί να λύσει το σύστημα εξισώσεων για την V^* χρησιμοποιώντας μία οποιαδήποτε μέθοδο επίλυσης μη γραμμικών εξισώσεων. Το ίδιο μπορεί να συμβεί και για την Q^*.

Εφόσον γνωρίζουμε την V^*, είναι σχετικά εύκολο να ορίσουμε τη βέλτιστη πολιτική. Για κάθε κατάσταση s, υπάρχουν μία ή περισσότερες ενέργειες που μας δίνουν τη μέγιστη τιμή στη βέλτιστη συνάρτηση Bellman. Μια οποιαδήποτε πολιτική που δίνει μη μηδενική πιθανότητα μόνο σε αυτές τις ενέργειες είναι βέλτιστη πολιτική. Εάν έχουμε τη βέλτιστη συνάρτηση αξίας, V^*, τότε οι ενέργειες που εμφανίζονται μετά από αναζήτηση ενός βήματος ως καλύτερες, θα είναι οι βέλτιστες ενέργειες. Αντίθετα η γνώση της Q^* κάνει την επιλογή των βέλτιστων ενεργειών ακόμη πιο εύκολη. Για κάθε κατάσταση s, μπορεί απλά να βρει οποιαδήποτε ενέργεια πού μεγιστοποιεί την $Q^*(s,a)$, χωρίς να κοιτάζει ένα βήμα μπροστά. Με τον τρόπο αυτό, η βέλτιστη συνάρτηση αξίας κατάστασης-ενέργειας μας επιτρέπει να επιλέγουμε τις βέλτιστες ενέργειες χωρίς να χρειάζεται να ξέρουμε τίποτα σχετικά με τις πιθανές διαδοχικές καταστάσεις και τις αξίες τους, δηλαδή χωρίς καμία γνώση του μοντέλου του περιβάλλοντος.

2.6 Εξερεύνηση και Εκμετάλλευση

Μια από τις προκλήσεις πού προκύπτουν στην ενισχυτική μάθηση σε αντίθεση με άλλα είδη μάθησης είναι η εξισορρόπηση (trade-off) μεταξύ εξερεύνησης (exploration) και εκμετάλλευσης (exploitation). Για την απόκτηση μεγαλύτερων ανταμοιβών, ο πράκτορας θα πρέπει να επιλέγει ενέργειες που επιλέχτηκαν στο παρελθόν και αποδείχτηκαν αποτελεσματικές στη παραγωγή ανταμοιβών. Αλλά για να ανακαλύψουμε τις ενέργειες αυτές, θα πρέπει να δοκιμάσουμε ενέργειες που δεν έχουν επιλεχθεί ακόμη. Ο πράκτορας πρέπει να εκμεταλλευτεί την ηδή αποκτηθείσα γνώση του για να πάρει καλές απολαβές, και επίσης πρέπει να εξερευνεί έτσι ώστε να κάνει καλύτερες επιλογές ενεργειών στο μέλλον. Το θέμα πού ανακύπτει είναι με ποιό τρόπο θα επιτύχουμε εξισορρόπηση μεταξύ της εξερεύνησης και της εκμετάλλευσης. Δηλαδή, κατά πόσο ο πράκτορας πρέπει να επιλέγει τύχαιες ενέργειες έτσι ώστε να μπορέσει να επισκεφθεί καινούριες (τυχόν καλύτερες) καταστάσεις ή να αξιοποιήσει την ήδη υπάρχουσα γνώση του ώστε να μεγιστοποιήσει την απολαβή του. Η εξερεύνηση είναι αρκετά πιθανό να οδηγήση τον πράκτορα σε καταστάσεις που θα του αποφέρουν ακόμη

μεγαλύτερες απολαβές σε σχέση με την εκμετάλλευση. Παρολά αυτά όμως, ο πράκτορας δεν είναι δυνατόν να κάνει εξερεύνηση επ'άπειρον διότι θα πρέπει να αξιοποιήσει την γνώση που έχει αποκτηθεί έως τώρα για να μεγιστοποιήσει τις απολαβές του. Αντίθετα, η αμιγής εκμετάλλευση μπορεί να οδηγήσει σε αποτελμάτωση. Έχει προταθεί μία μεγάλη ποικιλία μεθόδων επίλυσης του συγκεκριμένου προβλήματος, αλλά στη συνέχεια παρουσιάζονται δύο από τις σημαντικότερες μεθόδους για την αντιστάθμιση μεταξύ της εξερεύνησης και της εκμετάλλευσης.

2.6.1 ϵ-greedy επιλογή ενέργειας

Ο απλούστερος τρόπος επιλογής μίας ενέργειας είναι να επιλέξουμε την ενέργεια με την μεγαλύτερη εκτιμώμενη αξία, a^*, για την οποία $Q_t(s, a^*) = \max_a Q_t(s, a)$. Η συγκεκριμένη μέθοδος εκμεταλλεύεται πλήρως την αποκτηθείσα γνώση για να μεγιστοποιήσει την άμεση ανταμοιβή ενώ παράλληλα δε χρονοτριβεί ψάχνοντας για πιθανές καλύτερες ενέργειες. Μια απλή εναλλακτική είναι με πιθανότητα ϵ να επιλέξουμε ομοιόμορφα μία ενέργεια, από όλες τις ενέργειες (Εξερεύνηση) ενώ με πιθανότητα $1 - \epsilon$ να επιλέξουμε την ενέργεια με την μεγαλύτερη εκτιμώμενη αξία κατάστασης-ενέργειας (Εκμετάλλευση). Η παραπάνω μέθοδος ονομάζεται ϵ-greedy. Το κύριο πλεονέκτημα των μεθόδων αυτών είναι πως κάθε ενέργεια θα επιλεχθεί τουλάχιστον μία φορά.

2.6.2 Softmax επιλογή ενέργειας

Αν και η ϵ-greedy μέθοδος επιλογής ενέργειας είναι ένας αποτελεσματικός και δημοφιλής τρόπος εξισορρόπησης της εξερεύνησης και της αξιοποίησης στην ενισχυτική μάθηση, ένα μειονέκτημα της είναι πως κατά την εξερεύνηση επιλέγει ισοπίθανα όλες τις ενέργειες. Αυτό σημαίνει πως είναι εξίσου πιθανό να επιλέξει τόσο την χειρότερη ενέργεια όσο και την καλύτερη ενέργεια. Η πιο εμφανής λύση είναι να παραστήσουμε τις πιθανότητες επιλογής ενέργειας σαν μία συνάρτηση της εκτιμώμενης αξίας τους. Η μεγαλύτερη πιθανότητα επιλογής εξακολουθεί να δίνεται στην καλύτερη ενέργεια, ενώ όλες οι άλλες ταξινομούνται και σταθμίζονται σύμφωνα με την εκτιμώμενη αξία τους. Αυτοί οι κανόνες ονομάζονται softmax κανόνες επιλογής ενέργειας (Softmax action selection rules). Η πιο γνωστή softmax μέθοδος χρησιμοποιεί μία Gibbs, ή Boltzmann, κατανομή. Η συγκεκριμένη μέθοδος, επιλέγει της ενέργεια a με πιθανότητα:

$$P(a \mid s) = \frac{\exp^{\frac{Q_t(s,a)}{T}}}{\sum_{b=1}^{n} \exp^{\frac{Q_t(s,b)}{T}}}, \tag{2.12}$$

όπου T είναι μια θετική παράμετρος πού ονομάζεται θερμοκρασία (temperature). Υψηλές θερμοκρασίες κάνουν τις ενέργειες ισοπίθανες. Αντίθετα, χαμηλές θερμοκρασίες προκαλούν την αύξηση της διαφοράς στις πιθανότητες επιλογής των ενεργειών με διαφορετικές εκτιμώμενες αξίες. Καθώς $T \to 0$, η softmax επιλογή ενέργειας γίνεται ίδια με την άπληστη (greedy) επιλογή ενέργειας.

Κεφαλαιο 3

ΜΕΘΟΔΟΙ ΕΠΙΛΥΣΗΣ ΠΡΟΒΛΗΜΑΤΩΝ ΕΝΙΣΧΥΤΙΚΗΣ ΜΑΘΗΣΗΣ

3.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζονται οι τρείς θεμελιώδεις κλάσεις μεθόδων για την επίλυση του προβλήματος της ενισχυτικής μάθησης: ο δυναμικός προγραμματισμός (dynamic programming), οι μέθοδοι Monte Carlo (Monte Carlo methods), και η μάθηση χρονικών διαφορών (temporal-difference learning). Ολές αυτές οι μέθοδοι επιλύουν πλήρως το πρόβλημα της ενισχυτικής μάθησης.

Κάθε κλάση μεθόδων έχει τα πλεονεκτήματα και τις αδυναμίες της. Οι μέθοδοι δυναμικού προγραμματισμού είναι καλά μαθηματικά αναπτυγμένες, αλλά απαιτούν ένα πλήρες και ακριβές μοντέλο του περιβάλλοντος. Οι μέθοδοι Monte Carlo δε χρειάζονται το μοντέλο του περιβάλλοντος και είναι εννοιολογικά απλές, αλλά δεν είναι κατάλληλες γία να υλοποιηθούν επαυξητικά (incremental). Οι μέθοδοι χρονικών διαφορών δεν απαιτούν την γνώση του μοντέλου και είναι πλήρως επαυξητικές, αλλα είναι αρκετά περίπλοκες για να αναλυθούν. Επίσης οι μέθοδοι διαφέρουν στην αποδοτικότητα και στη ταχύτητα σύγκλισης τους.

Στη συνέχεια παρουσιάζονται τα ίχνη επιλεξιμότητας τα οποία χρησιμοποιούνται για την καταγραφή γεγονότων. Τέλος θα αναφερθούμε στο μηχανισμό γενίκευσης δηλαδή τον τρόπο προσέγγισης μιας συνάρτησης μεσώ παραδειγμάτων που λαμβάνονται.

3.2 Δυναμικός Προγραμματισμός

Με τον όρο δυναμικός προγραμματισμός (ΔΠ) [9], [15] αναφερόμαστε σε μία συλλογή αλγορίθμων που μπορούν να χρησιμοποιηθούν για να υπολογίσουμε βέλτιστες πολιτικές δοθέντος του μοντέλου του περιβάλλοντος. Οι κλασσικοί αλγόριθμοι ΔΠ είναι ένα περιορισμένο εργαλείο στην ενισχυτική μάθηση τόσο εξαιτίας της παραδοχής ενός τέλειου μοντέλου και εξαιτίας του υψηλού υπολογιστικού κόστους, αλλά αποτελούν τη θεωρητική βάση των μεθόδων ενισχυτικής μάθησης. Η κύρια ιδέα στίς μεθόδους ΔΠ είναι η χρήση των συναρτήσεων αξίας για την αναζήτηση καλών πολιτικών. Στη συνέχεια παρουσιάζονται δύο από τους πιο γνωστούς αλγόριθμους ΔΠ, ο αλγόριθμος επανάληψης ως προς τη πολιτική (policy iteration) και ο αλγόριθμος επανάληψης ως πρός την αξία (value iteration).

3.2.1 Επανάληψη ως προς την πολιτική

Ο αλγόριθμος επανάληψη ως προς τη πολιτική [9], [15] αποτελείται από δύο ταυτόχρονες, αλληλεπιδρόμενες διεργασίες, μια που υπολογίζει την συνάρτηση αξίας υπό την τρέχουσα πολιτική (policy evalution), και μια που πραγματοποιεί τη βελτίωση της πολιτικής ως προς τη τρέχουσα συνάρτηση αξίας (policy improvement). Οι δύο αυτές διεργασίες εναλλάσονται μεταξύ τους, ενώ η κάθε μία ολοκληρώνεται πριν ξεκινήσει η επόμενη (χωρίς αυτό να είναι πάντα απαραίτητο), μέχρι να υπάρξει σύγκλιση.

Η εκτίμηση της πολιτικής π επιτυγχάνεται με την εύρεση της συνάρτησης αξίας κατάστασης V^π για κάθε κατάσταση υπό την τρέχουσα πολιτική. Όπως προαναφέρθηκε στο Κεφάλαιο 2, η συνάρτηση αξίας κατάστασης V^π υπολογίζεται χρησιμοποιώντας την εξίσωση Bellman:

$$\begin{aligned} V^\pi(s) &= E_\pi\{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots \mid s_t = s\} \\ &= E_\pi\{r_{t+1} + \gamma V^\pi(s_{t+1}) \mid s_t = s\} \\ &= \sum_a \pi(s,a) \sum_{s'} \mathcal{T}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V^\pi(s')], \end{aligned} \tag{3.1}$$

για κάθε $s \in \mathcal{S}$. Εάν τα δυναμικά του περιβάλλοντος είναι γνωστά, τότε η Εξίσωση 3.1 είναι ένα σύστημα $|S|$ εξισώσεων με $|S|$ αγνώστους. Εφόσον το μοντέλο του περιβάλλοντος είναι γνωστό, η λύση είναι απλή παρόλο το μεγάλο υπολογιστικό κόστος. Στη συγκεκριμένη περίπτωση, οι επαναληπτικές μέθοδοι είναι περισσότερο κατάλληλες. Θεωρείστε μία ακολουθία απο εκτιμώμενες συναρτήσεις αξίας $V_0, V_1, V_2, \ldots$. Η αρχική προσέγγιση, V_0, επιλέγεται αυθαίρετα (εκτός από την τερματική κατάσταση, που πρεπει να είναι 0), ενώ κάθε διαδοχική προσέγγιση λαμβάνεται χρησιμοποιώντας την Εξίσωση Bellman 3.1 για την

V^π σαν κανόνα ενημέρωσης:

$$\begin{aligned} V_{k+1}(s) &= E_\pi\{r_{t+1} + \gamma V_k(s_{t+1}) \mid s_t = s\} \\ &= \sum_a \pi(s,a) \sum_{s'} \mathcal{T}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V_k(s')], \end{aligned} \tag{3.2}$$

για κάθε $s \in S$. Μπορεί εύκολα να δειχθεί ότι η ακολουθία $\{V_k\}$ συγκλίνει στην πραγματική V^π καθώς $k \to \infty$. Ο συγκεκριμένος αλγόριθμος ονομάζεται επαναληπτική εκτίμηση πολιτικής (iterative policy evaluation).

Για να δημιουργήσει μια διαδοχική προσέγγιση, $\{V_{k+1}\}$ από την $\{V_k\}$, η επαναληπτική εκτίμηση της πολιτικής εφαρμόζει την ίδια διαδικασία για κάθε κατάσταση s: αντικαθιστά τη παλία αξία της s με μία καινούρια αξία η οποία λαμβάνεται από τις παλίες αξίες όλων των πιθανών διάδοχων καταστάσεων της s και τις αναμενόμενες άμεσες ανταμοιβές. Η συγκεκριμένη διαδικασία ονομάζεται *full backup*. Σε κάθε επανάληψη η επαναληπτική εκτίμηση πολιτικής κρατά τις αξίες κάθε κατάστασης έτσι ώστε να παραχθεί η νέα προσεγγιστική συνάρτηση αξίας V_{k+1}.

Αφού ολοκληρωθεί η εκτίμηση της πολιτικής χρησιμοποιούμε τη συνάρτηση αξίας κατάστασης γία τη βελτίωση της πολιτικής. Ο λόγος υπολογισμού της συνάρτησης αξίας κατάστασης μίας πολιτικής είναι η εύρεση καλύτερων πολιτικών. Ας υποθέσουμε πως έχουμε ορίσει τη συνάρτηση αξίας V^π για μια τυχαία ντετερμινιστική πολιτική π. Για κάποια κατάσταση s θέλουμε να γνωρίζουμε, αν θα πρεπει ή όχι, να αλλάξουμε την πολιτική ώστε να επιλέγει μία ενέργεια $a \neq \pi(s)$. Γνωρίζουμε ήδη πόσο καλό είναι να ακολουθήσουμε την τωρινή πολιτική από την κατασταση s, η αξία της οποίας είναι $V^\pi(s)$. Η αξία της ενέργειας a υπολογίζεται με τον εξής τρόπο:

$$\begin{aligned} Q^\pi(s,a) &= E_\pi\{r_{t+1} + \gamma V^\pi(s_{t+1}) \mid s_t = s, a_t = a\} \\ &= \sum_{s'} \mathcal{T}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V^\pi(s')]. \end{aligned} \tag{3.3}$$

Το βασικό κριτήριο είναι αν η $Q^\pi(s,a)$ είναι μεγαλύτερη ή ίση από τη $V^\pi(s)$. Αν είναι μεγαλύτερη, θα είναι καλύτερα να επιλέξουμε την ενέργειας a στην κατάσταση s και μετά να ακολουθήσουμε τη πολιτική π από το να ακολουθούμε όλη την ώρα τη πολιτική π. Αναμένεται επίσης να είναι καλύτερα αν κάθε φορά πού συναντάμε την κατάσταση s να επιλέγουμε την ενέργεια a, άρα η νέα πολιτική θα είναι γενικά καλύτερη. Υποθέτουμε οτι π και π' είναι ένα ζευγάρι ντετερμινιστικών πολιτικών έτσι ώστε για κάθε $s \in S$,

$$Q^\pi(s, \pi'(s)) \geq V^\pi(s). \tag{3.4}$$

Τότε η πολιτική π' θα πρέπει να είναι το ίδιο καλή ή καλύτερη σε σχέση με τη πολιτική π. Αυτό σημαίνει πως θα πρέπει να εξασφαλίζει μεγαλύτερες ή ίσες αναμενόμενες απολαβές γία όλες τις καταστάσεις $s \in S$:

$$V^{\pi'}(s) \geq V^\pi(s). \tag{3.5}$$

Εως τώρα έχουμε δεί πως δοθέντος μίας πολιτικής και μίας συνάρτησης αξίας, μπορούμε εύκολα να αξιολογήσουμε την αλλαγή της πολιτικής σε μία κατάσταση για την επιλογή μίας

διαφορετικής ενέργειας. Σε κάθε κατάσταση καλύτερη ενέργεια είναι αυτή πού μεγιστοποιεί τη συνάρτηση $Q^\pi(s,a)$. Έτσι η νέα άπληστη πολιτική, π', υπολογίζεται με τον εξής τρόπο:

$$\begin{aligned}\pi'(s) &= \arg\max_a Q^\pi(s,a) \\ &= \arg\max_a E\{r_{t+1} + \gamma V^\pi(s_{t+1}) \mid s_t = s, a_t = a\} \\ &= \arg\max_a \sum_{s'} \mathcal{T}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V^\pi(s')]. \end{aligned} \tag{3.6}$$

Εφόσον μια πολιτική, π, έχει βελτιωθεί χρησιμοποιόντας την V^π αποφέροντας μια καλύτερη πολιτική, π', μπορούμε να ξαναυπολογίσουμε την $V^{\pi'}$ και να την ξαναβελτιώσουμε παίρνοντας μια ακόμη καλύτερη πολιτική $V^{\pi''}$. Κατ'αυτό το τρόπο μπορούμε να πάρουμε μία ακολουθία βελτιωμένων πολιτικών και συναρτήσεων αξίας:

$$\pi_0 \rightarrow^E V^{\pi_0} \rightarrow^I \pi_1 \rightarrow^E V^{\pi_1} \rightarrow^I \pi_2 \rightarrow^E \cdots \rightarrow^I \pi^* \rightarrow^E V^*.$$

Αλγόριθμος 1 Επανάληψη ως προς την πολιτικη

1. Αρχικοποίηση
   Η $V(s) \in \Re$ και η $\pi(s) \in \mathcal{A}(s)$ αρχικοποιούνται τυχαία για κάθε $s \in \mathcal{S}$

2. Εκτίμηση Πολιτικής
repeat
   $\Delta \leftarrow 0$
   for each $s \in \mathcal{S}$ **do**
      $\upsilon \leftarrow V(s)$
      $V(s) \leftarrow \sum_{s'} \mathcal{T}^{\pi(s)}_{ss'}[\mathcal{R}^{\pi(s)}_{ss'} + \gamma V(s')]$
      $\Delta \leftarrow \max(\Delta, |\upsilon - V(s)|)$
until $\Delta < \theta$

3. Βελτίωση Πολιτικής
policy-stable $\leftarrow true$
for each $s \in \mathcal{S}$ **do**
   $b \leftarrow \pi(s)$
   $\pi(s) \leftarrow \arg\max_a \sum_{s'} \mathcal{T}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V(s')]$
   if $b \neq \pi(s)$ **then**
      policy-stable $\leftarrow false$
if policy-stable $= true$ **then**
   stop
else
   goto 2

Κάθε πολιτική εγγυάται πώς είναι αυστηρά βελτιωμένη σε σχέση με τη προηγούμενη (διαφορετικά είναι βέλτιστη). Επειδή μια πεπερασμένη ΜΔΑ έχει πεπερασμένο αριθμό πολιτικών, αυτή η διαδικασία συγκλίνει σε μια βέλτιστη πολιτική και μια βέλτιστη συνάρτηση

αξίας, σε ενα πεπερασμένο αριθμό επαναλήψεων. Αυτός ο τρόπος εύρεσης της βέλτιστης πολιτικής ονομάζεται επανάληψη ως προς τη πολιτική. Ο αλγόριθμος 1 περιγράφει τον αλγόριθμο επανάληψης ως προς τη πολιτική.

3.2.2 Επανάληψη ως πρός την αξία

Ένα μειονέκτημα της επανάληψης ως προς τη πολιτική είναι ότι κάθε επανάληψη εμπεριέχει το βήμα της εκτίμησης πολιτικής το οποίο εισάγει μεγάλο υπολογιστικό κόστος. Ένας καλύτερος τρόπος να βρούμε τη βέλτιστη πολιτική είναι βρίσκοντας τη βέλτιστη συνάρτηση αξίας V. Αυτή μπορεί να υπολογιστεί εύκολα από ένα απλό επαναληπτικό αλγόριθμο (Αλγόριθμος 2) πού ονομάζεται αλγόριθμος επανάληψης ως πρός την αξία (Value Iteration) [9], [15] και ο οποίος συγκλίνει στις σώστες βέλτιστες αξίες V^* (Bellman,1957; Bertsekas, 1987) [1], [2].

Δέν είναι εμφανές πότε θα πρέπει να σταματά ο αλγόριθμος επανάληψης ως πρός την αξία. Όπως η επανάληψη ως πρός τη πολιτική, έτσι και η επανάληψη ως πρός την αξία απαιτεί ένα άπειρο αριθμό επαναλήψεων για να συγκλίνει ακριβώς στη V^*. Έχει δειχθεί ότι, όταν η διαφορά δύο διαδοχικών συναρτήσεων αξίας είναι μικρότερη από ϵ, τότε η αξία της άπληστης πολιτικής, διαφέρει από την αξία της βέλτιστης πολιτικής λιγότερο από $2\epsilon\gamma/(1-\gamma)$ σε κάθε κατάσταση (Williams & Baird, 1993) [20]. Αυτό είναι ένα αποτελεσματικό κριτήριο τερματισμού του αλγοριθμού.

Αλγόριθμος 2 Επανάληψη ως πρός την αξία

Αρχικοποιήσετε τα V αυθαίρετα, π.χ, $V(s) = 0$, για κάθε $s \in \mathcal{S}^+$

repeat
  $\Delta \leftarrow 0$
  for each $s \in \mathcal{S}$ **do**
    $v \leftarrow V(s)$
    $V(s) \leftarrow \max_a \sum_{s'} \mathcal{T}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V(s')]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$(μικρός θετικός αριθμός)

Παραγωγή ντετερμινιστικής πολιτικής, π, τέτοια ώστε
  $\pi(s) = \arg\max_a \sum_{s'} \mathcal{T}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V(s')]$

3.3 Μέθοδοι Monte Carlo

Σε αντίθεση με τις μεθόδους ΔΠ, οι μέθοδοι Monte Carlo (MC) [15] δεν απαιτούν τη γνώση του μοντέλου του περιβάλλοντος. Βασίζονται μόνο στη συλλεγόμενη εμπειρία τους (ακολουθίες καταστάσεων, ενεργειών και ανταμοιβών), που αποκτήθηκε από την απευθείας

(on-line) ή εξομοιούμενη (off-line) αλληλεπιδρασή τους με το περιβάλλον. Η μάθηση με απευθείας αλληλεπίδραση δεν απαιτεί καμία προηγούμενη γνώση των δυναμικών του περιβάλλοντος και μπορεί να επιτύχει βέλτιστη συμπεριφορά. Παρόλο που η μάθηση με εξομοιούμενη εμπειρία απαιτεί τη γνώση του μοντέλου, δεν χρειάζεται τη γνώση του πλήρους μοντέλου αλλά αρκεί μια προσέγγιση αυτού.

Οι μέθοδοι MC είναι τρόποι επίλυσης προβλημάτων ενισχυτικής μάθησης βασιζόμενοι στους μέσους όρους των απολαβών του πράκτορα. Για να εξασφαλίσουμε οτι οι απολαβές είναι καλά ορισμένες, ορίζουμε τις μεθόδους MC μόνο για εργασίες με επεισόδια. Μόνο μετά την ολοκλήρωση ενός επεισοδίου μεταβάλλονται οι εκτιμήσεις των αξιών και οι πολιτικές.

Για την εκτίμηση μίας πολιτικής, οι μέθοδοι MC χρησιμοποιούν την συνάρτηση αξίας κατάστασης, η οποία εκφράζει την αναμενόμενη απολαβή πού λαμβάνει ο πράκτορας ξεκινώντας από αύτη τη κατάσταση. Ο πίο προφανής τρόπος για τον υπολογισμό της από την εμπειρία είναι να υπολογίσουμε το μέσο όρο των απολαβών πού παρατηρήθηκαν μετά από κάθε επίσκεψη στη συγκεκριμένη κατάσταση. Όσες περισσότερες απολαβές παρατηρούνται, τόσο περισσότερο ο μέσος όρος συγκλίνει στην αναμενόμενη αξία. Η συγκεκριμένη ιδέα χαρακτηρίζει όλες τις μεθόδους MC.

Κάθε εμφάνιση της κατάστασης s σε ένα επεισόδιο ονομάζεται επίσκεψη($visit$) της s. Η μέθοδος κάθε επίσκεψης MC (every-visit MC method), εκτιμά τη $V^\pi(s)$ ως το μέσο όρο των απολαβών πού λαμβάνονται μετά από κάθε επίσκεψη της s σε κάθε επεισόδιο. Η πρώτη μας επίσκεψη στη κατάσταση s σε ένα επεισόδιο, ονομάζεται πρώτη επίσκεψη (first visit) της s. Η μέθοδος πρώτης επίσκεψης MC (first-visit MC method) γιά τον υπολογισμό του μέσου όρου τών απολάβων λαμβάνει υπόψη μόνο τις απολαβές που λαμβάνονται μετά τη πρώτη επίσκεψη στη κατάσταση s. Τόσο η μέθοδος κάθε επίσκεψης MC όσο και η μέθοδος πρώτης επίσκεψης MC συγκλίνουν στη $V^\pi(s)$, καθώς ο αριθμός των επισκέψεων (ή των πρώτων επισκέψεων) στην s πλησιάζει στο άπειρο. Ο επόμενος Αλγόριθμος 3 περιγράφει τη μέθοδο πρώτης επίσκεψης MC.

Αλγόριθμος 3 Μέθοδος πρώτης επίσκεψης MC

Αρχικοποίηση:
$\pi \leftarrow$ πολιτική προς εκτίμηση
$V \leftarrow$ τυχαία συνάρτηση αξίας κατάστασης
$Returns(s) \leftarrow$ κενή λίστα, για κάθε $s \in \mathcal{S}$

loop
(a) Παραγωγή ενός επεισοδίου χρησιμοποιόντας τη πολιτική π
(b) Για κάθε κατάσταση s πού εμφανίζεται στο επεισόδιο:
$D \leftarrow$ απολαβή πού λαμβάνεται μετά τη πρώτη εμφάνιση της s
Προσθήκη της D στη λίστα $Returns(s)$
$V(s) \leftarrow average(Returns(s))$

Στη περίπτωση που το μοντελό του περιβάλλοντος δεν είναι διαθέσιμο, είναι προτιμότερο

να εκτιμήσουμε τις αξίες κατάστασης-ενέργειας παρά τις αξίες κατάστασης. Γνωρίζοντας το μόντελο, οι αξίες κατάστασης είναι αρκετές για να προσδιορίσουμε μια πολιτική. Αντίθετα, θα πρέπει να εκτιμήσουμε την αξία κάθε ενέργειας έτσι ώστε να μπορέσουμε να ορίσουμε μια πολιτική. Για το λόγο αυτό, ένας από τους σημαντικότερους στόχους των μεθόδων MC είναι ο υπολογισμός της Q^*. Το πρόβλημα εκτίμησης πολιτικής για αξίες ενέργειας είναι ο υπολογισμός της $Q^\pi(s, a)$, δηλαδή της αναμενόμενης απολαβής που λαμβάνουμε ξεκινώντας από τη κατάσταση s, επιλέγοντας την ενέργεια a και έπειτα ακολουθώντας τη πολιτική π. Οι μέθοδοι MC παραμένουν στην ουσία ιδιές με τις μεθόδους που παρουσιάστικαν για την αξία κατάστασης.

Το μόνο πρόβλημα που προκύπτει είναι πώς αρκέτα ζεύγη κατάστασης-ενέργειας μπορεί να μήν επισκεφτούν ποτέ. Στη περίπτωση που η π είναι μια ντετερμινιστική πολιτική, ακολουθώντας κάποιος τη π θα παρατηρήσει απολαβές μόνο για μια ενέργεια για κάθε κατάσταση. Για να υπολογίσει το μέσο όρο χωρίς απολαβές, οι MC εκτιμήσεις για τις άλλες ενέργειες δε θα βελτιωθούν με την εμπειρία. Αυτό είναι ενα πολύ σημαντικό πρόβλημα διότι ο σκοπός της μάθησης των αξίων κατάστασης-ενέργειας είναι να βοηθήσει στην επιλογή ανάμεσα στις ενέργειες που είναι διαθέσιμες σε κάθε κατάσταση. Δηλαδή, χρειάζεται να υπολογίσουμε τις αξίες όλων των ενεργείων για κάθε κατάσταση.

Για να λειτουργήσει η εκτίμηση πολιτικής για αξίες κατάστασης-ενέργειας, θα πρέπει να εξασφαλίσουμε πλήρη εξερεύνηση. Ένας τρόπος για να συμβεί αυτό είναι υποθέτοντας πως το πρώτο βήμα κάθε επεισοδίου ξεκινά από ενα ζεύγος κατάστασης-ενέργειας και κάθε ζεύγος αυτής της μορφής έχει μη μηδενική πιθανότητα για να επιλεχθεί κατά το ξεκίνημα. Αυτό μας εγγυάται πως όλα τα ζεύγη κατάστασης-ενέργειας θα επισκεφθούν άπειρο αριθμό φορών καθώς ο αριθμός τών επεισοδίων γίνεται άπειρος. Αυτή η παραδοχή ονομάζεται εξερεύνηση αφετηρίας (exploring starts).

Η βελτίωση της πολιτικής (policy improvement) επιτυγχάνεται κάνοντας τη πολιτική άπληστη σε σχέση με τη τρέχουσα συνάρτηση αξίας. Στη συγκεκριμένη περίπτωση είναι διαθέσιμη η συνάρτηση αξίας κατάστασης-ενέργειας, με αποτέλεσμα να μη χρειάζεται το μοντέλο για τη παραγωγή της άπληστης πολιτικής. Για κάθε συνάρτηση αξίας κατάστασης-ενέργειας Q, άπληστη πολιτική είναι αυτή που γία κάθε κατάσταση, $s \in \mathcal{S}$, ντετερμινιστικά επιλέγει την ενέργεια με τη μεγαλύτερη αξία Q:

$$\pi(s) = \arg\max_a Q(s, a). \tag{3.7}$$

Είναι φυσιολογίκη για την εκτίμηση πολιτικής MC η εναλλαγή ανάμεσα στην εκτίμηση και τη βελτίωση, επεισόδιο ανα επεισόδιο. Μετά το τέλος κάθε επεισοδίου, οι παρατηρούμενες απολαβές χρησιμοποιούνται για την εκτίμηση πολιτικής, ενώ αμέσως μετά η πολιτική βελτιώνεται σε όλες τις καταστάσεις που έχουν επισκεφτεί κατά τη διάρκεια του επεισοδίου. Ο συγκεκριμένος αλγόριθμος (Αλγόριθμος 4) ονομάζεται Monte Carlo ES (MC με εξερεύνηση αφετηρίας) [15].

Στον Monte Carlo ES, όλες οι απολαβές για κάθε ζεύγος κατάστασης-ενέργειας συσσωρεύονται και υπολογίζεται ο μέσος όρος τους, ανεξάρτητα από ποια πολιτίκη ήταν σε ισχύ όταν παρατηρήθηκαν. Είναι εύκολο να δειχθεί ότι ο Monte Carlo ES δεν μπορεί να συγκλίνει σε μια μή βελτιστη πολιτική.

Αλγόριθμος 4 Monte Carlo ES

Αρχικοποιήσετε, για κάθε $s \in \mathcal{S}$, $a \in \mathcal{A}$:

$Q(s,a) \leftarrow$ τυχαίες τιμές

$\pi(s) \leftarrow$ τυχαίες τιμές

$Returns(s,a) \leftarrow$ κενή λίστα

loop

(a) Παραγωγή ενός επεισοδιού χρησιμοποιόντας την πολιτική π και την εξερεύνηση αφετηρίας

(b) Για κάθε ζεύγος s,a πού εμφανίζεται στο επεισόδιο:

$D \leftarrow$ απολαβή μετά τη πρώτη εμφάνιση των s,a

Προσθήκη της D στη λίστα $Returns(s,a)$

$Q(s,a) \leftarrow average(Returns(s,a))$

(c) Για κάθε κατάσταση s του επεισοδίου:

$\pi \leftarrow \arg\max_a Q(s,a)$

3.4 Μάθηση Χρονικών Διαφορών

Η μάθηση χρονικών διαφορών (Temporal-Difference Learning) [15], [16] είναι ένας συνδυασμό των μεθόδων Monte Carlo και Δυναμικού Προγραμματισμού. Όπως στις μεθόδους Monte Carlo, οι μέθοδοι χρονικών διαφορών (ΧΔ) δεν απαιτούν τη γνώση του μοντέλου του περιβάλλοντος. Όπως και στο ΔΠ, οι μέθοδοι ΧΔ κάνουν ενημέρωση των εκτιμήσεων βασιζόμενοι σε ήδη γνωστές εκτιμήσεις, χώρις να περιμένουν για τη τελική απολαβή.

Τόσο οι μέθοδοι ΧΔ όσο και οι μέθοδοι Monte Carlo χρησιμοποιούν την εμπειρία για να επιλύσουν το πρόβλημα της πρόβλεψης. Έαν τη χρονική στιγμή t επισκεφθούμε τη μη τερματική κατάσταση s_t, τότε και οι δύο μέθοδοι ενημερώνουν την εκτιμήση $V(s_t)$ βασιζόμενη στο τι θα συμβεί μετά την επίσκεψη. Οι μέθοδοι Monte Carlo περιμένουν μέχρι η απολαβή πού ακολουθεί την επίσκεψη αυτή να γίνει γνωστή, έπειτα χρησιμοποιεί αυτή την απολαβή ως στόχο για την $V(s_t)$. Αντίθετα, οι μέθοδοι ΧΔ χρείαζεται να περιμένουν μόνο μέχρι το επόμενο βήμα. Η πιο απλή μέθοδος ΧΔ είναι η TD(0) (Sutton, 1988) [14], ο κανόνας ενημέρωσης της οποίας είναι ο εξής:

$$V(s_t) \leftarrow V(s_t) + \alpha[\underbrace{r_{t+1} + \gamma V(s_{t+1}) - V(s_t)}_{\text{TD error}}], \tag{3.8}$$

όπου $0 < \alpha < 1$ είναι ο ρυθμός μάθησης (learning rate) και $0 < \gamma < 1$ ο ρυθμός έκπτωσης (discount rate).

Αν τη χρονική στιγμή t επισκεφθούμε τη κατάσταση s_t, η εκτιμώμενη αξία της ενημερώνεται ετσί ώστε να είναι πιο κοντά στη $r_{t+1} + \gamma V(s_{t+1})$, όπου r_{t+1} είναι η άμεση ανταμοιβή πού λαμβάνεται και $V(s_{t+1})$ είναι η εκτιμώμενη αξία της παρατηρόμενης επόμενης κατάστασης. Η κύρια ιδέα είναι ότι η $r_{t+1} + \gamma V(s_{t+1})$ είναι ένα δείγμα της $V(s_t)$ και είναι περισσότερο πιθανό να είναι σωστή διότι ενσωματώνει τη πραγματική άμεση ανταμοιβή r_{t+1}. Πιό συγκεκριμένα η ποσότητα $r_{t+1} + \gamma V(s_{t+1})$ είναι αυτή προς την οποία θέλουμε

να μετατοπίσουμε την αξία $V(s_t)$. Αν ο ρυθμός μάθησης έχει προσαρμοστεί κατάλληλα, τότε η TD(0) είναι σίγουρο πως συγκλίνει στη βέλτιστη συνάρτηση αξίας. Ο Αλγόριθμος 5 περιγράφει τη μέθοδο TD(0) για την εκτίμηση της συνάρτησης αξίας V^π.

Αλγόριθμος 5 TD(0) για εκτίμηση της συνάρτησης αξίας κατάστασης V^π

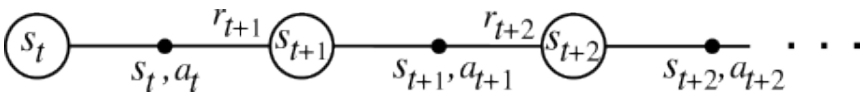
Αρχικοποιούμε αυθαίρετα τη V(s), π είναι η πολιτική προς εκτίμηση

repeat (για κάθε επεισόδιο):
  Αρχικοποιούμε την s
  repeat (για κάθε βήμα του επεισοδίου)
    $a \leftarrow$ ενέργεια πού επιλέγεται από τη π για τη κατάσταση s
    Εκτελούμε την ενέργεια a; Λαμβάνουμε την ανταμοιβή, r, και την επόμενη κατάσταση, s'
    $V(s) \leftarrow V(s) + \alpha[r + \gamma V(s') - V(s)]$
    $s \leftarrow s'$
  until s είναι τερματική κατάσταση

Ενα από τα πλεονεκτήματα των μεθόδων ΧΔ έναντι αυτών του ΔΠ είναι πως δεν απαιτείται η γνώση του μοντέλου του περιβάλλοντος. Το επόμενο πιο εμφανές πλεονέκτημα των μεθόδων ΧΔ σε σχέση με τις μεθόδους Monte Carlo ότι υλοποιούνται επαυξητικά (incremental) διότι θα πρέπει να περάσει ένα μόνο χρονικό βήμα για να γίνει η ενημέρωση. Αντίθετα στίς μεθόδους Monte Carlo θα πρέπει να περιμένουμε μέχρι το τέλος του επεισοδίου, αφού μόνο τότε γίνεται γνωστή η απολαβή. Έχει βρέθει πως στη πράξη οι μέθοδοι ΧΔ συγκλίνουν γρηγορότερα απο τις μεθόδους Monte Carlo για στοχαστικές εργασίες.

3.4.1 Sarsa

Στην ενότητα αυτή παρουσιάζεται ο αλγόριθμος Sarsa [11], [15], ενάς από τους πιό γνωστούς αλγορίθμους ΧΔ. Ο Sarsa είναι ένας αλγόριθμος εντός πολιτικής (on-policy), δηλαδή ένας αλγόριθμος στον οποίο η πολιτική η οποία αξιολογείται είναι και αυτή πού χρησιμοποιείται για να λάβουμε αποφάσεις. Αρχικά θα πρέπει να μάθουμε τη συνάρτηση αξίας κατάστασης-ενέργειας πάρα τη συνάρτηση αξίας κατάστασης. Δηλαδή, θα πρέπει να υπολογίσουμε την $Q^\pi(s, a)$ για τη τρέχουσα πολιτική π και για όλες τις καταστάσεις s και ενέργειες a. Αυτό επιτυγχάνεται χρησιμοποιώντας ακριβώς την ίδια μέθοδο ΧΔ πού περιγράφηκε προηγουμένως για τη μάθηση της V^π. Όπως προαναφέρθηκε, ένα επεισόδιο αποτελείται από μια ακολουθία καταστάσεων και ζευγών κατάστασης-ενέργειας:



Στη συνέχεια θα εξετάζουμε μεταβάσεις από ζεύγος κατάστασης-ενέργειας σε ζεύγος κατάστασης-ενέργειας και θα μαθαίνουμε την αξία των ζευγών κατάστασης-ενέργειας. Έστώ ότι ο πράκτορας τη χρονική στιγμή t βρίσκεται στη κατάσταση s_t, επιλέγει την

ενέργεια a_t και μεταβαίνει στη κατάσταση s_{t+1}. Στη συνέχεια λαμβάνει την ανταμοιβή r_{t+1} και επιλέγει την επόμενη ενέργεια a_{t+1}. Τότε οι αξίες κατάστασης-ενέργειας ενημερώνονται με τον εξής τρόπο:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]. \tag{3.9}$$

Μια τέτοια ενημέρωση γίνεται μετά από κάθε μετάβαση από μια μη τερματική κατάσταση s. Εάν η κατάσταση s_{t+1} είναι τερματική, τότε η αξία $Q(s_{t+1}, a_{t+1})$ είναι ίση με μηδέν. Ο σχεδιασμός ενός αλγορίθμου εντός πολιτικής πού βασίζεται στη μέθοδο πρόβλεψης Sarsa είναι αρκέτα απλός. Όπως σε όλες τις μεθόδους εντό πολιτικής, υπολογίζουμε την αξία Q^π για τη πολιτική π και ταυτόχρονα ενημερώνουμε άπληστα τη πολιτική π με βάση τη Q^π. Ο Αλγόριθμος 6 παρουσιάζει τη γενική μορφή του Sarsa. Ο Sarsa συγκλίνει με πιθανότητα 1 σε μια βέλτιστη πολιτική και συνάρτηση αξίας κατάσταστης-ενέργειας καθώς όλα τα ζεύγη κατάστασης-ενέργειας επισκέπτονται άπειρο αριθμό φορών.

Αλγόριθμος 6 Sarsa

Αρχικοποίηση της αξίας $Q(s, a)$

Repeat (για κάθε επεισόδιο):
  Αρχικοποίηση της κατάστασης s
  Επιλογή της ενέργειας a στη κατάσταση s χρησιμοποιώντας τη πολιτική που παράγεται από τη Q
  Repeat (για κάθε βήμα του επεισοδίου):
    Εκτέλεση της ενέργειας a και παρατήρηση των r, s'
    Επιλογή της ενέργειας a' στη κατάσταση s' χρησιμοποιώντας τη πολιτική που παράγεται από τη Q
    $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma Q(s', a') - Q(s, a)]$
    $s \leftarrow s'$
    $a \leftarrow a'$
  until η κατάσταση s να είναι τερματική

3.4.2 Q-Learning

Ένας επίσης πολύ γνωστός αλγόριθμος ΧΔ είναι ο Q-Learning (Watkins, 1989; Watkins & Dayan, 1992) [9], [15], [18], [19] . Ο Q-Learning είναι ένας αλγόριθμος εκτός πολιτικής (off-policy) δηλαδή η πολιτική που χρησιμοποιείται για τη λήψη αποφάσεων δεν χρειάζεται να είναι ίδια με αυτή που αξιολογείται και βελτιώνεται. Ο Q-Learning χρησιμοποιεί τον παρακάτω κανόνα ενημέρωσης:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right]. \tag{3.10}$$

Στη συγκεκριμένη περίπτωση, η συνάρτηση αξίας κατάστασης-ενέργειας, Q, προσεγγίζει άμεσα τη Q^*, τη βέλτιστη συνάρτηση αξίας κατάστασης-ενέργειας, ανεξάρτητα από τη πολιτική πού ακολουθείται. Αν κάθε ενέργεια εκτελείται σε κάθε κατάσταση άπειρο αριθμό

φορών και ο ρυθμός μάθησης φθίνει κατάλληλα, τότε οι αξίες Q θα συγκλίνουν με πιθανότητα 1 στην Q^* (Watkins, 1989; Tsitsiklis, 1994; Jaakkola, Jorda , & Singh, 1994). Ο Αλγόριθμος 7 περιγράφει τον Q-Learning.

Αλγόριθμος 7 Q-Learning

Αρχικοποίηση της αξίας $Q(s, a)$

Repeat (για κάθε επεισόδιο):
  Αρχικοποίηση της κατάστασης s
  Repeat (για κάθε βήμα του επεισοδίου):
    Επιλογή της ενέργειας a στη κατάσταση s χρησιμοποιώντας τη πολιτική που παράγεται από τη Q
    Εκτέλεση της ενέργειας a και παρατήρηση των r, s'
    $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$
    $s \leftarrow s'$
  until η κατάσταση s να είναι τερματική

3.5 Ίχνη Επιλεξιμότητας

Τα ίχνη επιλεξιμότητας (eligibility traces) (Singh and Sutton, 1996) [13], [15] είναι ένας βασικός μηχανισμός της ενισχυτικής μάθησης. Ένα ίχνος επιλεξιμότητας είναι μία προσωρινή καταγραφή ενός γεγονότος, όπως η επίσκεψη μιας κατάστασης ή η λήψη μίας ενέργειας. Το ίχνος μαρκάρει τις παραμέτρους (αποθηκεύονται στη μνήμη) πού σχετίζονται με το γεγονός σαν επιλεγμένες, για να υποβληθούν στη συνέχεια σε αλλαγές. Όταν συμβαίνει ένα σφάλμα ΧΔ, μόνο στις επιλεγμένες (eligible) καταστάσεις ή ενέργειες απονέμεται κάποιο κέρδος ή ποινή για το σφάλμα.

Στή συνέχεια ορίζουμε τον αλγόριθμο $TD(\lambda)$, το λ αναφέρεται στη χρήση των ίχνων επιλεξιμότητας. Ο $TD(\lambda)$ είναι χρήσιμος εξαιτίας τόσο της εννοιολογικής όσο και της υπολογιστικής του απλότητας. Κάθε κάτασταση στο συγκεκριμένο αλγόριθμο σχετίζεται με μία επιπρόσθετη μεταβλητή μνήμης, το ιχνος επιλεξιμότητας. Το ίχνος επιλεξιμότητας της κατάστασης s τη χρονική στιγμή t συμβολίζεται ως $e_t(s) \in \Re^+$. Σε κάθε βήμα, τα ίχνη επιλεξιμότητας όλων των καταστάσεων φθίνουν με συντελεστή $\gamma\lambda$, ενώ το ίχνος επιλεξιμότητας της κατάστασης που επισκεφθήκαμε σε αύτο το βήμα αυξάνεται κατά 1:

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & \text{if } s \neq s_t \\ \gamma\lambda e_{t-1}(s) + 1 & \text{if } s = s_t, \end{cases} \tag{3.11}$$

για κάθε $s \in \mathcal{S}$, όπου $0 < \gamma < 1$ είναι ο ρυθμός έκπτωσης (discount return) και $0 \leq \lambda \leq 1$ είναι ένα βάρος για το καθορισμό της μείωσης του ίχνους (trace-decay parameter). Κάθε χρονική στιγμή, τα ίχνη επιλεξιμότητας καταγράφουν ποίες καταστάσεις έχουμε επισκεφτεί

πρόσφατα. Το σφάλμα ΧΔ για την πρόβλεψη της αξίας κατάστασης είναι:

$$\delta_t = r_{t+1} + \gamma V_t(s_{t+1}) - V_t(s_t). \tag{3.12}$$

Στον αλγόριθμο $TD(\lambda)$, το σήμα του σφάλματος ΧΔ ενεργοποιεί αναλογίκα την εκτέλεση ενημερώσεων σε όλες τις πρόσφατα επισκεπτόμενες καταστάσεις, που σηματοδοτούνται από μη μηδενικά ίχνη:

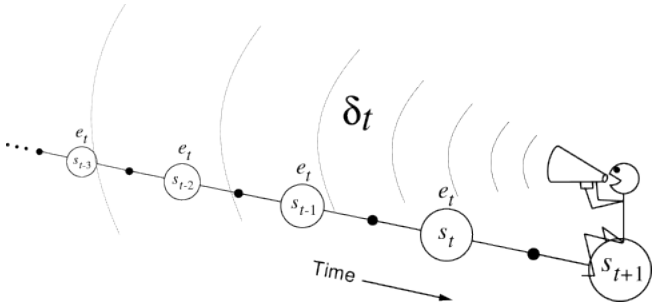
$$\Delta V_t(s) = \alpha \delta_t e_t(s), \text{για κάθε } s \in \mathcal{S}. \tag{3.13}$$

Ο αλγόριθμος 8 παρουσιάζει το ψευδοκώδικα για τον TD(λ).

Αλγόριθμος 8 TD(λ)

Αρχικοποιούμε τη $V(s)$ τυχαία, για κάθε $s \in S$
Repeat (για κάθε επεισόδιο)
  Αρχικοποιούμε $e(s) = 0$, για κάθε $s \in S$
  Αρχικοποιούμε τη κατάσταση s
  Repeat (για κάθε βήμα του επεισοδίου)
    $a \leftarrow$ η ενέργεια πού επιλέγει η πολιτική π στην s
    Εκτελούμε την ενέργεια a και λαμβάνουμε την ανταμοιβή, r, και τη κατάσταση, s'
    $\delta \leftarrow r + \gamma V(s') - V(s)$
    $e(s) \leftarrow e(s) + 1$
    for all s
      $V(s) \leftarrow V(s) + \alpha \delta e(s)$
      $e(s) \leftarrow \gamma \lambda e(s)$
    $s \leftarrow s'$
  until η κατάσταση s να είναι τερματική

Ο TD(λ) είναι προσανατολισμένος πίσω στο χρόνο. Κάθε χρονική στιγμή βρίσκουμε το τρέχον σφάλμα ΧΔ και το απονέμουμε προς τα πίσω σε κάθε προηγούμενη κατάσταση σύμφωνα με το ίχνος επιλεξιμότητας των καταστάσεων, τη συγκεκριμένη χρονική στιγμή. Στο Σχήμα 3.1 παρουσιάζεται σχηματικά η παραπάνω διαδικασία.



Σχήμα 3.1: Σχηματική περιγραφή του αλγορίθμου TD(λ)

Εαν $\lambda = 0$, τότε στην 3.11 όλα τα ίχνη θα είναι μηδέν τη χρονική στιγμή t εκτός από το ίχνος που αντιστοιχεί στη κατάσταση s_t. Τότε η ενημέρωση 3.13 του TD(λ) μετατρέπεται

στον απλό TD κανόνα, TD(0). Ο TD(0) είναι η περίπτωση όπου μόνο η κατάσταση που προηγείται της τρέχουσας κατάστασης αλλάζει από το σφάλμα ΧΔ. Για μεγαλύτερες τιμές του λ, αλλα $\lambda < 1$, περισσότερες από μία προηγούμενες καταστάσεις αλλάζουν, αλλά οι λιγότερες πρόσφατες καταστάσεις αλλάζουν λιγότερο καθώς το ίχνος επιλεξιμοτητάς τους είναι μικρότερο σε σχεση με τις πιό πρόσφατες. Εάν $\lambda = 1$, τότε η τιμή που δίνεται στις πρωγενέστερες καταστάσεις φθίνει μόνο κατα γ σε κάθε βήμα. Αυτό αποδεικνύεται πως κάνει ακριβώς το ίδιο πράγμα με την μέθοδο Monte Carlo.

3.5.1 Sarsa(λ)

Στην ενότητα αύτη θα δούμε πως τα ίχνη επιλεξιμότητας μπορούν να συνδυαστούν με απλό τρόπο με τον αλγόριθμο Sarsa, για την παραγωγή μιας νέας εντός πολιτικής μεθόδου ΧΔ. Ονομάζουμε Sarsa(λ) [15] τη συγκεκριμένη έκδοση του αλγορίθμου Sarsa με ίχνη επιλεξιμότητας. Η ιδέα στον Sarsa(λ) [15] είναι να εφαρμόσουμε τη μέθοδο TD(λ) για ζεύγη κατάστασης-ενέργειας αντί γία καταστάσεις. Έτσι χρειαζόμαστε ένα ίχνος όχι απλά για κάθε κατάσταση, αλλά για κάθε ζεύγος κατάστασης-ενέργειας. Συμβολίζουμε $e_t(s,a)$, το ίχνος για το ζεύγος κατάστασης-ενέργειας (s,a). Ο κανόνας ανανέωσης της μεθόδου για τις αξίες Q είναι:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha\delta_t e_t(s,a), \quad \text{για κάθε } s,a. \tag{3.14}$$

όπου

$$\delta_t = r_{t+1} + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \tag{3.15}$$

και

$$e_t(s,a) = \begin{cases} \gamma\lambda e_{t-1}(s,a) + 1 & \text{αν } s = s_t \text{και } a = a_t \\ \gamma\lambda e_{t-1}(s,a) & \text{διαφορετικά .} \end{cases} \tag{3.16}$$

Ο Αλγόριθμος 9 παρουσιάζει το ψευδοκώδικα για τον Sarsa(λ).

3.6 Γενίκευση με Προσέγγιση Συνάρτησης

Μέχρι τώρα θεωρούσαμε ότι οι συναρτήσεις αξίας που εκπαιδεύονται από τους πράκτορες αναπαριστόνται με τη μορφή πίνακα καθώς θεωρούμε διακριτές καταστάσεις, με μία καταχώρηση για κάθε κατάσταση ή για κάθε ζεύγος κατάστασης-ενέργειας. Μια τέτοια προσέγγιση περιορίζεται μόνο σε μικρούς χώρους καταστάσεων και ενεργειών. Το πρόβλημα δεν είναι μόνο η μνήμη πού απαιτείται για την αποθήκευση μεγάλων πινάκων, αλλά και ο χρόνος και τα δεδομένα που χρειάζονται για να γεμίσουμε τους πίνακες με τις σωστές τιμές. Για το λόγο αυτό κρίνεται αναγκαία η ύπαρξη ενός μηχανισμού γενίκευσης (generalization). Το είδος γενίκευσης που απαιτείται είναι γνωστό ως προσέγγιση συνάρτησης (function approximation) διότι προσπαθεί μέσω παραδειγμάτων που λαμβάνει να προσεγγίση μια συνάρτηση. Η προσέγγιση συνάρτησης καθιστά πρακτική την αναπαράσταση συναρτήσεων αξίας για πολύ μεγάλους χώρους καταστάσεων, αλλά αυτό δεν είναι το κύριο πλεονεκτημά

Αλγόριθμος 9 Sarsa(λ)

Αρχικοποιούμε τη $Q(s,a)$ τυχαία και $e(s,a)=0$, για κάθε s,a
Repeat (για κάθε επεισόδιο)
  Αρχικοποιούμε τη κατάσταση s και την ενέργεια a
  Repeat (για κάθε βήμα του επεισοδίου)
    Εκτελούμε την ενέργεια a και λαμβάνουμε την ανταμοιβή, r, και τη κατάσταση, s'
    Επιλέγουμε την ενέργεια a' στην s' σύμφωνα με την Q
    $\delta \leftarrow r + \gamma Q(s',a') - Q(s,a)$
    $e(s,a) \leftarrow e(s,a) + 1$
    for all s
      $Q(s,a) \leftarrow Q(s,a) + \alpha\delta e(s,a)$
      $e(s,a) \leftarrow \gamma\lambda e(s,a)$
    $s \leftarrow s'$
    $a \leftarrow a'$
  until η κατάσταση s να είναι τερματική

της. Η συμπίεση που επιτυγχάνεται από την προσέγγιση συνάρτησης επιτρέπει στον πράκτορα να κάνει γενίκευση από καταστάσεις που έχει επισκεφθεί σε καταστάσεις που δεν έχει επισκεφθεί.

Τη χρονική στιγμή t η κατά προσέγγιση συνάρτηση αξίας, V_t, δεν αναπαρίσταται ως πίνακας αλλά ως μια παραμετροποιημένη μορφή με διάνυσμα παραμέτρων $\vec{\theta}_t$. Στις μεθόδους βαθμωτής πτώσης (gradient-descent), το διάνυσμα παραμέτρων είναι ενα διάνυσμα στήλης με σταθερό αριθμό παραμέτρων, $\vec{\theta}_t = (\theta_t(1), \theta_t(2), \ldots, \theta_t(n))^T$ και η συνάρτηση V είναι διαφορίσιμη ως προς το διάνυσμα $\vec{\theta}$. Υποθέτουμε ότι σε κάθε βήμα παρατηρούμε παραδείγματα της μορφής $s_t \mapsto V^\pi(s_t)$ και ότι οι καταστάσεις εμφανίζονται στα παραδείγματα με την ίδια πιθανότητα. Οι μέθοδοι βαθμωτής πτώσης προσαρμόζουν το διάνυσμα παραμέτρων μετά από κάθε παράδειγμα πρός τη κατεύθυνση της αρνητικής κλίσης, η οποία θα μειώσει περισσότερο το σφάλμα για το συγκεκριμένο παράδειγμα:

$$\vec{\theta}_{t+1} = \vec{\theta}_t + \alpha[V^\pi(s_t) - V_t(s_t)]\nabla_{\vec{\theta}} V_t(s_t), \tag{3.17}$$

όπου α είναι ο ρυθμός μάθησης και $\nabla_{\vec{\theta}} V_t(s)$ είναι η μερική παράγωγος της V_t ως προς το διάνυσμα παραμέτρων $\vec{\theta}_t$, $\left(\frac{\partial V_t(s)}{\partial \theta_t(1)}, \frac{\partial V_t(s)}{\partial \theta_t(2)}, \cdots, \frac{\partial V_t(s)}{\partial \theta_t(n)}\right)^T$.

Ο κανόνας ενημέρωσης για τη μέθοδο gradient descent TD(λ) είναι ο εξής:

$$\vec{\theta}_{t+1} = \vec{\theta}_t + \alpha\delta_t \vec{e}_t, \tag{3.18}$$

όπου δ_t είναι το σφάλμα ΧΔ,

$$\delta_t = r_{t+1} + \gamma V_t(s_{t+1}) - V_t(s_t), \tag{3.19}$$

και $\vec{e}_t$ είναι ένα διάνυσμα στήλη, ενα για κάθε μεταβλητή του $\vec{\theta}_t$, το οποίο ενημερώνεται ως εξής

$$\vec{e}_t = \gamma\lambda\vec{e}_{t-1} + \nabla_{\vec{\theta}_t} V_t(s_t), \tag{3.20}$$

με $\vec{e}_0 = 0$.

Μια από τις σημαντικότερες κατηγορίες μεθόδων προσέγγισεις συνάρτησης είναι οι γραμμικές μέθοδοι. Σε αυτές τις μεθόδους η συνάρτηση προς προσέγγιση, V_t, είναι γραμμική ως προς το διάνυσμα παραμέτρων $\vec{\theta}_t$. Σε κάθε κατάσταση s αντιστοιχεί ενα διάνυσμα χαρακτηριστικών $\vec{\phi}_s = (\phi_s(1), \phi_s(2), \ldots, \phi_s(n))^T$, με ίδιο αριθμό παραμέτρων με το $\vec{\theta}_t$. Έτσι η συνάρτηση αξίας κατάστασης δίνεται ως εξής:

$$V_t(s) = \vec{\theta}_t^T \vec{\phi}_s = \sum_{i=1}^{n} \theta_t(i)\phi_s(i). \tag{3.21}$$

Η κατάλληλη εισαγωγή χαρακτηριστικών είναι ένας τρόπος για την εισαγωγή προγενέστερης αποκτηθείσας γνώσης στα συστήματα ενισχυτικής μάθησης. Τα χαρακτηριστικά πρέπει να αντιστοιχούν στα φυσικά χαρακτηριστικά της εργασιάς. Εάν αποτιμούμε τις καταστάσεις ενός ρομποτ, τότε θέλουμε να έχουμε χαρακτηριστικά για τις τοποθεσείες, τους βαθμούς της εναπομείναντος ισχύος της μπαταρίας, τις πρόσφατες μετρήσεις των αισθητήρων, κλπ. Γενικά, χρειαζόμαστε χαρακτηριστικά για το συνδυασμό αυτών των φυσικών ποσοτήτων. Στη συνέχεια εξετάζουμε κάποιους τρόπους για να γίνει αυτό.

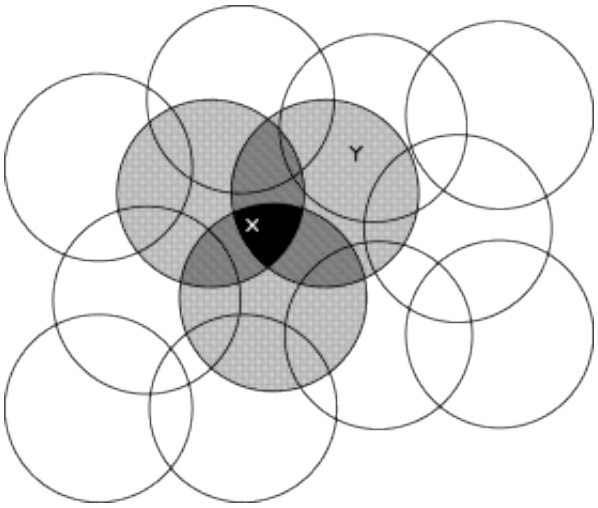
3.6.1 Χονδροειδής Κωδικοποίηση

Ας υποθέσουμε μια εργασία στην οποία το σύνολο καταστάσεων είναι συνεχές και δισδιάστατο. Σε αυτή τη περίπτωση μια κατάσταση είναι ένα σημείο στο διδιάσταστο χώρο, ένα διάνυσμα με δύο πραγματικούς συντελεστές. Ένα είδος χαρακτηριστικών για αυτή τη περίπτωση, είναι αυτά πού αντιστοιχούν σε κύκλους στο χώρο καταστάσεων όπως φαίνεται στο Σχημα 3.2. Εάν η κατάσταση είναι μέσα στο κύκλο, τότε το αντίστοιχο χαρακτηριστικό έχει τιμή 1 (present) διαφορετικά έχει τιμή 0 (absent). Αυτού του είδους τα χαρακτηριστικα ονομάζονται δυαδικά χαρακτηριστικά (binary features). Δοθέντος μίας κατάστασης, τα δυαδικά χαρακτηριστικά με τιμή 1 υποδεικνύουν τους κύκλους μέσα στους οποίους βρίσκεται η κατάσταση. Αυξάνοντας την ακτίνα των κύκλων έχει ως αποτέλεσμα καλύτερη γενικευτική ικανότητα. Η αναπαράσταση μιας κατάστασης με χαρακτηριστικά αυτου του τύπου ονομάζεται χονδροειδής κωδικοποίηση (coarse coding) [15].

3.6.2 Ακτινωτές Συναρτήσεις Βάσης

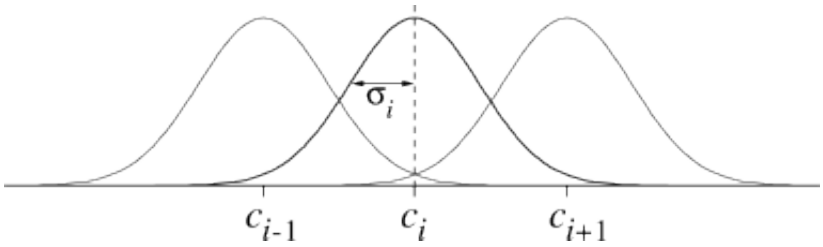
Οι ακτινωτές συναρτήσεις βάσης (Radial basis functions(RBF)) [4], [15] είναι η φυσική επέκταση της χονδροειδής κωδικοποίησης σε χαρακτηριστικά συνεχών τιμών. Αντί κάθε χαρακτηριστικό να λαμβάνει διακριτές τιμές 0 ή 1, μπορεί να παίρνει συνέχεις τιμές στο διάστημα $[0, 1]$, αντικατοπρίζοντας μεγάλη ποικιλία βαθμών που τα χαρακτηριστικά είναι παρών (present). Ένα τυπικό RBF χαρακτηριστικό, i, εξαρτάται μόνο από την απόσταση ανάμεσα στη κατάσταση s και το κέντρο του c_i:

$$\phi_s(i) = \exp\left(-\frac{\| s - c_i \|^2}{2\sigma_i^2}\right). \tag{3.22}$$



Σχήμα 3.2: Χονδροειδής Κωδικοποίηση(Coarse Coding)

Η νόρμα μπορεί να επιλεγεί με τέτοιο τρόπο ώστε να ταιριάζει τόσο στις καταστάσεις όσο και την εργασία. Το Σχήμα 3.3 δείχνει ένα μονοδιάστατο παράδειγμα με μετρική την Ευκλίδεια απόσταση.



Σχήμα 3.3: Μονοδιάστατες ακτινωτές συναρτήσης βάσης

Κεφαλαιο 4

ΕΝΙΣΧΥΤΙΚΗ ΜΑΘΗΣΗ ΜΕ ΓΚΑΟΥΣΙΑΝΕΣ ΔΙΑΔΙΚΑΣΙΕΣ

4.1 Εισαγωγή

Μια σημαντική συνιστώσα πολλών μεθόδων επίλυσης ενισχυτικής μάθησης είναι η εκτίμηση των αξιών κατάστασης ή κατάστασης ενέργειας για μια σταθερή πολιτική, μια εργασία γνωστή ως εκτίμηση πολιτικής. Στο παρόν κεφάλαιο, παρουσιάζεται μια Μπεϋσιανή προσέγγιση εκτίμησης πολιτικής σε γενικούς χώρους καταστάσεων και ενεργείων, η οποία χρησιμοποιεί στατιστικά γεννητικά μοντέλα μέσω Γκαουσιανών διαδικασιών (Gaussian Processes) για τις συναρτήσεις αξίας [5], [6], [7]. Η βασίζομενη σε ένα στατιστικό μοντέλο (βασιζόμενο σε Γκαουσιανές διαδικασίες) εκ των υστέρων κατανομή, μας παρέχει την εκτίμηση της συνάρτησης αξίας καθώς επίσης και τη διακύμανση αυτής της εκτίμησης.

Αρχικά παρουσιάζονται κάποιες βασικές έννοιες σχετικά με τη Γκαουσιανή Κατανομή και με τις Γκαουσιανές διαδικασίες, που αποτελούν τη βάση των μεθόδων που περιγράφονται στο κεφάλαιο αυτό. Στην ενότητα 4.4, παρουσιάζεται μια Μπεϋσιανή προσέγγιση για την εκτίμηση της συνάρτηση αξίας και πως αυτή μπορεί να επεκταθεί σε εργασίες με επεισόδια. Επίσης, περιγράφεται και μια αραιή έκδοση της συγκεκριμένης μεθόδου εκτίμησης

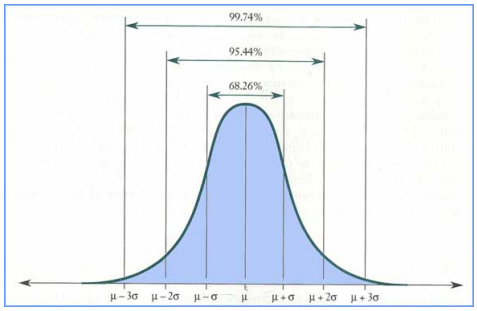
συνάρτησης αξίας. Στην ενότητα 4.5 περιγράφονται αλγόριθμοι βελτίωσης της πολιτικής που αξιοποιούν τις προηγούμενες μεθόδους.

4.2 Η Πολυμεταβλητή Γκαουσιανή Κατανομή

Μια από τις σημαντικότερες κατανομές πιθανότητας για συνεχείς μεταβλητές είναι η κανονική ή Γκαουσιανή κατανομή (Gaussian Distribution) [4], [10]. Στη περίπτωση μιας μονοδιάστατης πραγματικής μεταβλητής x, η Γκαουσιανή κατανομή ορίζεται ως

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{4.1}$$

και προσδιορίζεται από δυο παραμέτρους: το μέσο (mean) μ, και τη διακύμανση (variance) σ^2. Η τετραγωνική ρίζα της διακύμανσης γράφεται ως σ και ονομάζεται τυπική απόκλιση (standard deviation). Το αντίστροφο της διακύμανσης δίνεται ως $\beta = 1/\sigma^2$ και ονομάζεται ακρίβεια (precision). Η γραφική της παράσταση φαίνεται στο Σχήμα 4.1. Για ένα D-



Σχήμα 4.1: Γκαουσιανή κατανομή

διάστατο διάνυσμα $\mathbf{x}$, η πολυμεταβλητή Γκαουσιανή κατανομή ορίζεται ως

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \tag{4.2}$$

όπου $\boldsymbol{\mu}$ είναι ένα D-διάστατο διάνυσμα του μέσου, $\boldsymbol{\Sigma}$ είναι ο $D \times D$ πίνακας συνδιακύμανσης (covariance matrix) και το $|\boldsymbol{\Sigma}|$ ορίζει την ορίζουσα του $\boldsymbol{\Sigma}$.

4.3 Γκαουσιανές Διαδικασίες

Οι Γκαουσιανές Διαδικασίες (Gaussian Processes) [4], [10] εχούν χρησιμοποιηθεί εκτενώς τα τελευταία χρόνια σε προβλήματα ταξινόμησης (classification) και παλινδρόμησης (regression). Βασιζόμενες σε πιθανοτικά γεννητικά μοντέλα (probabilistic generative models), οι μέθοδοι Γκαουσιανών διαδικασιών είναι θεωρητικά ελκυστικές αφού επιτρέπουν τη Μπεϋσιανή μεταχείριση αυτών των προβλημάτων, παράγοντας πλήρες εκ των υστέρων κατανομές βασιζόμενες τόσο στις εκ των προτέρων πεποιθήσεις μας όσο και στα παρατηρούμενα δεδομένα. Εφόσον οι Γκαουσιανές διαδικασίες μπορούν να οριστούν απευθείας στο χώρο

της συνάρτησης, δεν είναι τόσο περιοριστικές όσο τα παραμετρικά μοντέλα σχετικά με τον υποθετικό χώρο στον οποίο η μάθηση λαμβάνει μέρος. Επιπλέον, όταν τόσο η εκ των προτέρων κατανομή όσο και η πιθανοφάνεια (likelihood) είναι Γκαουσιανές, η εκ των υστέρων κατανομή θα είναι επίσης Γκαουσιανή και ο κανόνας του Bayes παράγει εκφράσεις κλειστής μορφής.

Μια στοχαστική διαδικασία F είναι Γκαουσιανή εάν οι μεταβλητές τις που αντιστοιχούν σε ένα οποιοδήποτε πεπερασμένο υποσύνολο του $\mathcal{X}$ είναι από κοινού Γκαουσιανές. Προκειμένου να εφαρμόσουμε το Μπεϋσιανό συμπέρασμα χρησιμοποιώντας Γκαουσιανές διαδικασίες θα πρέπει πρώτα να ορίσουμε ένα στατιστικό γενεσιουργό μοντέλο (statistical generative model).Τέτοια μοντέλα αποτελούνται από :

1. Μια εξίσωση του μοντέλου που συσχετίζει τα παρατηρούμενα και τα απαρατήρητα συστατικά του μοντέλου μας. Συνήθως τα τελευταία μετασχηματίζονται και αλλοιώνονται από κάποιο προστιθέμενο θόρυβο για να παραγάγουν τα πρώτα. Οι απαρατήρητες(ή κρυμμένες) διαδικασίες είναι το αντικείμενο της Bayesian inference προσπάθειας μας.

2. Μια κατανομή για τον θόρυβο. Με τον όρο θόρυβο, αναφερόμαστε σε οποιαδήποτε προστιθέμενη στοχαστική διαδικασία στην εξίσωση του μοντέλου, οι παράμετροι τις οποίας είναι γνωστές και η οποία δεν αποτελεί το αντικείμενο του inference προβλήματος μας.

3. Μια εκ των προτέρων κατανομή για τις απαρατήρητες διαδικασίες. Αυτό είναι ένα απαραίτητο συστατικό που απαιτείται για την εφαρμογή του κανόνα Bayes.

Εφόσον η F είναι εκ των προτέρων Γκαουσιανή, η εκ των προτέρων πιθανότητα της προσδιορίζεται πλήρως από το μέσο (mean) της και τη συνδιακυμανσή (covariance) της,

$$\mathbf{E}[F(\mathbf{x})] \stackrel{\text{def}}{=} f_0(\mathbf{x}) \tag{4.3}$$

$$\mathbf{Cov}[F(\mathbf{x}), F(\mathbf{x}')] = \mathbf{E}[F(\mathbf{x}), F(\mathbf{x}')] - f_0(\mathbf{x}) f_0(\mathbf{x}') \stackrel{\text{def}}{=} k(\mathbf{x}, \mathbf{x}'), \tag{4.4}$$

αντίστοιχα, όπου το $\mathbf{E}$ υποδηλώνει την αναμενόμενη τιμή σε σχέση με την εκ των προτέρων κατανομή. Για να είναι ο $k(\cdot, \cdot)$ μια θεμιτή συνδιακύμανση θα πρέπει να είναι συμμετρικός και θετικά ορισμένος. Για να είναι συμμετρικός θα πρέπει $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$ για όλα τα $\mathbf{x}, \mathbf{x}' \in X$. Για πεπερασμένο X, η $k(\cdot, \cdot)$ είναι ένας πίνακας. Ένας πίνακας είναι θετικά ορισμένος εάν $\int_{X^2} g(\mathbf{x}) k(\mathbf{x}, \mathbf{x}') g(\mathbf{x}') \mathrm{d}x \mathrm{d}x' \geq 0, \forall g \in \ell_2$. Στις μεθόδους πυρήνα (kernel methods), η $k(\cdot, \cdot)$ αναφέρεται ως συνάρτηση πυρήνα (kernel function) και αντιμετωπίζεται ως ένα εσωτερικό γινόμενο σε κάποιο υψηλής διάστασης χώρο. Όπως έχει αποδειχθεί, οι δυο αυτές απόψεις είναι στη πραγματικότητα ταυτόσημες. Αυτός είναι ο λόγος που ο ίδιος όρος χρησιμοποιείται και για τις δύο συναρτήσεις.

Στη συνέχεια θα επανεξετάσουμε τη χρήση γκαουσιανών διαδικασιών για τη παλινδρόμηση με λευκό Γκαουσιανό θόρυβο. Στο σημείο αυτό έχουμε στη διάθεση μας ένα δείγμα από t παραδείγματα εκπαίδευσης $\{(\mathbf{x}_i, y_i)\}_{i=1}^t$. Η εξίσωση του μοντέλου για κάποιο $\mathbf{x} \in \mathcal{X}$ είναι:

$$Y(\mathbf{x}) = F(\mathbf{x}) + N(\mathbf{x}), \tag{4.5}$$

όπου F είναι η Γκαουσιανή διαδικασία που αντιστοιχεί στην άγνωστη συνάρτηση από την οποία παράγονται τα δεδομένα, N είναι η Γκαουσιανή διαδικασία του θορύβου (ανεξάρτητη από την F) και Y είναι η παρατηρούμενη διαδικασία, η οποία μοντελοποιείται ως μια θορυβώδης εκδοχή της F. Η F υποθέτουμε εκ των προτέρων πως είναι μια Γκαουσιανή διαδικασία με μέσο $f_0(\cdot)$ και συμεταβλητότητα που δίνεται από μια συνάρτηση πυρήνα $k(\cdot,\cdot)$ όπως στην Εξίσωση 4.4. Η Εξίσωση 4.5 που εκτιμάται για τα παραδείγματα εκπαίδευσης μπορεί να γραφεί συνοπτικά ως

$$Y_t = F_t + N_t, \tag{4.6}$$

όπου $Y_t = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_t))^T$, $F_t = (F(\mathbf{x}_1), \ldots, F(\mathbf{x}_t))^T$ και $N_t = (N(\mathbf{x}_1), \ldots, N(\mathbf{x}_t))^T$. Στο παραδειγμά μας, υποθέτουμε πως οι όροι του θορύβου που αλλοιώνουν κάθε δείγμα είναι ανεξάρτητοι και πανομοιότυποι κατανεμημένοι. Ως εκ τούτου έχουμε

$$N_t \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}),$$

όπου σ^2 είναι η διακύμανση κάθε όρου του θορύβου. Η απο κοινού κατανομή της $F(\mathbf{x})$ για οποιοδήποτε $\mathbf{x} \in \mathcal{X}$ με την Y_t είναι

$$\begin{pmatrix} F(\mathbf{x}) \\ Y_t \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} f_0(x) \\ \mathbf{f}_0 \end{pmatrix}, \begin{bmatrix} k(\mathbf{x},\mathbf{x}) & \mathbf{k}_t(\mathbf{x}) \\ \mathbf{k}_t(\mathbf{x})^T & \mathbf{K}_t + \sigma^2\mathbf{I} \end{bmatrix} \right\}, \tag{4.7}$$

όπου $(\mathbf{f}_0)_i = f_0(x_i)$, $[\mathbf{K}_t]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ και $(\mathbf{k}_t(\mathbf{x}))_i = k(\mathbf{x}_i, \mathbf{x})$, για $i = 1, 2, \ldots, t$. Στη συνέχεια επικαλούμαστε το νόμο Bayes για να βρούμε την εκ των υστέρων πιθανότηνα της F δεδομένου των παρατηρούμενων δεδομένων:

$$\begin{aligned} (F(\cdot) \mid Y_t) &\sim \mathcal{N}\left\{\hat{F}_t(\cdot), P_t(\cdot,\cdot)\right\}, \quad \text{όπου} \\ \hat{F}_t(\mathbf{x}) &= f_0(\mathbf{x}) + \mathbf{k}_t(\mathbf{x})^T(\mathbf{K}_t + \sigma^2\mathbf{I})^{-1}(Y_t - \mathbf{f}_0), \\ P_t(\mathbf{x},\mathbf{x}') &= k(\mathbf{x},\mathbf{x}') - \mathbf{k}_t(\mathbf{x})^T(\mathbf{K}_t + \sigma^2\mathbf{I})^{-1}\mathbf{k}_t(\mathbf{x}'). \end{aligned} \tag{4.8}$$

4.4 Μπεϋσιανή Προσέγγιση της Εκτιμήσης Συνάρτησης Αξίας

Μια σταθερή ή μόνιμη πολιτική (stationary policy) $\mu(\cdot|\mathbf{x}) \in \mathcal{P}(\mathcal{U})$ είναι μια χρονικά ανεξάρτητη απεικόνιση καταστάσεων σε πιθανότητες επιλογής ενεργειών. Δοθέντος μιας πολιτικής, η κατανομή πιθανότητας μετάβασης κατάστασης ορίζεται ως εξής:

$$p^\mu(\mathbf{x}'|\mathbf{x}) = \int_{\mathcal{U}} \mu(\mathbf{u}|\mathbf{x}) p(\mathbf{x}'|\mathbf{u},\mathbf{x}) d\mathbf{u}. \tag{4.9}$$

Ως εκ τούτου, για μια σταθερή πολιτική μ και μια σταθερή αρχική κατάσταση $\mathbf{x}_0$, η πιθανότητα παρατήρησης της ακολουθίας καταστάσεων $\xi_t = \mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t$ είναι $\mathbf{P}(\xi_t) = p_0(\mathbf{x}_0)\prod_{i=1}^t p^\mu(\mathbf{x}_i|\mathbf{x}_{i-1})$. Όπως προαναφέρθηκε, μια χρήσιμη ποσότητα είναι η εκπτώμενη απολαβή (discount return). Η εκπτώμενη απολαβή είναι μια στοχαστική διαδικασία, η οποία ορίζεται ως:

$$D^\mu(\mathbf{x}) = \sum_{i=0}^{\infty} \gamma^i R(\mathbf{x}_i)|(\mathbf{x}_0 = \mathbf{x}), \quad \text{όπου } \mathbf{x}_{i+1} \sim p^\mu(\cdot|\mathbf{x}_i) \text{ για όλα τα } i \geq 0. \tag{4.10}$$

Η τυχαιότητα στην $D^\mu(\mathbf{x}_0)$, για μια οποιαδήποτε κατάσταση $\mathbf{x}_0$, οφείλεται στη στοχαστικότητα της ακολουθίας των καταστάσεων που ακολουθούν την κατάσταση $\mathbf{x}_0$ και στη τυχαιότητα των ανταμοιβών $R(\mathbf{x}_0), R(\mathbf{x}_1), \ldots$κ.λ.π.

Η Εξίσωση 4.10 μαζί με τη σταθερότητα της Μαρκοβιανής διαδικασίας απόφασης παράγουν τον ακόλουθο αναδρομικό τύπο

$$D^\mu(\mathbf{x}) = R(\mathbf{x}) + \gamma D^\mu(\mathbf{x}'), \quad \text{όπου } \mathbf{x}' \sim p^\mu(\cdot|\mathbf{x}). \tag{4.11}$$

Ορίζουμε τον τελεστή μέσης τιμής $\mathbf{E}_\xi$ ως τη μέση τιμή για όλες τις πιθανές μεταβάσεις και όλες τις πιθανές ανταμοιβές που λαμβάνονται. Αυτό μας δίνει τη δυνατότητα να ορίσουμε τη συνάρτηση αξίας $V^\mu(\mathbf{x})$, ως το αποτέλεσμα της εφαρμογής του τελεστή μέσης τιμής στην εκπτώμενη απολαβή $D^\mu(\mathbf{x})$:

$$V^\mu(\mathbf{x}) = \mathbf{E}_\xi D^\mu(\mathbf{x}) \tag{4.12}$$

Έτσι, εφαρμόζοντας τον τελεστη $\mathbf{E}_\xi$ στην Εξίσωση 4.11 παίρνουμε:

$$\begin{aligned} V^\mu(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{E}_\xi D^\mu(\mathbf{x}) &= \mathbf{E}_\xi[R(\mathbf{x}) + \gamma D^\mu(\mathbf{x}')] \\ &= \bar{R}(\mathbf{x}) + \gamma \mathbf{E}_{x'|x} \mathbf{E}_\mu[D^\mu(\mathbf{x}'|\mathbf{x}')] \\ &= \bar{R}(\mathbf{x}) + \gamma \mathbf{E}_{x'|x} V^\mu(\mathbf{x}'), \end{aligned}$$

όπου

$$\begin{aligned} \mathbf{E}_{x'|x} V^\mu(\mathbf{x}') &= \int_{\mathcal{X}} p^\mu(\mathbf{x}'|\mathbf{x}) V^\mu(\mathbf{x}') d\mathbf{x}', \text{ και} \\ \bar{R}(\mathbf{x}) &= \int_{\mathbf{R}} r q(r|\mathbf{x}) dr, \end{aligned}$$

όπου $q(r|\mathbf{x})$ είναι η πιθανότητα στη κατάσταση $\mathbf{x}$ να πάρουμε την r ως ανταμοιβή. Η ισότητα που μόλις αποδείξαμε, δηλαδή η

$$V^\mu(\mathbf{x}) = \bar{R}(\mathbf{x}) + \gamma \mathbf{E}_{x'|x} V^\mu(\mathbf{x}') \forall \mathbf{x} \in \mathcal{X}, \tag{4.13}$$

είναι η εκδοχή σταθερής πολιτικής της εξίσωσης Bellman (Bellman, 1957) [1].

Πιο πάνω σε αυτή την ενότητα είδαμε ότι η συνάρτηση αξίας V είναι το αποτέλεσμα που λαμβάνουμε παίρνοντας τη μέση τιμή της εκπτώμενης απολαβής D σε σχέση με τη τυχαιότητα των μεταβάσεων και των ανταμοιβών που λαμβάνονται. Στη κλασσική προσέγγιση η $V(\cdot)$ δεν είναι πλέον τυχαία, δεδομένου πως είναι η πραγματική, εν τουτις άγνωστη συνάρτηση αξίας. Υιοθετώντας τη Μπεϋσιανή προσέγγιση, μπορούμε να θεωρήσουμε τη συνάρτηση αξίας V σαν μια τυχαία μεταβλητή, αναθέτοντας της επιπλέον τυχαιότητα που οφείλεται στην υποκειμενική αβεβαιότητα μας σχετικά με το μοντέλο μετάβασης (p,q). Δεν γνωρίζουμε ποιές είναι οι πραγματικές κατανομές των p και q, που σημαίνει πως δεν είμαστε βέβαιοι για τη πραγματική συνάρτηση αξίας. Προηγούμενες προσπάθειες εφαρμογής της Μπεϋσιανής λογικής στην ενισχυτική μάθηση, μοντελοποιούσαν αυτή την αβεβαιότητα τοποθετώντας εκ των προτέρων κατανομές στα μοντέλα μετάβασης και ανταμοιβής (p,q) και εφαρμόζοντας τον κανόνα Bayes για να υπολογίσουν την εκ των υστέρων κατανομή

βασιζόμενοι στις παρατηρούμενες μεταβάσεις. Το σημαντικότερο μειονέκτημα αυτής της προσέγγισης ήταν πως οι αλγόριθμοι που προέκυπταν περιοριζόταν στην επίλυση Μαρκοβιανών διαδικασιών απόφασης σε πεπερασμένο χώρο καταστάσεων και ενεργειών. Μια λύση σε αυτό το πρόβλημα είναι να μοντελοποιήσουμε την αβεβαιότητα μας σχετικά με τη ΜΔΑ χρησιμοποιώντας μια εκ των προτέρων κατανομή απευθείας στη V. Αυτό επιτυγχάνεται μοντελοποιώντας τη V ως μια στοχαστική διαδικασία και πιο συγκεκριμένα ως μια Γκαουσιανή διαδικασία.

Στη συνέχεια εξετάζουμε την ανάλυση της εκπτώμενης απολαβής στο μέσο (την αξία) της και σε μια μηδενικού μέσου διαφορά ΔV:

$$D(\mathbf{x}) = \mathbf{E}_\xi D(\mathbf{x}) + (D(\mathbf{x}) - \mathbf{E}_\xi D(\mathbf{x})) \stackrel{\text{def}}{=} V(\mathbf{x}) + \Delta V(\mathbf{x}). \tag{4.14}$$

Αυτή η ανάλυση είναι χρήσιμη, καθώς διαχωρίζει τις δύο πηγές προέλευσης της αβεβαιότητας που κληρονομούνται στην εκπτώμενη ανταμοιβή D. Για ένα γνωστό μοντέλο ΜΔΑ, η V είναι μια ντετερμινιστική συνάρτηση και η τυχαιώτητα στη D, μοντελοποιείται από την ΔV και οφείλεται στην εσωτερική τυχαιώτητα των μεταβάσεων που παράγονται από τη ΜΔΑ και τη πολιτική. Από την αλλή πλευρά, σε μια ΜΔΑ όπου οι μεταβάσεις και οι ανταμοιβές είναι ντετερμινιστικές αλλά παρόλα αυτά άγνωστες, η ΔV είναι ντετερμινιστική και η τυχαιώτητα στη D οφείλεται στην εξωτερική Μπεϋσιανή αβεβαιότητα που μοντελοποιείται από τη στοχαστική διαδικασία V.

4.4.1 Μοντέλο Γκαουσιανής Διαδικασίας για Συναρτήσεις Αξίας

Η αξία κατάστασης V είναι το αποτέλεσμα που λαμβάνουμε αν πάρουμε τη μέση τιμή της εκπτώμενης απολαβής D σε σχέση με την τυχαιώτητα των μεταβάσεων και των ανταμοιβών που λαμβάνουμε. Αντικαθιστώντας την Εξίσωση 4.14 στην Εξίσωση 4.11 και αναδιατάσσοντας έχουμε:

$$R(\mathbf{x}) = V(\mathbf{x}) - \gamma V(\mathbf{x}') + N(\mathbf{x}, \mathbf{x}'),$$
$$\text{όπου } \mathbf{x}' \sim p^\mu(\cdot|\mathbf{x}), \text{ και } N(\mathbf{x}, \mathbf{x}') \stackrel{\text{def}}{=} \Delta V(\mathbf{x}) - \gamma \Delta V(\mathbf{x}'). \tag{4.15}$$

Υποθέτωντας μια ακολουθία μετάβασης $\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_t$, γράφουμε τις εξισώσεις (4.15) για αυτά τα παραδείγματα, με αποτέλεσμα να προκύπτει το ακόλουθο σύνολο t εξισώσεων:

$$R(\mathbf{x}_i) = V(\mathbf{x}_i) - \gamma V(\mathbf{x}_{i+1}) + N(\mathbf{x}_i, \mathbf{x}_{i+1}) \text{ για } i = 0, \ldots, t-1. \tag{4.16}$$

Στη συνέχεια ορίζουμε τα διανύσματα:

$$R_t = (R(\mathbf{x}_0), R(\mathbf{x}_1), \ldots, R(\mathbf{x}_t))^T,$$
$$V_t = (V(\mathbf{x}_0), V(\mathbf{x}_1), \ldots, V(\mathbf{x}_t))^T,$$
$$N_t = (N(\mathbf{x}_0, \mathbf{x}_1), \ldots, N(\mathbf{x}_{t-1}, \mathbf{x}_t))^T. \tag{4.17}$$

Χρησιμοποιώντας τους παραπάνω ορισμούς, το σύνολο των εξισώσεων 4.16 μπορεί να γραφεί συνοπτικά ως:

$$R_{t-1} = \mathbf{H}_t V_t + \mathbf{H}_t \Delta V_t = \mathbf{H}_t V_t + N_t, \tag{4.18}$$

όπου

$$\mathbf{H}_t = \begin{bmatrix} 1 & -\gamma & 0 & \cdots & 0 \\ 0 & 1 & -\gamma & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & 1 & -\gamma \end{bmatrix}. \tag{4.19}$$

Η εξίσωση 4.18 είναι η βάση πάνω στην οποία στηρίζεται η Μπεϋσιανή προσέγγιση που περιγράφεται σε αυτό το κεφάλαιο. Για να ορίσουμε πλήρως ένα ολοκληρωμένο πιθανοτικό γεννητικό μοντέλο, χρειάζεται να προσδιορίσουμε την κατανομή της διαδικασίας N_t του θορύβου. Αυτό επιτυγχάνεται μοντελοποιώντας τις διαφορές $\Delta V_t = (\Delta V(x_0), \ldots, \Delta V(x_t))^T$ ως στοχαστικό Γκαουσιανό θόρυβο. Συγκεκριμένα, αυτό σημαίνει πως η κατανομή του διανύσματος ΔV_t προσδιορίζεται ολοκληρωτικά από το μέσο και τη συνδιακυμανσή της. Επίσης υποθέτουμε πως οι διαφορές $\Delta V(x_i)$ είναι ανεξάρτητες μεταξύ τους. Εξ'ορισμού $\mathbf{E}_\mu[\Delta V(\mathbf{x})] = 0$ για όλα τα $\mathbf{x}$, έτσι έχουμε $\mathbf{E}_\mu[N(\mathbf{x}_i, \mathbf{x}_{i+1})] = 0$. Στρεφόμενοι προς τη συνδιακύμανση, έχουμε

$$\mathbf{E}_\mu[N(\mathbf{x}_i, \mathbf{x}_{i+1})N(\mathbf{x}_j, \mathbf{x}_{j+1})] = \mathbf{E}_\mu[(\Delta V(\mathbf{x}_i) - \gamma\Delta V(\mathbf{x}_{i+1}))(\Delta V(\mathbf{x}_j) - \gamma\Delta V(\mathbf{x}_{j+1}))].$$

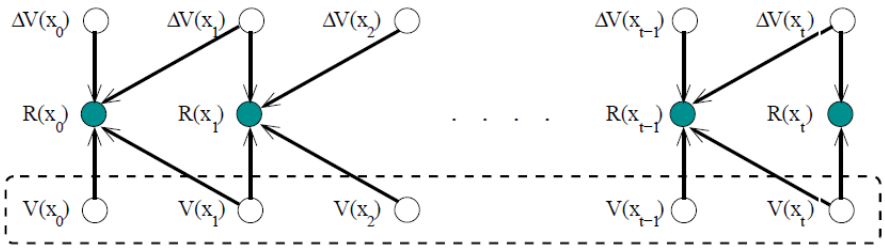
Συμφωνα με την παραδοχή σχετικά με την ανεξαρτησία των διαφόρων, $\mathbf{E}_\mu[\Delta V(\mathbf{x}_i)\Delta V(\mathbf{x}_j)] = 0$ για $i \neq j$. Αντίθετα, η $\mathbf{E}_\mu[\Delta V(\mathbf{x}_i)^2] = \mathbf{Var}_\mu[D(\mathbf{x}_i)]$ είναι γενικά μεγαλύτερη από μηδέν. Ορίζοντας $\sigma_i^2 = \mathbf{Var}[D(\mathbf{x}_i)]$, η κατανομή του ΔV_t δίνεται ως εξής:

$$\Delta V_t \sim \mathcal{N}(\mathbf{0}, diag(\sigma_t)),$$

όπου $\sigma_t = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_t^2)^T$ και $diag(\cdot)$ είναι ένα διαγώνιος πίνακα, τα διαγώνια στοιχεία του οποίου είναι οι μεταβλητές του διανύσματος που παίρνει ως όρισμα. Γνωρίζουμε ότι η οικογένεια των Γκαουσιανών κατανομών είναι κλειστή ως προς διάφορους γραμμικούς μετασχηματισμούς. Για το λόγο αύτο επειδή η ΔV_t ακολουθεί μια Γκαουσιανή κατανομή και $N_t = \mathbf{H}_t\Delta V_t$, τότε η N_t θα ακολουθεί μια κανονική κατανομή $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t)$ με,

$$\begin{aligned} \mathbf{\Sigma}_t &= \mathbf{H}_t diag(\sigma_t)\mathbf{H}_t^\top \\ &= \begin{bmatrix} \sigma_0^2 + \gamma^2\sigma_1^2 & -\gamma\sigma_1^2 & 0 & \cdots & 0 & 0 \\ -\gamma\sigma_1^2 & \sigma_1^2 + \gamma^2\sigma_2^2 & -\gamma\sigma_2^2 & 0 & \cdots & 0 \\ 0 & -\gamma\sigma_2^2 & \sigma_2^2 + \gamma^2\sigma_3^2 & \ddots & & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \vdots & & \ddots & \ddots & -\gamma\sigma_{t-1}^2 \\ 0 & 0 & \cdots & 0 & -\gamma\sigma_{t-1}^2 & \sigma_{t-1}^2 + \gamma^2\sigma_t^2 \end{bmatrix} \end{aligned} \tag{4.20}$$

Το Σχήμα 4.2 απεικονίζει τις υπό όρους σχέσεις ανεξαρτησίας ανάμεσα στις κρυφές μεταβλητές αξίας $V(\mathbf{x}_i)$, τις μεταβλητές θορύβου $\Delta V(\mathbf{x}_i)$ και τις παρατηρούμες ανταμοιβές $R(\mathbf{x}_i)$. Σε αντίθεση με την Γκαουσιανή παλινδρόμηση, υπάρχουν ακμές που συνδέουν μεταβλητές από διαφορετικά χρονικά βήματα, κάνοντας τη σειρά των δειγμάτων σημαντική. Αξίζει επίσης να σημειωθεί πως στη τελευταία κατάσταση κάθε επεισοδίου, η $R(\mathbf{x}_t)$ εξαρτάται μόνο από τη $V(\mathbf{x}_t)$ και την $\Delta V(\mathbf{x}_t)$.



Σχήμα 4.2: Απεικόνιση των υπό συνθήκη σχέσεων ανεξαρτησίας μεταξύ των μεταβλητών

Η Εξίσωση 4.18 μαζί με μια Γκαουσιανή εκ των προτέρων κατανομή στη V και μια Γκαουσιανή κατανομή θορύβου, ορίζουν ενα γραμμικό στατιστικό μοντέλο (linear statistical model)(Scharf, 1991) [12] που συνδέει τις στοχαστικές διαδικασίες της αξίας και της ανταμοιβής. Έτσι, μπορεί να χρησιμοποιήθει ο κανόνας του Bayes για τον υπολογισμό της εκ των υστέρων κατανομής της V δεδομένου της παρατηρούμενης ακολουθίας ανταμοιβών.

Στη συνέχεια υπολογίζουμε τις εκ των υστέρων στιγμές, δεδομένης της ακολουθίας κατάστασης-ανταμοιβής έως τη χρονική στιγμή t. Η εφαρμογή του κανόνα Bayes παράγει τη δεσμευμένη κατανομή της αξίας σε ένα σημείο $\mathbf{x}$, δοθέντος μιας ακολουθίας παρατηρούμενων ανταμοιβών $\mathbf{r}_{t-1} = (r_0, \ldots, r_{t-1})$ η οποία είναι κανονική:

$$(V(\mathbf{x})|R_{t-1} = \mathbf{r}_{t-1}) \sim \mathcal{N}\left\{\hat{V}_t(\mathbf{x}), P_t(\mathbf{x}, \mathbf{x}')\right\}, \tag{4.21}$$

όπου

$$\hat{V}_t(\mathbf{x}) = \mathbf{k}_t(\mathbf{x})^\top \mathbf{H}_t^\top \mathbf{Q}_t \mathbf{r}_{t-1}, \tag{4.22}$$
$$P_t(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_t(\mathbf{x})^\top \mathbf{H}_t^\top \mathbf{Q}_t \mathbf{H}_t \mathbf{k}_t(\mathbf{x}'), \tag{4.23}$$
$$\mathbf{Q}_t = (\mathbf{H}_t \mathbf{K}_t \mathbf{H}_t^\top + \mathbf{\Sigma}_t)^{-1},$$
$$\mathbf{\Sigma}_t = \mathbf{Cov}[N_t].$$

Οι παραπάνω εκφράσεις μπορουν να γραφούν σε μια πιό συνοπτική μορφή. Αυτο επιτυγχάνεται διαχωρίζοντας τους σταθερού όρους, από τους όρους που εκτιμούνται (παράμετροι):

$$\hat{V}_t(\mathbf{x}) = \boldsymbol{\alpha}_t^\top \mathbf{k}_t(\mathbf{x}),$$
$$P_t(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_t(\mathbf{x})^\top \mathbf{C}_t \mathbf{k}_t(\mathbf{x}'), \tag{4.24}$$

όπου $\boldsymbol{\alpha}_t = \mathbf{H}_t^\top \mathbf{Q}_t \mathbf{r}_{t-1}$ και $\mathbf{C}_t = \mathbf{H}_t^\top \mathbf{Q}_t \mathbf{H}_t$ οι οποίες είναι ανεξάρτητες τα $\mathbf{x}$, $\mathbf{x}'$.

Εξαιτίας της ειδικής μορφής του πίνακα συνδιακύμανσης $\mathbf{\Sigma}_t$, μπορούμε να εισάγουμε επαναληπτικές σχέσεις για τις παραμέτρους $\boldsymbol{\alpha}_t$ και το $\mathbf{C}_t$. Οι κανόνες ενημερώσεις γίνονται ως εξής ($\forall t$):

$$\boldsymbol{\alpha}_t = \begin{pmatrix} \boldsymbol{\alpha}_{t-1} \\ 0 \end{pmatrix} + \frac{\mathbf{c}_t}{s_t} d_t, \quad \mathbf{C}_t = \begin{bmatrix} \mathbf{C}_{t-1} & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{bmatrix} + \frac{1}{s_t} \mathbf{c}_t \mathbf{c}_t^\top,$$

όπου

$$\mathbf{c}_t = \frac{\gamma\sigma_{t-1}^2}{s_{t-1}}\begin{pmatrix}\mathbf{c}_{t-1}\\ 0\end{pmatrix} + \mathbf{h}_t - \begin{pmatrix}\mathbf{C}_{t-1}\boldsymbol{\Delta}\mathbf{k}_t\\ 0\end{pmatrix},$$

$$d_t = \frac{\gamma\sigma_{t-1}^2}{s_{t-1}}d_{t-1} + r_{t-1} - \boldsymbol{\Delta}\mathbf{k}_t^\top\boldsymbol{\alpha}_{t-1},$$

$$s_t = \sigma_{t-1}^2 + \gamma^2\sigma_t^2 - \frac{\gamma^2\sigma_{t-1}^4}{s_{t-1}} + \Delta k_{tt} - \boldsymbol{\Delta}\mathbf{k}_t^\top\mathbf{C}_{t-1}\boldsymbol{\Delta}\mathbf{k}_t + \frac{2\gamma\sigma_{t-1}^2}{s_{t-1}}\mathbf{c}_{t-1}^\top\boldsymbol{\Delta}\mathbf{k}_t.$$

Παραπάνω κάνουμε χρήση των ακόλουθων ορισμών:

$$\mathbf{h}_t = (0,\ldots,1,-\gamma)^\top,$$

$$\boldsymbol{\Delta}\mathbf{k}_t = \mathbf{k}_{t-1}(\mathbf{x}_{t-1}) - \gamma\mathbf{k}_{t-1}(\mathbf{x}_t),$$

$$\Delta k_{tt} = k(\mathbf{x}_{t-1},\mathbf{x}_{t-1}) - 2\gamma k(\mathbf{x}_{t-1},\mathbf{x}_t) + \gamma^2 k(\mathbf{x}_t,\mathbf{x}_t).$$

Η επαναληπτική διαδικασία αρχικοποιείται ως εξής:

$$\boldsymbol{\alpha}_0 = 0, \mathbf{C}_0 = 0, \mathbf{c}_0 = 0, d_0 = 0, \frac{1}{s_0} = 0.$$

Ο ψευδοκώδικας της παραπάνω μεθοδολογιάς παρουσιάζεται στον Αλγόριθμο 10.

Αλγόριθμος 10 Επαναληπτικός Monte-Carlo GPTD Αλγόριθμος

Αρχικοποίηση $\boldsymbol{\alpha}_0 = 0, \mathbf{C}_0 = 0, \mathbf{c}_0 = 0, d_0 = 0, \frac{1}{s_0} = 0$
for $t = 1, 2, \ldots$ **do**
  Παρατηρούμε $\mathbf{x}_{t-1}, r_{t-1}, \mathbf{x}_t$
  $\mathbf{h}_t = (0,\ldots,1,-\gamma)^\top$
  $\boldsymbol{\Delta}\mathbf{k}_t = \mathbf{k}_{t-1}(\mathbf{x}_{t-1}) - \gamma\mathbf{k}_{t-1}(\mathbf{x}_t)$
  $\Delta k_{tt} = k(\mathbf{x}_{t-1},\mathbf{x}_{t-1}) - 2\gamma k(\mathbf{x}_{t-1},\mathbf{x}_t) + \gamma^2 k(\mathbf{x}_t,\mathbf{x}_t)$
  $\mathbf{c}_t = \frac{\gamma\sigma_{t-1}^2}{s_{t-1}}\begin{pmatrix}\mathbf{c}_{t-1}\\ 0\end{pmatrix} + \mathbf{h}_t - \begin{pmatrix}\mathbf{C}_{t-1}\boldsymbol{\Delta}\mathbf{k}_t\\ 0\end{pmatrix}$
  $d_t = \frac{\gamma\sigma_{t-1}^2}{s_{t-1}}d_{t-1} + r_{t-1} - \boldsymbol{\Delta}\mathbf{k}_t^\top\boldsymbol{\alpha}_{t-1}$
  $s_t = \sigma_{t-1}^2 + \gamma^2\sigma_t^2 - \frac{\gamma^2\sigma_{t-1}^4}{s_{t-1}} + \Delta k_{tt} - \boldsymbol{\Delta}\mathbf{k}_t^\top\mathbf{C}_{t-1}\boldsymbol{\Delta}\mathbf{k}_t + \frac{2\gamma\sigma_{t-1}^2}{s_{t-1}}\mathbf{c}_{t-1}^\top\boldsymbol{\Delta}\mathbf{k}_t$
  $\boldsymbol{\alpha}_t = \begin{pmatrix}\boldsymbol{\alpha}_{t-1}\\ 0\end{pmatrix} + \frac{\mathbf{c}_t}{s_t}d_t$
  $\mathbf{C}_t = \begin{bmatrix}\mathbf{C}_{t-1} & \mathbf{0}\\ \mathbf{0}^\top & 0\end{bmatrix} + \frac{1}{s_t}\mathbf{c}_t\mathbf{c}_t^\top$
  $\mathcal{D}_t = \mathcal{D}_{t-1}\cup\{\mathbf{x}_t\}$
return $\boldsymbol{\alpha}_t, \mathbf{C}_t, \mathcal{D}_t$

4.4.2 Εργασίες με Επεισόδια

Στην ενότητα αυτή θα δούμε πως το GPTD [5], [6], [7] μοντέλο που περιγράφηκε παραπάνω, μπορεί να τροποποιηθεί, έτσι ώστε να χειρίζεται εργασίες μάθησης με επεισόδια. Στις συνεχής εργασίες (continual tasks), ο πράκτορας τοποθετείται σε κάποια αρχική κατάσταση και στη συνέχεια αφήνεται να περιπλανηθεί επ'αόριστον. Αντίθετα, στις εργασίες με επεισόδια, ο χώρος καταστάσεων περιέχει μια απορροφητική κατάσταση. Ο πράκτορας καταλήγει σε αυτή τη κατάσταση, μετά από πεπερασμένο αριθμό βημάτων. Όταν φθάνουμε σε μια τερματική κατάσταση, το επεισόδιο τερματίζει και ο πράκτορας τοποθετείται τυχαία (συνήθως) σε μια νέα κατάσταση για να ξεκινήσει ένα νέο επεισόδιο.

Όταν φθάνουμε σε μια τερματική κατάσταση, όλα οι μεταγενέστερες ανταμοιβές παραλείπονται. Ως εκ τούτου, τόσο η εκπτώμενη απολαβή όσο και η αξία μιας τερματικής κατάστασης παραλείπονται. Αυτό έχει ως αποτέλεσμα, η εκπτώμενη απολαβή και η αξία της κατάστασης που προηγείται της τερματικής κατάστασης, να είναι ίσες με τη ανταμοιβή και τη αναμενόμενη ανταμοιβή αυτής της κατάστασης, αντίστοιχα. Ειδικότερα, εάν το τελικό βήμα σε ένα επεισόδιο είναι τη χρονική στιγμή t ($\mathbf{x}_{t+1}$ είναι η τερματική κατάσταση), η τελευταία εξίσωση στο σύστημα εξισώσεων 4.18 γράφεται ως:

$$R(\mathbf{x}_t) = V(\mathbf{x}_t) + N(\mathbf{x}). \tag{4.25}$$

Ως εκ τούτου, για το πρώτο επεισόδιο, ο $\mathbf{H}_{t+1}$ είναι ένας $(t+1) \times (t+1)$ τετραγωνικός αντιστρέψιμος πίνακας (η οριζουσά του είναι ίση με ένα),

$$\mathbf{H}_{t+1} = \begin{bmatrix} 1 & -\gamma & 0 & \cdots & 0 \\ 0 & 1 & -\gamma & \cdots & 0 \\ \cdots & & & & \cdots \\ 0 & 0 & \cdots & 1 & -\gamma \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}. \tag{4.26}$$

Ο πίνακας συνδιακύμανσης του θορύβου για το πρώτο επεισόδιο γίνεται

$$\begin{aligned} \mathbf{\Sigma}_{t+1} &= \mathbf{H}_{t+1}\mathrm{diag}(\sigma_{\mathbf{t}})\mathbf{H}_{t+1}^{\top} \\ &= \begin{bmatrix} \sigma_0^2 + \gamma^2\sigma_1^2 & -\gamma\sigma_1^2 & 0 & \cdots & 0 & 0 \\ -\gamma\sigma_1^2 & \sigma_1^2 + \gamma^2\sigma_2^2 & -\gamma\sigma_2^2 & 0 & \cdots & 0 \\ 0 & -\gamma\sigma_2^2 & \sigma_2^2 + \gamma^2\sigma_3^2 & \ddots & & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \vdots & & \ddots & \ddots & -\gamma\sigma_t^2 \\ 0 & 0 & \cdots & 0 & -\gamma\sigma_t^2 & \sigma_t^2 \end{bmatrix} \end{aligned} \tag{4.27}$$

Τελικά, το σύνολο των εξισώσεων του μοντέλου είναι

$$R_t = \mathbf{H}_{t+1}V_t + N_{t+1}. \tag{4.28}$$

Μετά από μια ακολουθία επεισοδίων μάθησης, που το καθένα τελειώνει σε μια τερματική κατάσταση, ο $\mathbf{H}_{t+1}$ είναι ένας τετραγωνικός block-διαγώνιος πίνακας. Ο πίνακας συνδιακύμανσης του θορύβου διατηρεί μια αντίστοιχη block-διαγώνια δομή, με το κάθε block του $\mathbf{\Sigma}_t$ να είναι ένας τριδιαγώνιος (tridiagonal) πίνακας της μορφής 4.27.

Προκειμένου να παράγουμε τις ενημερώσεις που αντιστοιχούν στη τελευταία μετάβαση κάθε επεισοδίου, είναι χρήσιμο να παρατηρήσουμε ότι οι Εξισώσεις 4.25, 4.26 και 4.28 θα μπορούσαν να ληφθούν απλά, αν προσωρινά τοποθετήσουμε το παράγοντα έκπτωσης γ στο 0, μόνο για τη μετάβαση από την $\mathbf{x}_t$ στη τερματική κατάσταση. Ωστόσο, εφόσον οι εξισώσεις ενημέρωσης περιέχουν παράγοντες έκπτωσης για δύο διαδοχικές χρονικές στιγμές, πρέπει να ληφθεί κάποια επιπλέον μέριμνα. Βάζοντας ως ετικέτα σε κάθε γ το αντίστοιχο χρονικό βήμα, παίρνουμε το ακόλουθο σύνολο εξισώσεων για την ενημέρωση που αντιστοιχεί στη τελική μετάβαση ενός επεισοδίου (ορίζουμε $\mathbf{e} = (0, \cdots, 0, 1)^\top$):

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + \frac{\mathbf{c}_{t+1}}{s_{t+1}} d_{t+1}, \quad \mathbf{C}_{t+1} = \mathbf{C}_t + \frac{1}{s_{t+1}} \mathbf{c}_{t+1} \mathbf{c}_{t+1}^\top, \tag{4.29}$$

όπου

$$\begin{aligned}
\mathbf{c}_{t+1} &= \frac{\gamma \sigma_t^2}{s_t} \mathbf{c}_t + \mathbf{e} - \mathbf{C}_t \mathbf{k}_t(\mathbf{x}_t), \\
d_{t+1} &= \frac{\gamma \sigma_t^2}{s_t} d_t + r_t - \mathbf{k}_t(\mathbf{x}_t)^\top \boldsymbol{\alpha}_t, \\
s_{t+1} &= \sigma_t^2 + \mathbf{k}_t(\mathbf{x}_t)^\top \left(\mathbf{c}_{t+1} + \frac{\gamma \sigma_t^2}{s_t} \mathbf{c}_t \right) - \frac{\gamma^2 \sigma_t^4}{s_t}.
\end{aligned} \tag{4.30}$$

4.4.3 Αραιοί Αλγόριθμοι Άμεσης Απόκρισης

Ο Αλγόριθμος 10 που περιγράφηκε παραπάνω, αν και επαναληπτικός, δεν υπάγεται στους αλγορίθμους άμεσης απόκρισης. Αυτό οφείλεται στο γεγονός ότι το κόστος υπολογισμού της ενημέρωσης το χρονικό βήμα t είναι $O(t^2)$ τόσο σε χρόνο όσο και σε μνήμη. Στους αλγορίθμους άμεσης απόκρισης ή πραγματικού χρόνου, απαιτούμε το κόστος υπολογισμού το χρονικό βήμα t να είναι ανεξάρτητο από το t. Υπάρχουν δύο γενικές προσεγγίσεις για τη προσαρμογή των αλγορίθμων που περιγράφηκαν παραπάνω στη κατηγορία αυτή. Η μια βασίζεται στην εξαγωγή και χρήση παραμετρικών προσεγγίσεων αυτών των GPTD αλγορίθμων. Η άλλη προσέγγιση, που περιγράφεται σε αυτή την ενότητα, βασίζεται σε μια αποτελεσματική ακολουθιακή μέθοδο σποραδικότητας του πυρήνα.

Αυτή η μέθοδος βασίζεται στην ακόλουθη παρατήρηση: Εξαιτίας του θεωρήματος Mercer, η συνάρτηση συνδιακύμανσης πυρήνα $k(\cdot, \cdot)$, μπορεί να εκληφθεί ως ένα εσωτερικό γινόμενο σε έναν γενικά άπειρης διάστασης Hilbert χώρο $\mathcal{H}$. Αυτό σημαίνει πως υπάρχει, μια γενικά μη γραμμική απεικόνιση $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{H}$ για την οποία $\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle_{\mathcal{H}} = k(\mathbf{x}, \mathbf{x}')$. Παρόλο που η διάσταση του $\mathcal{H}$ μπορεί να είναι υπερβολικά υψηλή, η διάσταση του χώρου που ορίζεται από το σύνολο διανυσμάτων $\{\boldsymbol{\phi}(\mathbf{x}_i)\}_{i=0}^t$ είναι το πολύ t. Μπορεί να είναι χαμηλότερη, εάν υπάρχουν γραμμικές εξαρτήσεις ανάμεσα σε διαφορετικά διανύσματα $\boldsymbol{\phi}(\mathbf{x}_i)$. Κατά συνέπεια, οποιαδήποτε έκφραση που περιγράφεται σαν γραμμικός συνδυασμός αυτών των διανυσμάτων, μπορεί να εκφραστεί σε σχέση με ένα αυθαίρετο σύνολο γραμμικών ανεξάρτητων διανυσμάτων που παράγουν το χώρο (βάση του χώρου). Όταν μια τέτοια βάση αποτελείται από ένα υποσύνολο του $\{\boldsymbol{\phi}(\mathbf{x}_i)\}_{i=0}^t$, αναφερόμαστε σε αυτό, καθώς επίσης και στο αντίστοιχο σύνολο των εισόδων του, ως λεξικό (dictionary). Επιπλέον, αντί να κατασκευάζουμε πλήρη λεξικά που παράγουν ακριβώς το χώρο, μπορούμε να χρησιμοποιήσουμε

μερικά ή προσεγγιστικά λεξικά όπου κάθε μια άλλη λέξη (εκτός λεξικού) μπορεί να παραχθεί ως ένας γραμμικός ή μη-γραμμικός συνδυασμός των λέξεων του λεξικού.

Εκμεταλλευόμενοι την ιδέα αυτή, η μέθοδος μας ξεκινά με ένα κενό λεξικό $\mathcal{D}_0 = \{\}$. Έστω μια ακολουθία καταστάσεων $\mathbf{x}_0, \mathbf{x}_1, \ldots$, δηλαδή μια κατάσταση κάθε χρονική στιγμή. Η κατάσταση $\mathbf{x}_t$ εισάγεται στο λεξικό, μονό εάν η εικόνα της, $\boldsymbol{\phi}(\mathbf{x}_t)$, δεν προσεγγίζεται αρκετά ικανοποιητικά από κάποιο συνδυασμό των εικόνων των καταστάσεων $\boldsymbol{\phi}(\tilde{\mathbf{x}}_t)$ που ήδη βρίσκονται στο λεξικό, $\mathcal{D}_{t-1} = \{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{m_{t-1}}\}$. Αυτό επιτυγχάνεται με τον κανόνα ελαχίστων τετραγώνων της $\boldsymbol{\phi}(\mathbf{x}_t)$ από το λεξικό. Έτσι αν υπάρχουν m_{t-1} συντελεστές a_j τέτοιοι ώστε:

$$\left\| \sum_{j=1}^{m_{t-1}} a_j \boldsymbol{\phi}(\tilde{\mathbf{x}}_j) - \boldsymbol{\phi}(\mathbf{x}_t) \right\|^2 \leq \nu, \tag{4.31}$$

όπου το ν είναι ένα θετικό κατώφλι που προσδιορίζει την ακρίβεια της προσέγγισης. Η εύρεση του βέλτιστου διανύσματος $\boldsymbol{a}_t$ μπορεί να επιτευχθεί, λύνοντας ένα πρόβλημα ελαχιστοποίησης,

$$\delta_t \stackrel{\text{def}}{=} \min_{\boldsymbol{a}} \left\| \sum_{j=1}^{m_{t-1}} a_j \boldsymbol{\phi}(\tilde{\mathbf{x}}_j) - \boldsymbol{\phi}(\mathbf{x}_t) \right\|^2. \tag{4.32}$$

Εάν η συνθήκη ALD διατηρείται στη 4.32, η $\boldsymbol{\phi}(\boldsymbol{x}_t)$ μπορεί να προσεγγιστεί από κάποιο γραμμικό συνδυασμό των τρεχόντων μελών του λεξικού, με ένα τετραγωνικό σφάλμα v. Ελαχιστοποιώντας την 4.32, μπορούμε ταυτόχρονα, να ελέγξουμε εάν διατηρείται αυτή η συνθήκη και να βρούμε το βέλτιστο διάνυσμα συντελεστών $\boldsymbol{a}_t$ που την ικανοποιεί, ελαχιστοποιώντας το τετραγωνικό σφάλμα. Αναλύοντας την 4.32 παίρνουμε:

$$\delta_t = \min_{\boldsymbol{a}} \left\{ \sum_{i,j=1}^{m_{t-1}} a_i a_j \langle \boldsymbol{\phi}(\tilde{\mathbf{x}}_i), \boldsymbol{\phi}(\tilde{\mathbf{x}}_j) \rangle - 2 \sum_{j=1}^{m_{t-1}} a_j \langle \boldsymbol{\phi}(\tilde{\mathbf{x}}_j), \boldsymbol{\phi}(\mathbf{x}_t) \rangle + \langle \boldsymbol{\phi}(\mathbf{x}_t), \boldsymbol{\phi}(\mathbf{x}_t) \rangle \right\}. \tag{4.33}$$

Στη συνέχεια αντικαθιστούμε τα εσωτερικά γινόμενα μεταξύ των διανυσμάτων με την συνάρτηση πυρήνα. Ως εκ τούτου, κάνοντας την αντικατάσταση $\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}) \rangle = k(\mathbf{x}, \mathbf{x}')$, παίρνουμε:

$$\begin{aligned} \delta_t &= \min_{\boldsymbol{a}} \left\{ \sum_{i,j=1}^{m_{t-1}} a_i a_j k(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) - 2 \sum_{j=1}^{m_{t-1}} a_j k(\tilde{\mathbf{x}}_j, \mathbf{x}_t) + k(\mathbf{x}_t, \mathbf{x}_t) \right\} \\ &= \min_{\boldsymbol{a}} \{ \boldsymbol{a}^\top \tilde{\mathbf{K}}_{t-1} \boldsymbol{a} - 2 \boldsymbol{a}^\top \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t) + k_{tt} \}, \end{aligned} \tag{4.34}$$

όπου $\tilde{\mathbf{K}}_{t-1}$ είναι ο πίνακας πυρήνα των καταστάσεων του λεξικού τη χρονική στιγμή $t-1$, $\tilde{\mathbf{k}}_{t-1}(\mathbf{x}) = (k(\tilde{\mathbf{x}}_1, \mathbf{x}), \ldots, k(\tilde{\mathbf{x}}_{m_{t-1}}, \mathbf{x}))^\top$ και $k_{tt} = k(\mathbf{x}_t, \mathbf{x}_t)$. Η λύση στο πρόβλημα βελτιστοποίησης 4.34 είναι η $\boldsymbol{a}_t = \tilde{\mathbf{K}}_{t-1}^{-1} \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t)$. Αντικαθιστώντας τη λύση αυτή στην Εξίσωση 4.34, παίρνουμε το τετραγωνικό σφάλμα που προκύπτει από την προσέγγιση:

$$\delta_t = k_{tt} - \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t)^\top \boldsymbol{a}_t = k_{tt} - \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t)^\top \tilde{\mathbf{K}}_{t-1}^{-1} \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t). \tag{4.35}$$

Εάν $\delta_t > \nu$, όπου ν είναι η παράμετρος κατωφλίου, προσθέτουμε το $\mathbf{x}_t$ στο λεξικό. Επίσης ορίζουμε $\boldsymbol{a}_t = (0, \ldots, 1)^\top$ και $\delta_t = 0$, εφόσον η $\boldsymbol{\phi}(\mathbf{x}_t)$ αναπαρίσταται ακριβώς από έναν όρο

του λεξικού (δηλαδή από τον εαυτό του). Εάν $\delta_t \leq \nu$, το λεξικό παραμένει το ίδιο. Είτε έτσι είτε αλλιώς, μπορούμε να διασφαλίσουμε πως όλα τα διανύσματα που αντιστοιχούν στις καταστάσεις μέχρι τη χρονική στιγμή t, μπορούν να προσεγγιστούν από το λεξικό, με μέγιστο τετραγωνικό σφάλμα ν, δηλαδή

$$\boldsymbol{\phi}(\mathbf{x}_i) = \sum_{j=1}^{m_i} a_{i,j}\boldsymbol{\phi}(\tilde{\mathbf{x}}_j) + \boldsymbol{\phi}_i^{res}, \quad \text{όπου } \|\boldsymbol{\phi}_i^{res}\|^2 \leq \nu. \tag{4.36}$$

Εναλλακτικά, το δ_t μπορεί να ερμηνευθεί ως η διακύμανση της αξίας $V(\mathbf{x}_t)$ της παρούσας κατάστασης $\mathbf{x}_t$, δεδομένων των αξιών $V(\tilde{\mathbf{x}}_\mathbf{1}), \ldots, V(\tilde{\mathbf{x}}_{|\mathcal{D}_{t-1}|})$ του τρέχοντος λεξικού. Το $\mathbf{x}_t$ θα προστεθεί στο λεξικό εάν, υποθέτοντας οτι οι αξίες των εγγράφων του τρέχοντος λεξικού είναι γνωστές, η απομένουσα αβεβαιότητα στην αξία του $\mathbf{x}_t$ εξακολουθεί να είναι μεγαλύτερη απο ν. Εάν το $\mathbf{x}_t$ προστεθεί στο λεξικό, αυτή η υπό συνθήκη διακύμανση παραλείπεται και το ίδιο κάνει και το δ_t.

Προκειμένου να είμαστε σε θέση να υπολογίσουμε το $\boldsymbol{a}_t$ σε κάθε χρονική στιγμή, χρειάζεται να ενημερώνουμε το $\tilde{\mathbf{K}}_t^{-1}$ οποτεδήποτε μια καινούρια κατάσταση εισάγεται στο λεξικό. Αυτό επιτυγχάνεται ως εξής:

$$\tilde{\mathbf{K}}_t = \begin{bmatrix} \tilde{\mathbf{K}}_{t-1} & \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t) \\ \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t)^\top & k_{tt} \end{bmatrix} \quad \Rightarrow \quad \tilde{\mathbf{K}}_t^{-1} = \frac{1}{\delta_t}\begin{bmatrix} \delta_t\tilde{\mathbf{K}}_{t-1}^{-1} + \hat{\boldsymbol{a}}_t\hat{\boldsymbol{a}}_t^\top & -\hat{\boldsymbol{a}}_t \\ -\hat{\boldsymbol{a}}_t^\top & 1 \end{bmatrix} \tag{4.37}$$

όπου $\hat{\boldsymbol{a}}_t = \tilde{\mathbf{K}}_{t-1}^{-1}\tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t)$. Παρατηρούμε ότι το $\hat{\boldsymbol{a}}_t$ είναι πανομοιότυπο με το διάνυσμα $\boldsymbol{a}_t$ που υπολογίζεται επιλύοντας την Εξίσωση 4.34 (πρίν την επαναφορά της στο $(0, 0, \ldots, 1)^\top$), έτσι δεν χρειάζεται να το επαναυπολογίσουμε.

Ορίζοντας τους πίνακες $[\mathbf{A}_t]_{i,j} = a_{i,j}$, $\boldsymbol{\Phi}_t = [\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_t)]$ και $\boldsymbol{\Phi}_t^{res} = [\boldsymbol{\phi}_1^{res}, \ldots, \boldsymbol{\phi}_t^{res}]$, μπορούμε να γράψουμε την Εξίσωση 4.36 για όλα τα βήματα μέχρι το t, συνοπτικά ως

$$\boldsymbol{\Phi}_t = \tilde{\boldsymbol{\Phi}}_t\mathbf{A}_t^\top + \boldsymbol{\Phi}_t^{res}. \tag{4.38}$$

Πολλαπλασιάζοντας την 4.38 με τον ανάστροφό της, αναλύουμε τον $t \times t$ πίνακα πυρήνα $\mathbf{K}_t = \boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t$ σε δύο πίνακες:

$$\mathbf{K}_t = \mathbf{A}_t\tilde{\mathbf{K}}_t\mathbf{A}_t^\top + \mathbf{K}_t^{\text{res}}. \tag{4.39}$$

όπου $\tilde{\mathbf{K}}_t = \tilde{\boldsymbol{\Phi}}_t^\top\tilde{\boldsymbol{\Phi}}_t$. Ο πίνακας $\mathbf{A}_t\tilde{\mathbf{K}}_t\mathbf{A}_t^\top$ είναι προσέγγιση τάξης m_t του $\mathbf{K}_t$. Μπορεί να δειχθεί ότι, η νόρμα του πίνακα διαφοράς $\mathbf{K}_t^{\text{res}}$ είναι φραγμένη από πάνω, κατά ένα παράγοντα γραμμικό στο $\sqrt{\nu}$. Κατά συνέπεια, κάνουμε τις ακόλουθες προσεγγίσεις:

$$\mathbf{K}_t \overset{\sqrt{\nu}}{\approx} \mathbf{A}_t\tilde{\mathbf{K}}_t\mathbf{A}_t^\top \quad , \quad \mathbf{k}_t(\mathbf{x}) \overset{\sqrt{\nu}}{\approx} \mathbf{A}_t\tilde{\mathbf{k}}_t(\mathbf{x}). \tag{4.40}$$

Αξίζει να σημειωθεί, ότι το υπολογιστικό κόστος του συγκεκριμένου αλγορίθμου για κάθε χρονικό βήμα είναι $O(m_t^2)$. Υποθέτουμε ότι το m_t δεν εξαρτάται ασυμπτωτικά με το t, αλλά είναι ανεξάρτητο από το χρόνο. Τώρα είμαστε έτοιμοι να ενσωματώσουμε αυτή την αραιή μέθοδο, στις επαναληπτικές ενημερώσεις των εκ των υστέρων GPTD στιγμών

που εξάγαμε στις προηγούμενες υποενότητες. Αντικαθιστώντας τις προσεγγίσεις 4.40 στη λύση 4.24, παίρνουμε:

$$\begin{aligned}\hat{V}_t(\mathbf{x}) &\overset{\sqrt{\nu}}{\approx} \boldsymbol{\alpha}_t^\top \mathbf{A}_t \tilde{\mathbf{k}}_t(\mathbf{x}) = \tilde{\boldsymbol{\alpha}}_t^\top \tilde{\mathbf{k}}_t(\mathbf{x}),\\ P_t(\mathbf{x},\mathbf{x}') &\overset{\sqrt{\nu}}{\approx} k(\mathbf{x},\mathbf{x}') - \mathbf{k}_t(\mathbf{x})^\top \mathbf{A}_t^\top \mathbf{C}_t \mathbf{A}_t \mathbf{k}_t(\mathbf{x}') = k(\mathbf{x},\mathbf{x}') - \tilde{\mathbf{k}}_t(\mathbf{x})^\top \tilde{\mathbf{C}}_t \tilde{\mathbf{k}}_t(\mathbf{x}'),\end{aligned} \tag{4.41}$$

όπου χρησιμοποιούνται οι ορισμοί

$$\begin{aligned}\tilde{\mathbf{H}}_t &= \mathbf{H}_t \mathbf{A}_t,\\ \tilde{\mathbf{Q}}_t &= \left(\tilde{\mathbf{H}}_t \tilde{\mathbf{K}}_t \tilde{\mathbf{H}}_t^\top + \boldsymbol{\Sigma}_t\right)^{-1},\\ \tilde{\boldsymbol{\alpha}}_t &= \tilde{\mathbf{H}}_t^\top \tilde{\mathbf{Q}}_t \mathbf{r}_{t-1},\\ \tilde{\mathbf{C}}_t &= \tilde{\mathbf{H}}_t^\top \tilde{\mathbf{Q}}_t \tilde{\mathbf{H}}_t.\end{aligned} \tag{4.42}$$

Οι παράμετροι που απαιτούνται να αποθηκεύονται και να ενημερώνονται, ώστε να υπολογίζουμε το μέσο και τη διακύμανσή είναι οι $\tilde{\boldsymbol{\alpha}}_t$ και $\tilde{\mathbf{C}}_t$. Οι διαστάσεις των οποίων είναι $m_t \times 1$ και $m_t \times m_t$, αντίστοιχα. Εδώ παραθέτουμε τις ενημερώσεις για το MC-GPTD μοντέλο.

Σε κάθε χρονικό βήμα, η τρέχουσα κατάσταση $\mathbf{x}_t$ ενδέχεται είτε να μην προστεθεί στο λεξικό ($\mathcal{D}_t = \mathcal{D}_{t-1}$) είτε να προστεθεί σε αυτό ($\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{\mathbf{x}_t\}$). Στη δεύτερη περίπτωση, οι διαστάσεις των $\tilde{\boldsymbol{\alpha}}$ και $\tilde{\mathbf{C}}$ αυξάνονται κατά 1. Σε κάθε περίπτωση οι ενημερώσεις, διαφέρουν ελάχιστα μεταξύ τους. Σε κάθε μια από τις δύο περιπτώσεις, χρησιμοποιούμε τον ορισμό:

$$\boldsymbol{\Delta}\tilde{\mathbf{k}}_t = \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_{t-1}) - \gamma \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t).$$

Περίπτωση 1. Το λεξικό παραμένει αμετάβλητο: $\mathcal{D}_t = \mathcal{D}_{t-1}$:

$$\tilde{\boldsymbol{\alpha}}_t = \tilde{\boldsymbol{\alpha}}_{t-1} + \frac{\tilde{\mathbf{c}}_t}{s_t} d_t, \quad \tilde{\mathbf{C}}_t = \tilde{\mathbf{C}}_{t-1} + \frac{1}{s_t} \tilde{\mathbf{c}}_t \tilde{\mathbf{c}}_t^\top, \tag{4.43}$$

όπου

$$\begin{aligned}\tilde{\mathbf{c}}_t &= \frac{\gamma \sigma_{t-1}^2}{s_{t-1}} \tilde{\mathbf{c}}_{t-1} + \tilde{\mathbf{h}}_t - \tilde{\mathbf{C}}_{t-1} \boldsymbol{\Delta}\tilde{\mathbf{k}}_t,\\ d_t &= \frac{\gamma \sigma_{t-1}^2}{s_{t-1}} d_{t-1} + r_{t-1} - \boldsymbol{\Delta}\tilde{\mathbf{k}}_t^\top \tilde{\boldsymbol{\alpha}}_{t-1},\\ s_t &= \sigma_{t-1}^2 + \gamma^2 \sigma_t^2 + \boldsymbol{\Delta}\tilde{\mathbf{k}}_t^\top \left(\tilde{\mathbf{c}}_t + \frac{\gamma \sigma_{t-1}^2}{s_{t-1}} \tilde{\mathbf{c}}_{t-1}\right) - \frac{\gamma^2 \sigma_{t-1}^4}{s_{t-1}},\end{aligned} \tag{4.44}$$

με τους ορισμούς $\tilde{\mathbf{h}}_t = \boldsymbol{a}_{t-1} - \gamma \boldsymbol{a}_t$ και $\Delta k_{tt} = \tilde{\mathbf{h}}_t^\top \boldsymbol{\Delta}\tilde{\mathbf{k}}_t$.

Περίπτωση 2. Στο λεξικό προστίθεται η $\mathbf{x}_t$: $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{\mathbf{x}_t\}$:

$$\tilde{\boldsymbol{\alpha}}_t = \begin{pmatrix} \tilde{\boldsymbol{\alpha}}_{t-1} \\ 0 \end{pmatrix} + \frac{\tilde{\mathbf{c}}_t}{s_t} d_t, \quad \tilde{\mathbf{C}}_t = \begin{bmatrix} \tilde{\mathbf{C}}_{t-1} & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{bmatrix} + \frac{1}{s_t} \tilde{\mathbf{c}}_t \tilde{\mathbf{c}}_t^\top, \tag{4.45}$$

όπου

$$\tilde{\mathbf{c}}_t = \frac{\gamma\sigma_{t-1}^2}{s_{t-1}}\begin{pmatrix}\tilde{\mathbf{c}}_{t-1}\\ 0\end{pmatrix} + \tilde{\mathbf{h}}_t - \begin{pmatrix}\tilde{\mathbf{C}}_{t-1}\boldsymbol{\Delta}\tilde{\mathbf{k}}_t\\ 0\end{pmatrix},$$

$$d_t = \frac{\gamma\sigma_{t-1}^2}{s_{t-1}}d_{t-1} + r_{t-1} - \boldsymbol{\Delta}\tilde{\mathbf{k}}_t^\top\tilde{\boldsymbol{\alpha}}_{t-1},$$

$$s_t = \sigma_{t-1}^2 + \gamma^2\sigma_t^2 + \Delta k_{tt} - \boldsymbol{\Delta}\tilde{\mathbf{k}}_t^\top\tilde{\mathbf{C}}_{t-1}\boldsymbol{\Delta}\tilde{\mathbf{k}}_t + \frac{2\gamma\sigma_{t-1}^2}{s_{t-1}}\tilde{\mathbf{c}}_{t-1}^\top\boldsymbol{\Delta}\tilde{\mathbf{k}}_t - \frac{\gamma^2\sigma_{t-1}^4}{s_{t-1}}, \tag{4.46}$$

με τους ορισμούς

$$\tilde{\mathbf{h}}_t = \begin{pmatrix}\boldsymbol{a}_{t-1}\\ 0\end{pmatrix} - \gamma\boldsymbol{a}_t = \begin{pmatrix}\boldsymbol{a}_{t-1}\\ -\gamma\end{pmatrix}, \quad \Delta k_{tt} = \tilde{\boldsymbol{a}}_{t-1}^\top\left(\tilde{\mathbf{k}}_{t-1}(\mathbf{x}_{t-1}) - 2\gamma\tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t)\right) + \gamma^2 k_{tt}.$$

Ο ψευδοκώδικας του συγκεκριμένου αλγορίθμου παρουσιάζεται στον Αλγόριθμο 11

Αλγόριθμος 11 Επαναληπτικός Αραιός Monte-Carlo GPTD Αλγόριθμος

Παράμετροι: ν

Αρχικοποίηση: $\mathcal{D}_0 = \{\mathbf{x}_0\}$, $\tilde{\mathbf{K}}_0^{-1} = 1/k(\mathbf{x}_0, \mathbf{x}_0)$, $\boldsymbol{a}_0 = (1)$, $\tilde{\boldsymbol{\alpha}}_0 = 0$, $\tilde{\mathbf{C}}_0 = 0$, $\tilde{\mathbf{c}}_0 = 0$, $d_0 = 0, 1/s_0 = 0$

for $t = 1, 2, \ldots$ **do**

Παρατηρούμε: $\mathbf{x}_{t-1}$, r_{t-1}, $\mathbf{x}_t$

$\boldsymbol{a}_t = \tilde{\mathbf{K}}_{t-1}^{-1} \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t)$

$\delta_t = k(\mathbf{x}_t, \mathbf{x}_t) - \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t)^\top \boldsymbol{a}_t$

$\boldsymbol{\Delta}\tilde{\mathbf{k}}_t = \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_{t-1}) - \gamma \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t)$

$d_t = \frac{\gamma \sigma_{t-1}^2}{s_{t-1}} d_{t-1} + r_{t-1} - \boldsymbol{\Delta}\tilde{\mathbf{k}}_t^\top \tilde{\boldsymbol{\alpha}}_{t-1}$

if $\delta_t > \nu$ **then**

Υπολογισμός $\tilde{\mathbf{K}}_t^{-1}$ (4.37)

$\boldsymbol{a}_t = (0, \ldots, 1)^\top$

$\tilde{\mathbf{h}}_t = (\boldsymbol{a}_{t-1}, -\gamma)^\top$

$\Delta k_{tt} = \tilde{\boldsymbol{a}}_{t-1}^\top \left(\tilde{\mathbf{k}}_{t-1}(\mathbf{x}_{t-1}) - 2\gamma \tilde{\mathbf{k}}_{t-1}(\mathbf{x}_t) \right) + \gamma^2 k_{tt}$

$\tilde{\mathbf{c}}_t = \frac{\gamma \sigma_{t-1}^2}{s_{t-1}} \begin{pmatrix} \tilde{\mathbf{c}}_{t-1} \\ 0 \end{pmatrix} + \tilde{\mathbf{h}}_t - \begin{pmatrix} \tilde{\mathbf{C}}_{t-1} \boldsymbol{\Delta}\tilde{\mathbf{k}}_t \\ 0 \end{pmatrix}$

$s_t = \sigma_{t-1}^2 + \gamma^2 \sigma_t^2 + \Delta k_{tt} - \boldsymbol{\Delta}\tilde{\mathbf{k}}_t^\top \tilde{\mathbf{C}}_{t-1} \boldsymbol{\Delta}\tilde{\mathbf{k}}_t + \frac{2\gamma \sigma_{t-1}^2}{s_{t-1}} \tilde{\mathbf{c}}_{t-1}^\top \boldsymbol{\Delta}\tilde{\mathbf{k}}_t - \frac{\gamma^2 \sigma_{t-1}^4}{s_{t-1}}$

$\tilde{\boldsymbol{\alpha}}_t = \begin{pmatrix} \tilde{\boldsymbol{\alpha}}_{t-1} \\ 0 \end{pmatrix}$

$\tilde{\mathbf{C}}_t = \begin{bmatrix} \tilde{\mathbf{C}}_{t-1} & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{bmatrix}$

else

$\tilde{\mathbf{h}}_t = \boldsymbol{a}_{t-1} - \gamma \boldsymbol{a}_t$

$\Delta k_{tt} = \tilde{\mathbf{h}}_t^\top \boldsymbol{\Delta}\tilde{\mathbf{k}}_t$

$\tilde{\mathbf{c}}_t = \frac{\gamma \sigma_{t-1}^2}{s_{t-1}} \tilde{\mathbf{c}}_{t-1} + \tilde{\mathbf{h}}_t - \tilde{\mathbf{C}}_{t-1} \boldsymbol{\Delta}\tilde{\mathbf{k}}_t$

$s_t = \sigma_{t-1}^2 + \gamma^2 \sigma_t^2 + \boldsymbol{\Delta}\tilde{\mathbf{k}}_t^\top \left(\tilde{\mathbf{c}}_t + \frac{\gamma \sigma_{t-1}^2}{s_{t-1}} \tilde{\mathbf{c}}_{t-1} \right) - \frac{\gamma^2 \sigma_{t-1}^4}{s_{t-1}}$

end if

$\tilde{\boldsymbol{\alpha}}_t = \tilde{\boldsymbol{\alpha}}_{t-1} + \frac{\tilde{\mathbf{c}}_t}{s_t} d_t$

$\tilde{\mathbf{C}}_t = \tilde{\mathbf{C}}_{t-1} + \frac{1}{s_t} \tilde{\mathbf{c}}_t \tilde{\mathbf{c}}_t^\top$

end for

return $\mathcal{D}_t, \tilde{\boldsymbol{\alpha}}_t, \tilde{\mathbf{C}}_t$

4.5 Βελτίωση Πολιτικής

Έως τώρα, έχουμε περιορίσει τη προσοχή μας, στο πρόβλημα της εκτίμησης πολιτικής. Σε αυτή την ενότητα, αξιοποιούμε τους προηγούμενους αλγορίθμους για τη κατασκευή αλγορίθμων που παράλληλα βελτιώνουν την πολιτική, αντί να την εκτιμούν. Αυτό μας επιτρέπει να επιλύουμε το πλήρες πρόβλημα της ενισχυτικής μάθησης, δηλαδή, το πρόβλημα εύρεσης μιας βέλτιστης πολιτικής. Στη συνέχεια προτείνονται δύο τύποι αλγορίθμων. Ο πρώτος βασίζεται στον αλγόριθμο επανάληψης ως προς τη πολιτική (Bertsekas and Tsitsiklis, 1996) [3], ενώ ο δεύτερος είναι βασισμένος στον αλγόριθμο SARSA (Sutton and Barto, 1998) [15]. Και στις δυο περιπτώσεις, χρησιμοποιείται ο GPTD (Gaussian Process Temporal Difference) [5], [6], [7] αλγόριθμος εκτίμησης πολιτικής. Προκειμένου να αποφευχθεί η μάθηση και η χρήση του μοντέλου μετάβασης p κατά τη βελτίωση της πολιτικής, θα πρέπει να κάνουμε μια απλή τροποποίηση στους GPTD αλγορίθμους έτσι ώστε να τους επιτρέπουν τη μάθηση των αξίων των ζευγών κατάστασης-ενέργειας.

Οι αλγόριθμοι που παρουσιαστήκαν μέχρι τώρα, σε αυτό το κεφάλαιο, βασίζονται στην ιδέα μιας ΜΔΑ που ελέγχεται από μια σταθερή πολιτική μ ως μια Μαρκοβιανή διαδικασία ανταμοιβής. Ο χώρος καταστάσεων αυτής της ΜΔΑ είναι ο $\mathcal{X}$ και η συνάρτηση πυκνότητας πιθανότητας μετάβασης είναι η $p^{\mu}(\mathbf{x}'|\mathbf{x}) = \int_{\mathcal{U}} \mu(\mathbf{u}|\mathbf{x})p(\mathbf{x}'|\mathbf{u},\mathbf{x})d\mathbf{u}$. Ωστόσο, μπορούμε να ορίσουμε μια άλλη ΜΔΑ $\mathcal{M}'$ εμπλουτίζοντας το χώρο καταστάσεων εισάγοντας και τις ενέργειες, δηλαδή $\mathcal{X}' = \mathcal{X} \times \mathcal{U}$. Έτσι η συνάρτηση πυκνότητας πιθανότητας μετάβασης είναι η $p'(\mathbf{x}', \mathbf{u}'|\mathbf{x}, \mathbf{u}) = p(\mathbf{x}'|\mathbf{x}, \mathbf{u})\mu(\mathbf{u}', \mathbf{x}')$, η συνάρτηση πυκνότητας της αρχικής κατάστασης είναι $p'_0(\mathbf{x}, \mathbf{u}) = p_0(\mathbf{x})\mu(\mathbf{u}|\mathbf{x})$ και η συνάρτηση πυκνότητας πιθανότητας της ανταμοιβής είναι $q'(r|\mathbf{x}, \mathbf{u}) = q(r|\mathbf{x})$. Εφαρμόζοντας ένα αλγόριθμο εκτίμησης της πολιτικής στη ΜΔΑ $\mathcal{M}'$, επιτυγχάνουμε την εκτίμηση των αξιών κατάστασης-ενέργειας $Q(\mathbf{x}, \mathbf{u})$. Το κύριο πλεονέκτημα του υπολογισμού των αξιών κατάστασης-ενέργειας $Q(\mathbf{x}, \mathbf{u})$ είναι πως αντί να εκτιμούμε την $\max_{\mathbf{u}} \mathbf{E}_{\mathbf{x}'|\mathbf{u},\mathbf{x}} \hat{\mathbf{V}}(\mathbf{x}')$ για τη βελτιώση της πολιτικής στη κατάσταση $\mathbf{x}$, αρκεί η εκτίμηση της $\max_{\mathbf{u}} \hat{Q}(\mathbf{x}, \mathbf{u})$ η οποία δεν απαιτεί τη γνώση του μοντέλου. Στα μη παραμετρικά μοντέλα, χρειάζεται να ορίσουμε μια συνάρτηση πυρήνα για τα ζεύγη κατάστασης-ενέργειας, $k : (\mathcal{X}\times\mathcal{U})\times(\mathcal{X}\times\mathcal{U}) \to \mathbb{R}$. Εφόσον οι καταστάσεις και οι ενέργειες είναι εντελώς διαφορετικές οντότητες, η συνάρτηση πυρήνα k προκύπτει από το γινόμενο μιας συνάρτησης πυρήνα k_x για την κατάσταση και μιας συνάρτησης πυρήνα k_u για την ενέργεια:

$$k(\mathbf{x}, \mathbf{u}, \mathbf{x}', \mathbf{u}') = k_x(\mathbf{x}, \mathbf{x}')k_u(\mathbf{u}, \mathbf{u}'). \tag{4.47}$$

Συνήθως, η συνάρτηση πυρήνα της κατάστασης, k_x, επιλέγεται να είναι Γκαουσιανή

$$k_x = k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma_x^2}\right)$$

Ένα σημαντικό ζήτημα στη Γκαουσιανή συνάρτηση πυρήνα είναι η εύρεση της κατάλληλης τιμής της παραμέτρου πλάτους σ_x. Στην παρούσα διατριβή προτείνεται μια περισσότερο αποτελεσματική συνάρτηση πυρήνα η οποία προσπαθεί να εξουδετερώσει το παραπάνω πρόβλημα. Συγκεκριμένα θεωρούμε τη συνάρτηση πυρήνα της κατάστασης ως ένα γραμμικό συνδυασμό N Γκαουσιανών συναρτήσεων πυρήνα με διαφορετικές τιμές πλάτους σ_i. Με το

τρόπο αυτό, λαμβάνουμε υπόψη όλο το εύρος τιμών των καταστάσεων και επιτυγχάνουμε τον ορθότερο υπολογισμό ομοιότητας μεταξύ δύο καταστάσεων. Η συνάρτηση κατάστασης πυρήνα παίρνει τη μορφή:

$$k_x = k(\mathbf{x}, \mathbf{x}') = \sum_{i=0}^{N} w_i \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma_i^2}\right), \quad \text{ισχύει ότι} \quad \sum_{i=0}^{N} w_i = 1.$$

Στα πειράματα χρησιμοποιούνται συνήθως $N = 10$ διαφορετικές τιμές πλάτους σ_i. Επειδή οι ενέργειες είναι συνήθως διακριτές τιμές δεν χρησιμοποιούνται Γκαουσιανές συναρτήσεις πυρήνα. Παρακάτω στη παρούσα ενότητα περιγράφεται μια συνάρτηση πυρήνα των ενεργειών.

Ο πρώτος αλγόριθμος είναι μια απλή προσαρμογή του αλγορίθμου επανάληψης ως προς την πολιτική (Ενότητα 3.2.1). Ο αλγόριθμος λειτουργεί διατηρώντας δύο GPTD εκτιμήσεις της πολιτικής, τη G_0 και τη G_1. Η συνάρτηση αξίας κατάστασης-ενέργειας που διατηρείται από την G_0 χρησιμοποιείται για να καθορίσει την πολιτική σύμφωνα με την οποία επιλέγονται οι ενέργειες, ενώ η G_1 χρησιμοποιείται για την εκτίμηση αυτής της πολιτικής. Εφόσον η G_1 εκτιμά με αρκετά ικανοποιητική ακρίβεια την συνάρτηση αξίας κατάστασης-ενέργειας, οι ρόλοι εναλλάσσονται. Τώρα η G_0 χρησιμοποιείται για εκτίμηση της πολιτικής που καθορίζεται από τις αξίες κατάστασης-ενέργειας της G_1. Οι πολιτικές που χρησιμοποιούνται σε αυτές τις επαναλήψεις θα πρέπει να προσεγγίζουν άπληστες πολιτικές σε σχέση με τις εκτιμήσεις αξίας, που διατηρούνται από τους αντίστοιχους εκτιμητές πολιτικής. Είναι συνεπώς χρήσιμο η συνάρτηση πυρήνα κατάστασης-ενέργειας, να επιλεχθεί με τέτοιο τρόπο, ώστε να επιτρέπει τον εύκολο υπολογισμο της $\max_{\mathbf{u}} \hat{Q}(\mathbf{x}, \mathbf{u})$. Εάν το σύνολο των ενεργειών $\mathcal{U}$ είναι πεπερασμένο, αυτό μπορεί να γίνει σε γραμμικό χρόνο ως προς το μέγεθος του $\mathcal{U}$. Ο ψευδοκώδικας του βασιζόμενου στη GPTD αλγορίθμου επανάληψης ως προς τη πολιτική (GPTD-API) [5], [6], [7] παρουσιάζεται στον Αλγόριθμο

Αλγόριθμος 12 Αλγόριθμος Επανάληψης ως προς τη Πολιτική Βασιζόμενος στη GPTD(GPTD-API)

Είσοδος: ΜΔΑ $\mathcal{M}$, κατώφλι σύγκλισης η
Αρχικοποίηση:
while not done **do**
  iter = iter + 1, i = iter mod 2, $j = 1 - i$
  $\mu(G_j, \epsilon)$ = ημι-άπληστη πολιτική σύμφωνα με τον G_j
  $G_i = \mathbf{GPTD}(\mathcal{M}, \mu(G_j, \epsilon))$
  done = $\|G_i - G_j\| \leq \eta$
end while
return G_i

Μια άλλη προσέγγιση εκτίμησης πολιτικής βασίζεται στον αλγόριθμο SARSA. Οι αλγόριθμοι αυτού του είδους αναφέρονται ως αισιόδοξη επανάληψη πολιτικής. Ο SARSA είναι μια σχετικά απλή επέκταση του αλγορίθμου ΧΔ στον οποίο εκτιμούνται οι αξίες κατάστασης-ενέργειας, ενώ την ίδια χρονική στιγμή οι ενέργειες επιλέγονται (semi-greedy)

βασιζόμενες στις τρέχουσες εκτιμήσεις των αξιών κατάστασης-ενέργειας. Ο λόγος που αναφέρεται ως αισιόδοξη επανάληψη πολιτικής, οφείλεται στο γεγονός ότι η πολιτική ενημερώνεται συνέχεια. Στη συνέχεια χρησιμοποιείτε ο GPTD ως εκτιμητής πολιτικής, καταλήγοντας στον αλγόριθμο GPSARSA [5], [6], [7]. Ο Αλγόριθμος 13, είναι ο ψευδοκώδικας του GPTD αλγορίθμου που βασίζεται στον μη παραμετρικό MC-GPTD αλγόριθμο (Αλγόριθμος 10).

Αλγόριθμος 13 Μη Παραμετρικός GPSARSA Αλγόριθμος

Αρχικοποίηση: $\boldsymbol{\alpha}_0 = \mathbf{0}, \mathbf{C}_0 = 0, \mathcal{D}_0 = \{\mathbf{x}_0, \mathbf{u}_0\}, \mathbf{c}_0 = \mathbf{0}, d_0 = 0, 1/s_0 = 0$

for $t = 1, 2, \ldots$ **do**

Παρατηρούμε $\mathbf{x}_{t-1}, \mathbf{u}_{t-1}, r_{t-1}, \mathbf{x}_t$

$\mathbf{u}_t = \textbf{SemiGreedyAction}(\mathbf{x}_t, \mathcal{D}_{t-1}, \boldsymbol{\alpha}_{t-1}, \mathbf{C}_{t-1})$

$\mathbf{h}_t = (0, \ldots, 1, -\gamma)^\top$

$\boldsymbol{\Delta}\mathbf{k}_t = \mathbf{k}_{t-1}(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}) - \gamma \mathbf{k}_{t-1}(\mathbf{x}_t, \mathbf{u}_t)$

$\Delta k_{tt} = k((\mathbf{x}_{t-1}, \mathbf{u}_{t-1}), (\mathbf{x}_{t-1}, \mathbf{u}_{t-1})) - 2\gamma k((\mathbf{x}_{t-1}, \mathbf{u}_{t-1}), (\mathbf{x}_t, \mathbf{u}_t)) + \gamma^2 k((\mathbf{x}_t, \mathbf{u}_t), (\mathbf{x}_t, \mathbf{u}_t))$

$\mathbf{c}_t = \frac{\gamma \sigma_{t-1}^2}{s_{t-1}} \begin{pmatrix} \mathbf{c}_{t-1} \\ 0 \end{pmatrix} + \mathbf{h}_t - \begin{pmatrix} \mathbf{C}_{t-1} \boldsymbol{\Delta}\mathbf{k}_t \\ 0 \end{pmatrix}$

$d_t = \frac{\gamma \sigma_{t-1}^2}{s_{t-1}} d_{t-1} + r_{t-1} - \boldsymbol{\Delta}\mathbf{k}_t^\top \boldsymbol{\alpha}_{t-1}$

$s_t = \sigma_{t-1}^2 + \gamma^2 \sigma_t^2 - \frac{\gamma^2 \sigma_{t-1}^4}{s_{t-1}} + \Delta k_{tt} - \boldsymbol{\Delta}\mathbf{k}_t^\top \mathbf{C}_{t-1} \boldsymbol{\Delta}\mathbf{k}_t + \frac{2\gamma \sigma_{t-1}^2}{s_{t-1}} \mathbf{c}_{t-1}^\top \boldsymbol{\Delta}\mathbf{k}_t$

$\boldsymbol{\alpha}_t = \begin{pmatrix} \boldsymbol{\alpha}_{t-1} \\ 0 \end{pmatrix} + \frac{\mathbf{c}_t}{s_t} d_t$

$\mathbf{C}_t = \begin{bmatrix} \mathbf{C}_{t-1} & \mathbf{0} \\ \mathbf{0}^\top & 0 \end{bmatrix} + \frac{1}{s_t} \mathbf{c}_t \mathbf{c}_t^\top$

$\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(\mathbf{x}_t, \mathbf{u}_t)\}$

end for

return $\boldsymbol{\alpha}_t, \mathbf{C}_t, \mathcal{D}_t$

Στον Αλγόριθμο 13, η συνάρτηση **SemiGreedyAction** χρησιμοποιείται ως ένα απροσδιόριστο μαύρο κουτί. Σε πολλούς semi-greedy κανόνες απαιτείται η λυσή της $\arg\max_{\mathbf{u}} \hat{Q}(\mathbf{x}_t, \mathbf{u})$. Ήδη προαναφέρθηκε, ότι ακόμη και όταν το $\mathcal{U}$ είναι συνεχές, μπορούμε να επιλύσουμε το συγκεκριμένο πρόβλημα μεγιστοποίησης σε κλειστή μορφή, με κατάλληλη επιλογή του τρόπου επιλογής των ενεργειών και της συνάρτησης πυρήνα. Για να το αποδείξουμε αυτό, εξετάζουμε μια εργασία πλοήγησης, που αντιμετωπίζεται από ενα ρομπότ. Για απλότητα, υποθέτουμε ότι χώρος καταστάσεων $\mathcal{X}$ είναι ένα υποσύνολο ενός n-διάστατου Ευκλίδειου χώρου (το n συνήθως είναι 2 ή 3) και οι ενέργειες είναι βήματα μήκους 1 προς οποιαδήποτε κατεύθυνση. Ως εκ τούτου, ο χώρος ενεργειών $\mathcal{U}$ είναι μια n-διάστατη μοναδιαία σφαίρα. Στις πραγματικές μεταβάσεις καταστάσεων εμπεριέχεται θόρυβος λόγω των περιορισμών που επιβάλλονται από το περιβάλλον, όπως τοίχους, εμπόδια, κλπ. Επιλέγουμε να αναπαραστήσουμε μια ενέργεια από το αντίστοιχο μοναδιαίο διάνυσμα της, $\mathbf{u}$. Αφήνουμε τον πυρήνα κατάστασης k_x απροσδιόριστο και εστιάζουμε στον πυρήνα ενέργειας. Ορίζουμε

τον k_u ως,

$$k_u(\mathbf{u}, \mathbf{u}') = \frac{1-b}{2}\mathbf{u}^\top\mathbf{u}' + \frac{1+b}{2},$$

όπου b είναι μια σταθερά στο $[0, 1]$. Εφόσον το $\mathbf{u}^\top\mathbf{u}'$ είναι το συνημίτονο της γωνίας ανάμεσα στο $\mathbf{u}$ και το $\mathbf{u}'$. Το $k_u(\mathbf{u}, \mathbf{u}')$ επιτυγχάνει τη μέγιστη τιμή(1) όταν οι δύο ενέργειες είναι ίδιες και την ελάχιστη τιμή (b) όταν οι δύο ενέργειες οδηγούν σε αντίθετες κατευθύνσεις. Το πολυτιμότερο χαρακτηριστικό αυτού του πυρήνα είναι η γραμμικότητά του, που καθιστά δυνατή τη μεγιστοποίηση της εκτιμώμενης αξίας για όλες τις ενέργειες.

Υποθέτουμε ότι ο πράκτορας εκτελεί τον αραιό μη παραμετρικό GPSARSA αλγόριθμο, με αποτέλεσμα να διατηρεί ένα λεξικό των ζευγών κατάστασης-ενέργειας $\mathcal{D}_t = \{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{u}}_i)\}_{i=1}^m$. Η εκτίμηση της αξίας για την τρέχουσα κατάσταση $\mathbf{x}$ και μια αυθαίρετη ενέργεια $\mathbf{u}$ είναι

$$\begin{aligned}\hat{V}(\mathbf{x}, \mathbf{u}) &= \sum_{i=1}^m \tilde{\boldsymbol{\alpha}}_i k_x(\tilde{\mathbf{x}}_i, \mathbf{x}) k_u(\tilde{\mathbf{u}}_i, \mathbf{u}) \\ &= \sum_{i=1}^m \tilde{\boldsymbol{\alpha}}_i k_x(\tilde{\mathbf{x}}_i, \mathbf{x}) \frac{1-b}{2}\mathbf{u}^\top\mathbf{u}' + \frac{1+b}{2}\end{aligned}$$

Η μεγιστοποίηση αυτής έκφρασης σε σχέση με το $\mathbf{u}$ οδηγεί στη μεγιστοποίηση της $\sum_{i=0}^m \beta_i(\mathbf{x})\tilde{\mathbf{u}}_i^\top\mathbf{u}$ υπό τον περιορισμό $\|\mathbf{u}\| = 1$, όπου $\beta_i(\mathbf{x}) \stackrel{\text{def}}{=} \tilde{\boldsymbol{\alpha}}_i k_x(\tilde{\mathbf{x}}_i, \mathbf{x})$. Επιλύοντας το συγκεκριμένο πρόβλημα χρησιμοποιώντας τη μέθοδο Lagrance, παίρνουμε την άπληστη ενέργεια $\mathbf{u}^* = \frac{1}{\lambda}\sum_{i=1}^m \beta_i(\mathbf{x})\tilde{\mathbf{u}}_i$, όπου λ είναι μια σταθερά κανονικοποίησης.

4.6 Επεκτάσεις

4.6.1 1η Επέκταση

Η πρώτη επέκταση που προτείνεται στη συγκεκριμένη διατριβή βασίζεται στη χρήση των RVM (Tipping, 2001) [17]στη παραπάνω Μπεϋσιανή προσέγγιση εκτίμησης πολιτικής για τη δημιουργία αραιότερων μοντέλων εκτίμησης της συνάρτησης αξίας. Όπως προαναφέρθηκε οι προβλέψεις μας βασίζονται στη συνάρτηση αξίας κατάστασης V. "Μάθηση" είναι η διαδικασία εξαγωγής, αυτής της συνάρτησης. Ένα ευέλικτο και δημοφιλές σύνολο υποψηφίων μοντέλων για την V, είναι αυτό της μορφής:

$$V(x) = \sum_{i=1}^M \mathbf{w}^\top\boldsymbol{\phi}(x) \tag{4.48}$$

όπου η έξοδος είναι ένα γραμμικά σταθμισμένο άθροισμα M, γενικά μη γραμμικών συναρτήσεων βάσης $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))^\top$. Στη περίπτωση μας, οι συναρτήσεις βάσης ορίζονται από πυρήνες, με κάθε πυρήνα να σχετίζεται με μια κατάσταση. Στη συνέχεια υιοθετούμε ένα πλήρως πιθανοτικό μοντέλο και εισάγουμε μια εκ των προτέρων κατανομή στά βάρη του μοντέλου που διέπεται από ένα σύνολο υπερπαραμέτρων. Κάθε υπερπαράμετρος συνδέεται με κάποιο βάρος, οι τιμές των οποίων υπολογίζονται επαναληπτικά από τα δεδομένα. Η σποραδικότητα επιτυγχάνεται διότι οι εκ των υστέρων κατανομές

αρκετών βαρών πηγαίνουν απότομα κοντά στο μηδέν. Όπως έχει προαναφερθεί (Ενότητα 4.4.1) η ανταμοιβή μιας συνάρτησης δίνεται ως εξής:

$$R_{t-1} = \mathbf{H}_t V_t + N_t, \tag{4.49}$$

όπου $N_t \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Έτσι η πιθανοφάνεια του συνόλου των ανταμοιβών εκφράζεται ως

$$p(R_{t-1}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-t/2} \exp\left\{-\frac{1}{2\sigma^2}\|R_{t-1} - \mathbf{H}_t\mathbf{\Phi}\mathbf{w}\|^2\right\}, \tag{4.50}$$

όπου $R_{t-1} = (R(\mathbf{x}_0), \cdots, R(\mathbf{x}_{t-1}))^\top$, $\mathbf{w} = (\mathbf{w}_0, \cdots, \mathbf{w}_{t-1})^\top$ και $\mathbf{\Phi}$ είναι ένας $t \times t$ πίνακας σχεδίασης με $\mathbf{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_0), \cdots, \boldsymbol{\phi}(\mathbf{x}_{t-1})]^\top$, όπου $\boldsymbol{\phi}(\mathbf{x}_n) = [k(\mathbf{x}_n, \mathbf{x}_0), \cdots, k(\mathbf{x}_n, \mathbf{x}_{t-1})]^\top$. Έπειτα εισάγουμε μια ξεχωριστή υπερπαράμετρο a_i για κάθε βάρος w_i αντί για μια κοινή υπερπαράμετρο. Έτσι η εκ των προτέρων κατανομή των βαρών έχει τη μορφή

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{t-1} \mathcal{N}(w_i|0, \alpha_i^{-1}), \tag{4.51}$$

όπου το α_i αναπαριστά την ακρίβεια της αντίστοιχης παραμέτρου w_i και $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{t-1})^\top$. Στη συνέχεια ορίζουμε τις υπερπαραμέτρους $\boldsymbol{\alpha}$ και τη διακύμανση του θορύβου σ^2, ως Γάμμα κατανομές:

$$p(\boldsymbol{\alpha}) = \prod_{i=0}^{N} \text{Gamma}(\alpha_i|a, b),$$
$$p(\beta) = \text{Gamma}(\beta|c, d),$$

όπου $\beta \equiv \sigma^{-2}$. Για να γίνουν αυτές οι εκ των προτέρων κατανομές μη ενημερωτικές(non-informative), θα πρέπει να ορίσουμε τις παραμέτρους τους με μικρές τιμες: π.χ $a = b = c = d = 10^{-4}$.

Η εκ των υστέρων κατανομή των βαρών είναι Γκαουσιανή και έχει τη μορφή

$$p(\mathbf{w}|R_{t-1}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \mathbf{\Sigma}) \tag{4.52}$$

όπου η εκ των υστέρων κατανομή και το μέσο είναι αντίστοιχα:

$$\mathbf{\Sigma} = (\sigma^{-2}\mathbf{\Phi}^\top\mathbf{H}_t^\top\mathbf{H}_t\mathbf{\Phi} + \mathbf{A})^{-1}, \tag{4.53}$$
$$\boldsymbol{\mu} = \sigma^{-2}\mathbf{\Sigma}\mathbf{\Phi}^\top\mathbf{H}_t^\top R_{t-1}, \tag{4.54}$$

με $A = diag(\alpha_0, \alpha_1, \ldots, \alpha_N)$.

Στη συνέχεια βρίσκουμε τις υπερπαραμέτρους α και β που μεγιστοποιούν την πιθανοφάνεια. Η περιθώρια πιθανοφάνεια λαμβάνεται ολοκληρώνοντας ως προς τις παραμέτρους των βαρών:

$$\begin{aligned} p(R_{t-1}|\boldsymbol{\alpha}, \sigma^2) &= \int p(R_{t-1}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\alpha)d\mathbf{w} \\ &= (2\pi)^{N/2}|\sigma^2\mathbf{I} + \mathbf{H}_t\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^\top\mathbf{H}_t^\top|^{-1/2} \exp\left\{-\frac{1}{2}R_{t-1}^\top(\sigma^2\mathbf{I} + \mathbf{H}_t\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^\top\mathbf{H}_t^\top)^{-1}R_{t-1}\right\}. \end{aligned} \tag{4.55}$$

Στόχος μας είναι η μεγιστοποίηση του λογαρίθμου της 4.55 ως προς τις υπερπαραμέτρους α και β. Θέτοντας απλά τις παραγώγους της περιθώριας κατανομής, ως προς το α και β ίσες με το μηδέν, παίρνουμε τις παρακάτω εξισώσεις επανεκτίμησης :

$$\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2} \tag{4.56}$$

$$\beta_i^{new} = \frac{\|R_{t-1} - \mathbf{H}_t \mathbf{\Phi} \boldsymbol{\mu}\|^2}{N - \sum_i \gamma_i} \tag{4.57}$$

Η ποσότητα γ_i εκτιμά πόσο καλά ορίζεται η παράμετρος w_i από τα δεδομένα και ορίζεται ως εξής:

$$\gamma_i \equiv 1 - \alpha_i \Sigma_{ii} \tag{4.58}$$

Οι τιμές των υπερπαραμέτρων υπολογίζονται κάθε φορά που μια καινούρια κατάσταση εισέρχεται στο λεξικό. Έχοντας βρεί τις τιμές α_{MP} και β_{MP} για τις υπερπαραμέτρους που μεγιστοποιούν την περιθώρια πιθανοφάνεια, μπορούμε να εκτιμήσουμε την κατανομή πρόβλεψης για ένα καινούριο δεδομένο $\mathbf{x}$. Αυτό γίνεται ως εξής

$$\begin{aligned} p(t_*|R_{t-1}, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP}) &= \int p(t_*|\mathbf{w}, \sigma^2_{MP}) p(\mathbf{w}|R_{t-1}, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP}) d\mathbf{w} \\ &= \mathcal{N}(t_*|y_*, \sigma^2_*), \end{aligned} \tag{4.59}$$

όπου

$$y_* = \boldsymbol{\mu}^\top \boldsymbol{\phi}(\mathbf{x}_*), \tag{4.60}$$

$$\sigma_0^2 = \sigma^2_{MP} + \boldsymbol{\phi}(\mathbf{x}_*)^\top \mathbf{\Sigma} \boldsymbol{\phi}(\mathbf{x}_*). \tag{4.61}$$

Παρατηρούμε ότι ο πίνακας συνδιακύμανσης δίνεται ως εξής:

$$\mathbf{K} = \mathbf{\Phi} \mathbf{A}^{-1} \mathbf{\Phi}^\top \quad \text{ή} \quad \mathbf{K}_{pq} = \sum_{j=1}^{M} \frac{1}{a_j} \phi_j(\mathbf{x}_p) \phi_j(\mathbf{x}_q) \tag{4.62}$$

Το διάνυσμα των διακυμάνσεων ανάμεσα στη νέα πρόβλεψη και στα δεδομένα εκπαίδευσης δίνεται ως:

$$\mathbf{k}(\mathbf{x}^*) = \mathbf{\Phi} \mathbf{A}^{-1} \boldsymbol{\phi}^* \quad \text{ή} \quad [\mathbf{k}(\mathbf{x}^*)]_p = \sum_{j=1}^{M} \frac{1}{a_j} \phi_j(\mathbf{x}_p) \phi_j(\mathbf{x}^*) \tag{4.63}$$

όπου ορίζουμε ότι $\boldsymbol{\phi}^* = \boldsymbol{\phi}(\mathbf{x}^*)$. Τέλος, η διακύμανση της νέας πρόβλεψης δίνεται ως $k = {\boldsymbol{\phi}^*}^\top \mathbf{A}^{-1} \boldsymbol{\phi}^*$.

Έτσι αντικαθιστώντας τις 4.62 και 4.63 στην Εξίσωση 4.41 δημιουργούμε μια αραιή Μπεϋσιανή προσέγγιση για την εκτίμηση της συνάρτησης αξίας με τη χρήση RVM.

4.6.2 2η Επέκταση

Στη συνέχεια περιγράφεται η δεύτερη προσέγγιση που προτείνεται στη παρούσα διατριβή η οποία τροποποιεί το πρόβλημα ελαχιστοποίησεις που περιγράφει η εξίσωση 4.32. Η συγκεκριμένη προσέγγιση προσθέτουμε ένα περιορισμό σε σχέση με τα βάρη στην εξίσωση 4.32

(μέθοδος Lasso [8]). Έτσι η εξίσωση θα έχει την εξής μορφή:

$$\delta_t \stackrel{\text{def}}{=} \min_{\boldsymbol{a}} \left\| \sum_{j=1}^{m_{t-1}} a_j \boldsymbol{\phi}(\tilde{\mathbf{x}}_j) - \boldsymbol{\phi}(\mathbf{x}_t) \right\|^2 + \lambda \sum_{j=1}^{m_{t-1}} |a_j|. \tag{4.64}$$

Με τον τρόπο αυτό οι περισσότεροι από τους συντελεστές a_i θα τείνουν στο μηδέν. Αυτό έχει ως αποτέλεσμα να λαμβάνουμε αραιότερες λύσεις αφού οι εικόνες των καταστάσεων του λεξικού που συνεισφέρουν ώστε να προσεγγίσουμε την εικόνα μιας νέας κατάστασεις θα είναι ελάχιστες.

Κεφαλαιο 5

ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ

5.1 Εισαγωγή

Στο κεφάλαιο αυτό αξιολογούμε κάποιες από τις κυριότερες μεθόδους επίλυσης προβλημάτων ενισχυτικής μάθησης που περιγράψαμε στην παρούσα διατριβή. Αυτές οι μέθοδοι εφαρμόζονται και αξιολογούνται σε πειραματικά περιβάλλοντα αυτόνομης πλοήγησης ρομποτικών συστημάτων. Συγκεκριμένα, οι μέθοδοι που αξιολογούνται είναι οι SARSA, SARSA(λ), GPTD. Αρχικά, υλοποιούμε και συγκρίνουμε την απόδοση των παραπάνω μεθόδων για τα προβλήματα ελέγχου Mountain Car και Cart Pole(ανεστραμμένο εκρεμμές), αντίστοιχα. Τέλος, η σημαντικότερη καινοτομία της συγκεκριμένης διατριβής είναι η εφαρμογή και η υλοποίηση των προηγούμενων μεθόδων στο πρόβλημα πλοήγησης ενός πραγματικού αυτόνομου ρομποτικού συστήματος. Οι αξιολόγηση των μεθόδων βασίζεται στο μέσο όρο βημάτων που χρειάζεται για να φθάσουν σε ένα συγκεκριμένο στόχο και στο μέσο όρο των απολαβών που λαμβάνει ο πράκτορας, σε κάθε επεισόδιο. Έτσι έχουμε τη δυνατότητα να εκτιμήσουμε το ρυθμό σύγκλισης κάθε μεθόδου ενώ ταυτόχρονα μπορούμε να συγκρίνουμε τις πολιτικές που παράγονται από τις συγκεκριμένες μεθόδους.

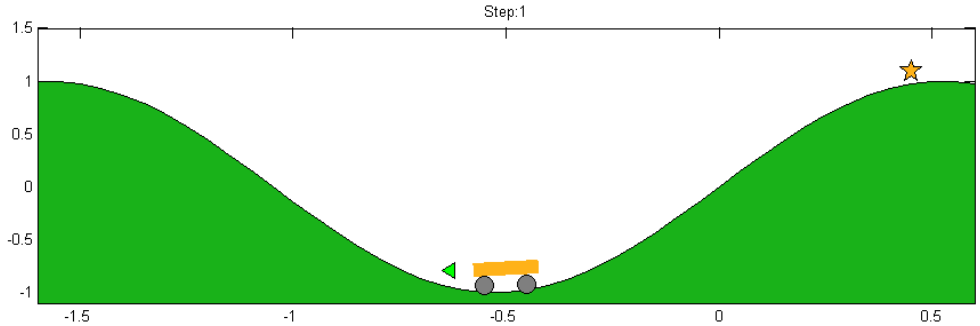
5.2 Πειραματικά Περιβάλλοντα

5.2.1 Mountain Car

Στην ενότητα αυτή μελετάμε ένα από τα δημοφιλέστερα προβλήματα ελέγχου, το Mountain Car. Στο συγκεκριμένο πρόβλημα, προσπαθούμε να βρούμε τον τρόπο με τον οποίο ένα αυτοκίνητο μπορεί να ανέβει ένα λόφο, ξεκινώντας από μια συγκεκριμένη θέση (Σχημα 5.1). Το αυτοκίνητο μπορεί να επιλέξει ανάμεσα από τρείς ενέργειες, $\boldsymbol{a}$. Είτε να επιταχύνει προς τα εμπρός ($\boldsymbol{a} = 1$) είτε να επιταχύνει προς τα πίσω ($\boldsymbol{a} = -1$) είτε να μην επιταχύνει προς καμία από τις δύο κατευθύνσεις ($\boldsymbol{a} = 0$). Κάθε χρονική στιγμή το αυτοκίνητο βρίσκεται σε μια συγκεκριμένη θέση, έχοντας μια συγκεκριμένη ταχύτητα. Έτσι, οι καταστάσεις του περιβάλλοντος ορίζονται ως διανύσματα δύο διαστάσεων, $\mathbf{x} = (x_1, x_2)$. Η συνιστώσα x_1 προσδιορίζει τη θέση, ενώ η x_2 την ταχύτητα του αυτοκινήτου. Επιλέγοντας και εκτελώντας μια ενέργεια το αυτοκίνητο μεταβαίνει από τη κατάσταση $\mathbf{x}$ στην οποία βρίσκεται εκείνη τη στιγμή, σε μια νέα κατάσταση $\mathbf{x}'$. Η κατάσταση στην οποία μεταβαίνουμε εξαρτάται τόσο από την ενέργεια όσο και από τη κατάσταση στην οποία βρισκόμαστε. Στα πειράματα μας, θεωρούμε την παρακάτω εξίσωση κίνησης:

$$x_2' = 0.999(x_2 + 0.001a + (-0.0025\cos(3.0x_1)))$$
$$x_1' = x_1 + x_2'$$

Η θέση του αυτοκινήτου παίρνει τιμές στο διάστημα $[-1.5, 0.5]$, ενώ η ταχύτητα στο διάστημα $[-0.07, 0.07]$. Στόχος του πράκτορα είναι να φθάσει στη θέση 0.45 (αστεράκι στο σχήμα) με όσο το δυνατόν λιγότερα βήματα, ξεκινώντας από τη θέση -0.5 έχοντας μηδενική ταχύτητα. Η ανταμοιβή που λαμβάνει ο πράκτορας μετά από κάθε μετάβαση σε μια νέα κατάσταση είναι -1, εκτός από τη περίπτωση που φθάνει στο στόχο, όπου παίρνει 100. Η συγκεκριμένη εργασία εκτελείται σε επεισόδια. Έτσι, αν ο πράκτορας δεν φθάσει στο στόχο σε λιγότερο από 1000 βήματα, ξεκινά ένα καινούριο επεισόδιο τοποθετώντας το αυτοκίνητο στην αρχική κατάσταση.



Σχήμα 5.1: Πειραματικό Περιβάλλον του Mountain Car

Αρχικά, υλοποιήσαμε και εφαρμόσαμε στο συγκεκριμένο πρόβλημα τον αλγόριθμο ΧΔ, Sarsa. Διαπιστώσαμε πως οι βέλτιστες τιμές των παραμέτρων του αλγορίθμου Sarsa για

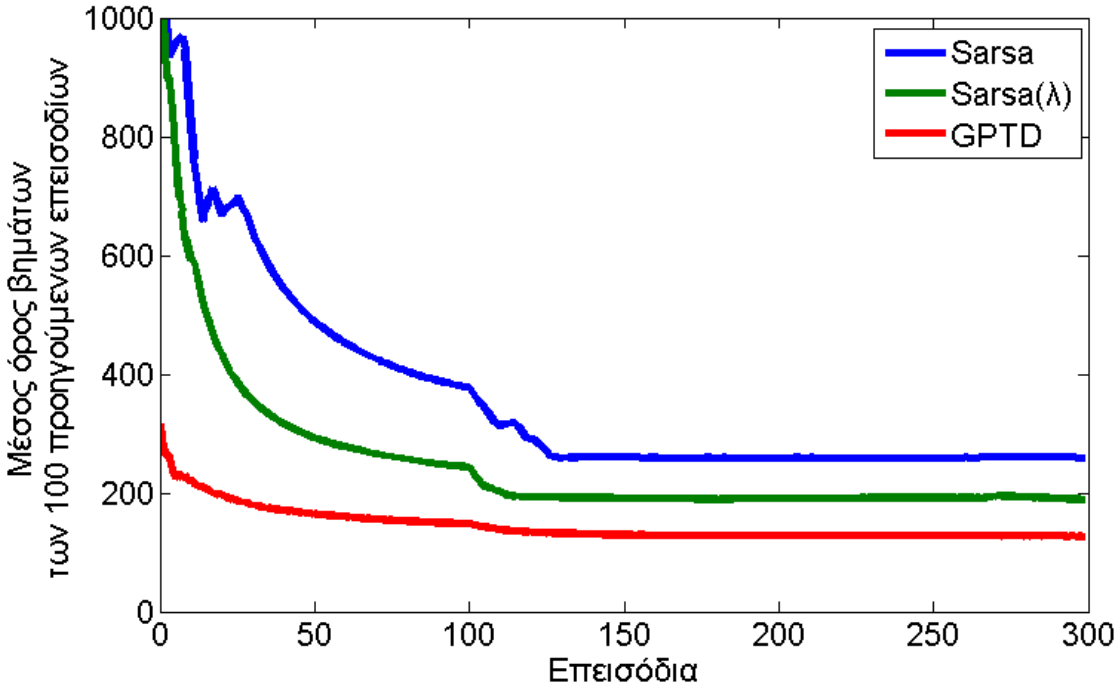
το συγκεκριμένο πρόβλημα είναι: ρυθμός μάθησης $\alpha = 0.5$, ρυθμός έκπτωσης $\gamma = 1$ και πιθανότητα επιλογής τυχαίας ενέργειας $\epsilon = 0.01$. Στη συνέχεια, εφαρμόσαμε στο παρών πρόβλημα τον αλγόριθμο ΧΔ με ίχνη επιλεξιμότητας, Sarsa(λ). Στο συγκεκριμένο πείραμα βρήκαμε ότι οι βέλτιστες τιμές των παραμέτρων του αλγορίθμου Sarsa(λ) είναί: ρυθμός μάθησης $\alpha = 0.5$, ρυθμός έκπτωσης $\gamma = 1$, παράμετρος μείωσης του ίχνους $\lambda = 0.95$ και τη πιθανότητα επιλογής τυχαίας ενέργειας $\epsilon = 0.01$. Τέλος, εφαρμόσαμε τον αλγόριθμο GPTD στο πρόβλημα ελέγχου του Mountain Car. Ανακαλύψαμε ότι οι βέλτιστες τιμές των παραμέτρων του αλγορίθμου GPTD για το πρόβλημα του MountainCar είναι οι εξής: ρυθμός έκπτωσης $\gamma = 0.999$, διακύμανση του θορύβου $\sigma^2 = 1$ και πιθανότητα επιλογής τυχαίας ενέργειας $\epsilon = 0.01$. Ως συνάρτηση πυρήνα κατάστασης, k_x, χρησιμοποιούμε μια Γκαουσιάνη συνάρτηση πυρήνα. Επειδή όμως οι συνιστώσες του διανύσματος κατάστασης x αντιπροσωπεύουν διαφορετικές έννοιες, τροποποιούμε την Γκαουσιανή συνάρτηση πυρήνα, παίρνοντας την

$$k_x(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{(x_1 - x_1')^2}{2\sigma_1^2} - \frac{(x_2 - x_2')^2}{2\sigma_2^2}\right),$$

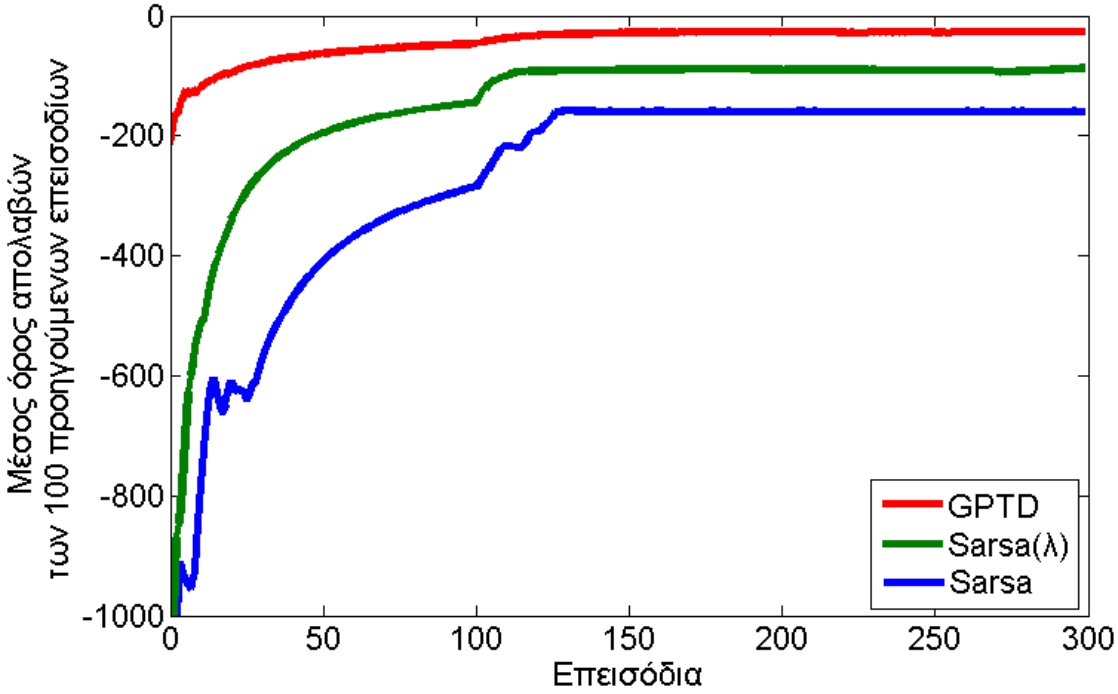
όπου σ_1^2 είναι η διακύμανση της θέσης και σ_2^2 είναι η διακύμανση της ταχύτητας του αυτοκινήτου. Οι καταστάσεις του περιβάλλοντος για να είναι όμοιες θα πρέπει τόσο η μεταξύ τους απόσταση να μην είναι μεγάλη ($|x_1 - x_1'| < 0.2$) όσο και η ταχύτητα που έχει το αυτοκίνητο στις συγκεκριμένες καταστάσεις να διαφέρει ελάχιστα ($|x_2 - x_2'| < 0.02$). Η εύρεση των βέλτιστων διακυμάνσεων για αυτό το πρόβλημα, έγινε με μια πειραματική προσέγγιση. Έτσι θεωρήσαμε τη διακύμανση της θέσης του αυτοκινήτου $\sigma_1^2 = 0.05$ και τη διακύμανση της ταχύτητας $\sigma_2^2 = 0.0005$. Τέλος, εξαιτίας του γεγονότος ότι οι ενέργειες είναι διακριτές, θεωρούμε τη παρακάτω συνάρτηση πυρήνα των ενεργειών:

$$k_a(\boldsymbol{a}, \boldsymbol{a}') = \begin{cases} 0 & \text{εάν } |\boldsymbol{a} + \boldsymbol{a}'| = 0 \\ 0.1 & \text{εαν } |\boldsymbol{a} + \boldsymbol{a}'| = 1 \\ 1 & \text{εαν } |\boldsymbol{a} + \boldsymbol{a}'| = 2 \end{cases}$$

Συγκριτικά αποτελέσματα των παραπάνω αλγορίθμων για το πρόβλημα του Mountain Car, παρουσιάζονται στα Σχήματα 5.2(a) και 5.2(b). Στο Σχήμα 5.2(a), βλέπουμε το μέσο όρο βημάτων που απαιτούνται για την επίτευξη του στόχου, εφαρμόζοντας τις μεθόδους GPTD, SARSA και SARSA(λ) στο πρόβλημα. Γίνεται εύκολα αντιληπτό ότι ο αλγόριθμος GPTD συγκλίνει γρηγορότερα σε σχέση με τους δύο άλλους αλγορίθμους. Ταυτόχρονα, ο αλγόριθμος GPTD βρίσκει μια καλύτερη πολιτική σε σχέση με τους SARSA και SARSA(λ), με αποτέλεσμα να χρειάζεται λιγότερα βήματα για να φθάσει στο στόχο. Στο Σχήμα 5.2(b), βλέπουμε το μέσο όρο των απολαβών που λαμβάνει ο πράκτορας εφαρμόζοντας τις μεθόδους GPTD, SARSA και SARSA(λ)στο πρόβλημα. Μπορούμε εύκολα να παρατηρήσουμε ότι ο αλγόριθμος GPTD, οδηγεί σε αρκετά καλύτερες ανταμοιβές συγκριτικά με τους δύο άλλους αλγορίθμους.



(a) Μέσος όρος βημάτων που απαιτούνται για επίτευξη του στόχου, εφαρμόζοντας τις μεθόδους GPTD, SARSA και SARSA(λ) στο πρόβλημα Mountain Car

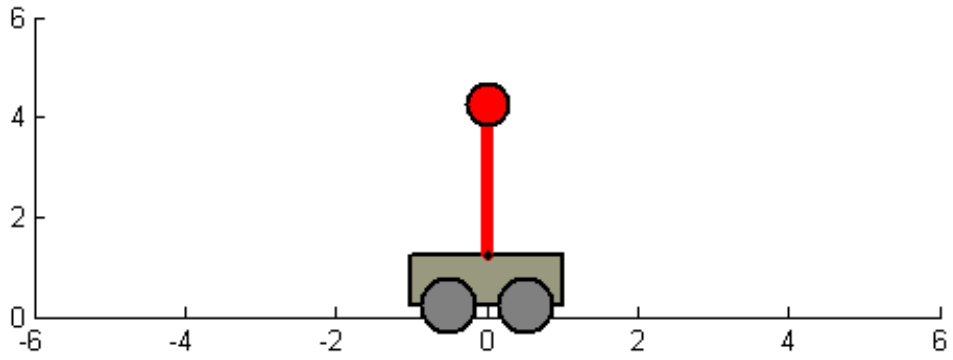


(b) Μέσος όρος απολαβών που λαμβάνει ο πράκτορας, εφαρμόζοντας τις μεθόδους GPTD, SARSA και SARSA(λ)στο πρόβλημα Mountain Car

Σχήμα 5.2: Σύγκριση των μεθόδων GPTD, SARSA και SARSA(λ) στο πρόβλημα Mountain Car

5.2.2 Το Πρόβλημα του Ανεστραμμένου Εκκρεμούς (Cart Pole Problem)

Στη συγκεκριμένη ενότητα μελετούμε το πρόβλημα εξισορρόπησης της άμαξας και του κονταριού (Cart-pole), που είναι γνωστό ως ανεστραμμένο εκκρεμές (inverted pendulum) (Εικόνα 5.3). Το πρόβλημα είναι ο έλεγχος της θέσης $\mathbf{x}$ της άμαξας έτσι ώστε το κοντάρι να είναι σχεδόν κατακόρυφο ενώ παράλληλα θα βρίσκεται μέσα στα όρια της διαδρομής της άμαξας, σύμφωνα με την εικόνα. Η άμαξα έχει τη δυνατότητα να κινηθεί στο διάστημα $[-4, 4]$.



Σχήμα 5.3: Ανεστραμμένο Εκκρεμές

Στο συγκεκριμένο πρόβλημα οι ενέργειες $\boldsymbol{a}$ είναι διακριτές: σπρωξιά προς τα αριστερά ή σπρωξιά προς τα δεξιά, η ονομαζόμενη συνταγή ελέγχου με κοφτά χτυπήματα (bang-bang control). Συγκεκριμένα, έχουμε στη διάθεσή μας 10 διακριτές ενέργειες $\boldsymbol{a}$ προς κάθε κατεύθυνση. Έτσι μπορούμε να επιλέξουμε ανάμεσα από συνολικά 21 ενέργειες, όπου $\boldsymbol{a} = -10, -9, \ldots, 0, \ldots, 9, 10$. Οι ενέργειες αυτές διαφέρουν μεταξύ λόγο τις διαφορετικής δύναμης που ασκείται σε κάθε χτύπημα προς την ίδια κατεύθυνση. Κάθε χρονική στιγμή η άμαξα βρίσκεται σε μια συγκεκριμένη θέση έχοντας μια συγκεκριμένη ταχύτητα. Την ιδία στιγμή, το κοντάρι σχηματίζει μια γωνία θ, με μια νοητή ευθεία κάθετη στη άμαξα, έχοντας μια συγκεκριμένη ταχύτητα κλίσης. Έτσι, παρατηρούμε ότι, οι μεταβλητές κατάστασης είναι συνέχεις και οι καταστάσεις ορίζονται ως διανύσματα τεσσάρων μεταβλητών $\mathbf{x} = (x_1, x_2, x_3, x_4)$. Η συνιστώσα x_1 του διανύσματος προσδιορίζει τη θέση ενώ η x_2 την ταχύτητα της άμαξας. Η θέση της άμαξας παίρνει τιμές στο διάστημα $[-4, 4]$ και η ταχύτητα στο διάστημα $[-1, 1]$. Επίσης, η συνιστώσα x_3 προσδιορίζει την κλίση του κονταριού και η x_4 ορίζει την ταχύτητα με την οποία μεταβάλλεται η συγκεκριμένη κλίση. Οι δύο παραπάνω συνιστώσες παίρνουν τιμές στα διαστήματα, $[-45°, 45°]$ και $[-0.1745, 0.1745]$, αντίστοιχα. Η επιλογή και εκτέλεση μιας ενέργειας αναγκάζουν τον πράκτορα να μεταβεί σε μια καινούρια κατάσταση $\mathbf{x}' = (x'_1, x'_2, x'_3, x'_4)$ κάθε φορά. Η κατάσταση $\mathbf{x}'$ στην οποία μεταβαίνουμε, εξαρτάται τόσο από τη κατάσταση $\mathbf{x}$ που βρισκόμαστε εκείνη τη στιγμή όσο και από την ενέργεια που επιλέξαμε. Η εξίσωση κίνησης που χρησιμοποιείται για να προσομοιώσει το

συγκεκριμένο περιβάλλον δίνεται ως εξής:

$$\begin{aligned} x'_1 &= x_1 + T \cdot x_2 \\ x'_2 &= x_2 + T \cdot x_{\text{acc}} \\ x'_3 &= x_3 + T \cdot x_4 \\ x'_4 &= x_4 + T \cdot theta_{\text{acc}}, \end{aligned}$$

όπου,

$$\begin{aligned} temp &= (force + PoleMass_{\text{Length}} \cdot x_4^2 \cdot \sin(x_3))/Total_{\text{Mass}}, \\ theta_{\text{acc}} &= (g \cdot \sin(x_3) - \cos(x_3) \cdot temp)/(Length \cdot ((4.0/3.0) - ((Mass_{\text{Pole}} \cdot \cos^2(x_3))/Total_{\text{Mass}}))), \\ x_{\text{acc}} &= temp - (PoleMass_{\text{Length}} \cdot theta_{\text{acc}} \cdot \cos(x_3))/Total_{\text{Mass}} \end{aligned}$$

και

$$\begin{aligned} g &= 9.8 \text{ (Βαρύτητα)} , \\ Mass_{\text{Cart}} &= 1.0 \text{ (Βάρος της άμαξας σε κιλά)} , \\ Mass_{\text{Pole}} &= 0.1 \text{ (Βάρος του κονταριού σε κιλά)} , \\ Total_{\text{Mass}} &= Mass_{\text{Cart}} + Mass_{\text{Pole}} \text{ (Συνολικό Βάρος)} , \\ Length &= 0.5 \text{ (Το μισό μήκος του κονταριού)} , \\ PoleMass_{\text{Length}} &= Mass_{\text{Pole}}, \cdot Length \\ Force_{\text{Mag}} &= 10.0 \text{ (Το μέγεθος της δύναμης)} , \\ T &= 0.02 \text{ (Χρονικό διάστημα ανάμεσα σε δυο ενημερώσεις)} , \\ force &= \boldsymbol{a} \cdot Force_{\text{Mag}} \end{aligned}$$

Οι ανταμοιβές που λαμβάνει ο πράκτορας σε κάθε βήμα, εξαρτώνται από το πόσο μακριά βρίσκεται η άμαξα από τη θέση μηδέν καθώς επίσης και από την κλίση του κονταριού. Κατ'αυτό το τρόπο, ο πράκτορας λαμβάνει καλύτερες ανταμοιβές όταν η άμαξα βρίσκεται πολύ κοντά στη θέση μηδέν και την ίδια στιγμή το κοντάρι είναι κάθετα στο επίπεδο της άμαξας. Όταν το κοντάρι δεν βρίσκεται μέσα στα όρια της διαδρομής της άμαξας ή όταν η κλίση του είναι μεγαλύτερη από 45 μοίρες από την κατακόρυφο, τότε ο πράκτορας λαμβάνει τη χειρότερη ανταμοιβή (αρνητική) και το επεισόδιο τερματίζει. Ο πράκτορας λαμβάνει ανταμοιβές σύμφωνα με την Εξίσωση 5.1. Στόχος του πράκτορα είναι να καταφέρει να ισορροπήσει τη ράβδο πάνω στην άμαξα για 1000 βήματα. Τότε, υποθέτουμε πως φθάνουμε στο στόχο μας και ξεκινάμε ένα καινούριο επεισόδιο. Κάθε φορά που ξεκινάμε ένα καινούριο επεισόδιο τοποθετούμε την άμαξα στη αρχική κατάσταση $\mathbf{x} = (0, 0, 0, 0.01)$.

$$\mathbf{r} = \begin{cases} -10000 - 50|x_1| - 10|x_3| & \text{εάν } x_1 > 4.0 \mid x_1 < -4.0 \mid x_3 > 45^\circ \mid x_3 < 45^\circ \\ 10 - 10|10x_3|^2 - |x_1| - 10x_4 & \text{διαφορετικά} \end{cases} \tag{5.1}$$

Αρχικά, στο πρόβλημα του ανεστραμμένου εκκρεμούς εφαρμόσαμε τη μέθοδο ΧΔ, Sarsa. Οι βέλτιστες τιμές των παραμέτρων του αλγορίθμου Sarsa για το συγκεκριμένο

πρόβλημα είναι: ρυθμός μάθησης $\alpha = 0.3$, ρυθμός έκπτωσης $\gamma = 1$ και πιθανότητα επιλογής τυχαίας ενέργειας $\epsilon = 0.001$. Στη συνέχεια εφαρμόσαμε στο συγκεκριμένο πρόβλημα τη μέθοδο ΧΔ με ίχνη επιλογής, Sarsa(λ). Βρήμαμε ότι οι βέλτιστες τιμές των παραμέτρων του αλγορίθμου Sarsa(λ) είναι: ρυθμός μάθησης $\alpha = 0.3$, ρυθμός έκπτωσης $\gamma = 1$, παράμετρος μείωσης του ίχνους $\lambda = 0.5$ και πιθανότητα επιλογής τύχαιας ενέργειας $\epsilon = 0.001$. Τέλος, υλοποιήσαμε τον αλγόριθμο GPTD στο πρόβλημα ελέγχου του Cart Pole. Ανακαλύψαμε πως οι βέλτιστες τιμές των παραμέτρων του συγκεκριμένου αλγορίθμου είναι οι εξής: ρυθμός έκπτωσης $\gamma = 0.999$, διακύμανση του θορύβου $\sigma^2 = 100$ και πιθανότητα επιλογής τύχαιας ενέργειας $\epsilon = 0.001$. Ως συνάρτηση κατάστασης πυρήνα χρησιμοποιούμε μια Γκαουσιάνη συνάρτηση πυρήνα. Επειδή όμως οι συνιστώσες του διανύσματος κατάστασης αντιπροσωπεύουν διαφορετικές έννοιες, τροποποιούμε την Γκαουσιανή συνάρτηση πυρήνα για να ταιριάζει περισσότερο στις ανάγκες μας:

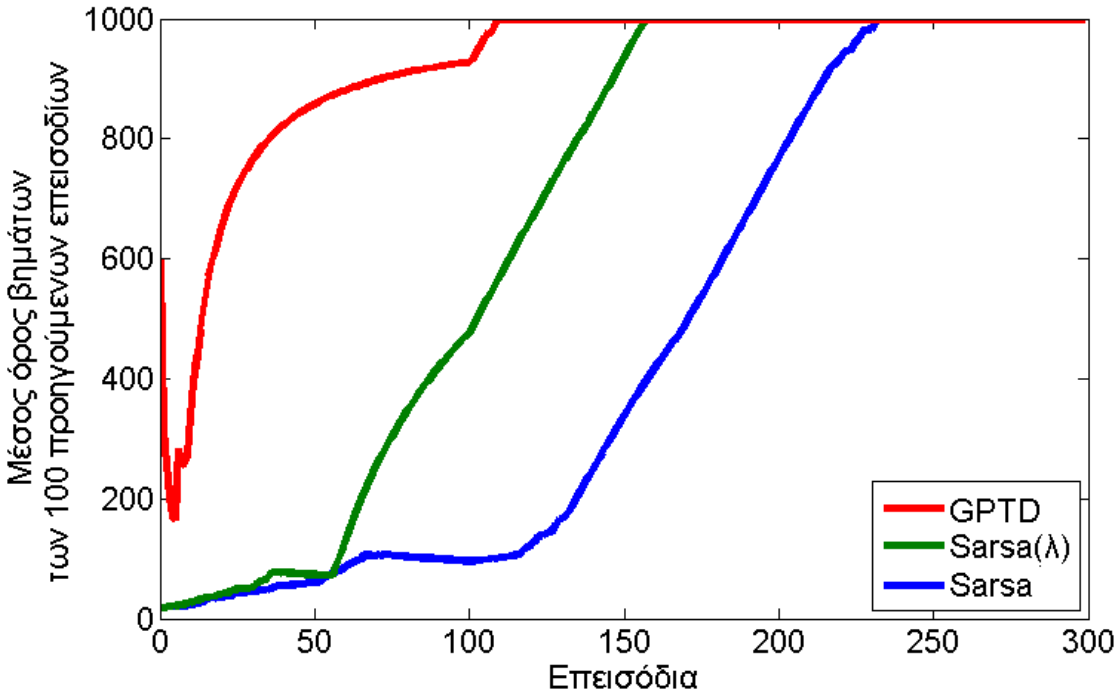
$$k_x(\mathbf{x}, \mathbf{x}') = \exp\left(-\sum_{i=1}^{4} \frac{\|x_i - x_i'\|^2}{2\sigma_i^2}\right) \tag{5.2}$$

Παρατηρούμε διαισθητικά ότι για να υπάρχει ομοιότητα ανάμεσα στις καταστάσεις $\mathbf{x} = (x_1, x_2, x_3, x_4)$ και $\mathbf{x}' = (x_1', x_2', x_3', x_4')$, θα πρέπει: το διάστημα που απέχουν να μην είναι πολύ μεγάλο ($|x_1 - x_1'| < 1.5$), η ταχύτητα που έχει η άμαξα στις δυο καταστάσεις να διαφέρει το πολύ κατά 0.5 ($|x_2 - x_2'| < 0.5$), η διαφορά στις γωνίες που σχηματίζει το κοντάρι από την κατακόρυφο να είναι μικρότερη από 5 μοίρες ($|x_3 - x_3'| < 5^\circ$) και η ταχύτητα κλίσης να διαφέρει το πολύ κατά 0.25 ($|x_4 - x_4'| < 0.25$). Η εύρεση των βέλτιστων διακυμάνσεων για το συγκεκριμένο πρόβλημα, έγινε πειραματικά. Έτσι θεωρούμε τη διακύμανση της θέσης της άμαξας $\sigma_1^2 = 2$, τη διακύμανση της ταχύτητας $\sigma_2^2 = 0.25$, τη διακύμανση της κλίσης της ράβδου $\sigma_3^2 = 0.008$ και τη διακύμανση της ταχύτητας κλίσης της ράβδου $\sigma_4^2 = 0.085$. Επίσης θεωρούμε ότι η συνάρτηση πυρήνα των ενεργειών είναι μια Γκαουσιανή συνάρτηση πυρήνα της μόρφης

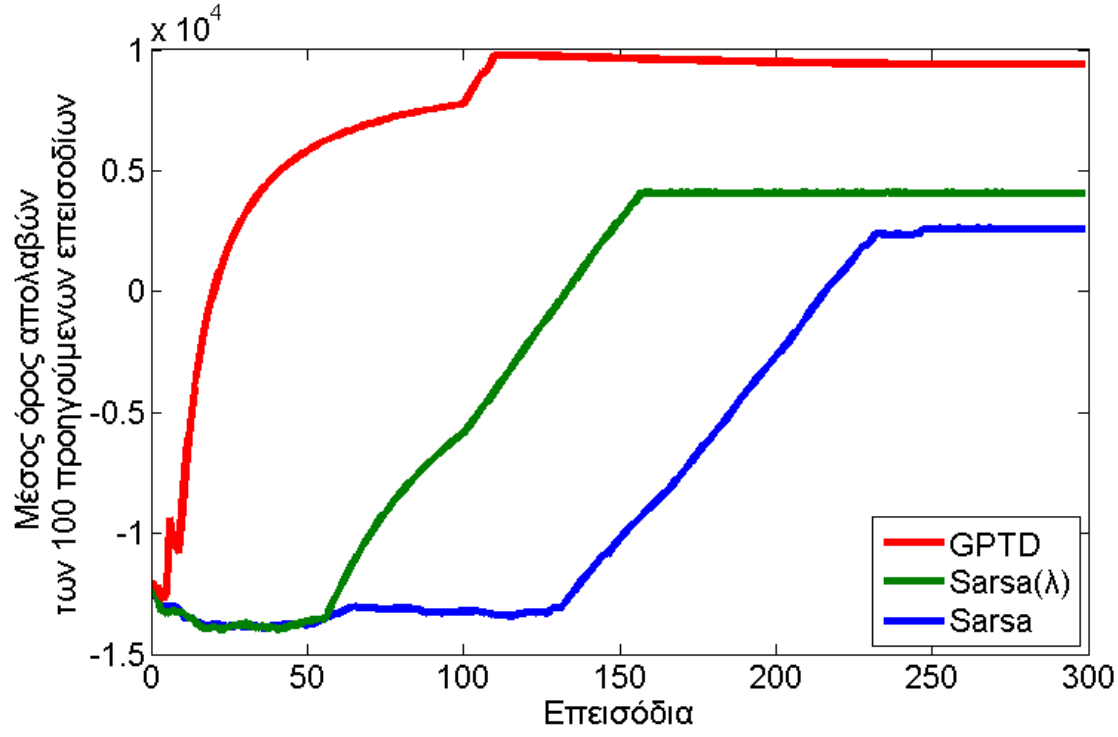
$$k_a(\boldsymbol{a}, \boldsymbol{a}') = \exp\left(-\frac{\|\boldsymbol{a} - \boldsymbol{a}'\|^2}{2\sigma_a^2}\right),$$

όπου $\sigma_a^2 = 1$. Αυτό σημαίνει πως η κάθε ενέργεια παρουσιάζει ομοιότητα μόνο με τις γειτονικές ενέργειες της.

Στα Σχήματα 5.4(a) και 5.4(b), παρουσιάζεται η απόδοση των συγκεκριμένων μεθόδων για το πρόβλημα του ανεστραμμένου εκρεμμούς τόσο ως προς τον αριθμό βημάτων που παραμένει όρθια η ράβδος όσο και ως προς τις απολαβές που λαμβάνει ο πράκτορας. Συγκεκριμένα, στο Σχήμα 5.4(a) παρουσιάζεται ο μέσος όρος των βημάτων που παραμένει όρθια η ράβδος σε κάθε επεισόδιο, αν εφαρμόσουμε τις μεθόδους GPTD, SARSA και SARSA(λ). Αντίστοιχα, στο Σχήμα 5.4(b) παρουσιάζεται ο μέσος όρος των απολαβών που λαμβάνει ο πράκτορας σε κάθε επεισόδιο για κάθε μια από τις τρεις μεθόδους. Γίνεται εύκολα αντιληπτό, πως ο αλγόριθμος GPTD οδηγεί σε πολύ καλές απολαβές, ενώ την ίδια ώρα φαίνεται πως συγκλίνει αρκετά ταχύτερα σε σχέση με τις δυο άλλες μεθόδους.



(a) Μέσος όρος βημάτων που παραμένει όρθια η ράβδος, εφαρμόζοντας τις μεθόδους GPTD, SARSA και SARSA(λ) στο πρόβλημα Cart Pole



(b) Μέσος όρος απολαβών που λαμβάνει ο πράκτορας, εφαρμόζοντας τις μεθόδους GPTD, SARSA και SARSA(λ)στο πρόβλημα Cart Pole

Σχήμα 5.4: Σύγκριση των μεθόδων GPTD, SARSA και SARSA(λ) στο πρόβλημα Cart Pole

5.2.3 Αυτόνομη Πλοήγηση Ρομποτικού Συστήματος

Σχήμα 5.5: Ρομποτικό Σύστημα PeopleBot

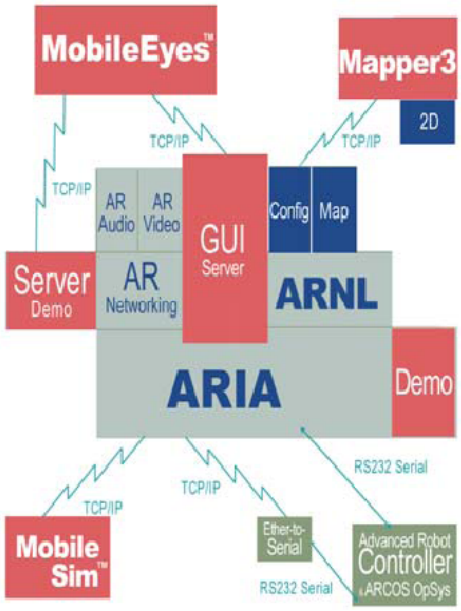
Γενική περιγραφή ρομποτικού συστήματος

Σε αυτή την ενότητα, θα μελετήσουμε το πρόβλημα της αυτόνομης πλοήγησης ενός ρομποτικού συστήματος, καθώς και πως αυτό επιλύεται με τη χρήση μεθόδων ενισχυτικής μάθησης. Συγκεκριμένα, θα προσπαθήσουμε να εκπαιδεύσουμε ένα πραγματικό ρομπότ έτσι ώστε να κινείται αυτόνομα σε το χώρο χωρίς συγκρούσεις. Για την μελέτη του συγκεκριμένου προβλήματος χρησιμοποιήθηκε ένα πραγματικό ρομποτικό σύστημα τύπου Pioneer PeopleBot (Σχημα 5.5), πάνω στο οποίο εφαρμόστηκαν και αξιολογήθηκαν οι μέθοδοι ενισχυτικής μάθησης που μελετούμε στο παρών κεφάλαιο. Το PeopleBot είναι ένα ρομπότ σχεδιασμένο για εργασίες υπηρεσιών και αλληλεπίδρασης με ανθρώπους. Είναι βασισμένο στην ισχυρή βάση P3-DX και έχει ένα βραχίονα για να διευκολύνει την επικοινωνία με τους ανθρώπους. Επίσης, το ρομπότ εξοπλίζεται με μια stereo camera που μπορεί να χρησιμοποιηθεί σε διάφορα προβλήματα υπολογιστικής όρασης.

Το ρομπότ έχει στη διάθεσή του διάφορους τύπους αισθητήρων για την αποφυγή εμποδίων. Στις εφαρμογές μας χρησιμοποιούμε τους αισθητήρες laser και sonar του ρομπότ για τον εντοπισμό εμποδίων. Το sonar έχει τη δυνατότητα εντοπισμού κάποιου εμποδίου σε απόσταση πέντε μέτρων σε αντίθεση με το laser που μπορεί να εντοπίσει κάποιο εμπόδιο μέχρι και είκοσι πέντε μέτρα μακριά. Η διαφορά τους είναι ότι το sonar μπορεί να εντοπίσει κάποιο εμπόδιο 360° γύρω από το ρομπότ σε αντίθεση με το laser που δεν μπορεί να εντοπίσει εμπόδια που βρίσκονται πίσω από το ρομπότ. Στο Σχήμα 5.7, με μπλε χρώμα φαίνονται οι ακτίνες που εκπέμπει το laser, ενώ οι ακμές ανήκουν στο sonar. Οι δύο παραπάνω τύποι αισθητήρων επιστρέφουν τη θέση του κοντινότερου εμποδίου που συναντάνε κάθε φορά.

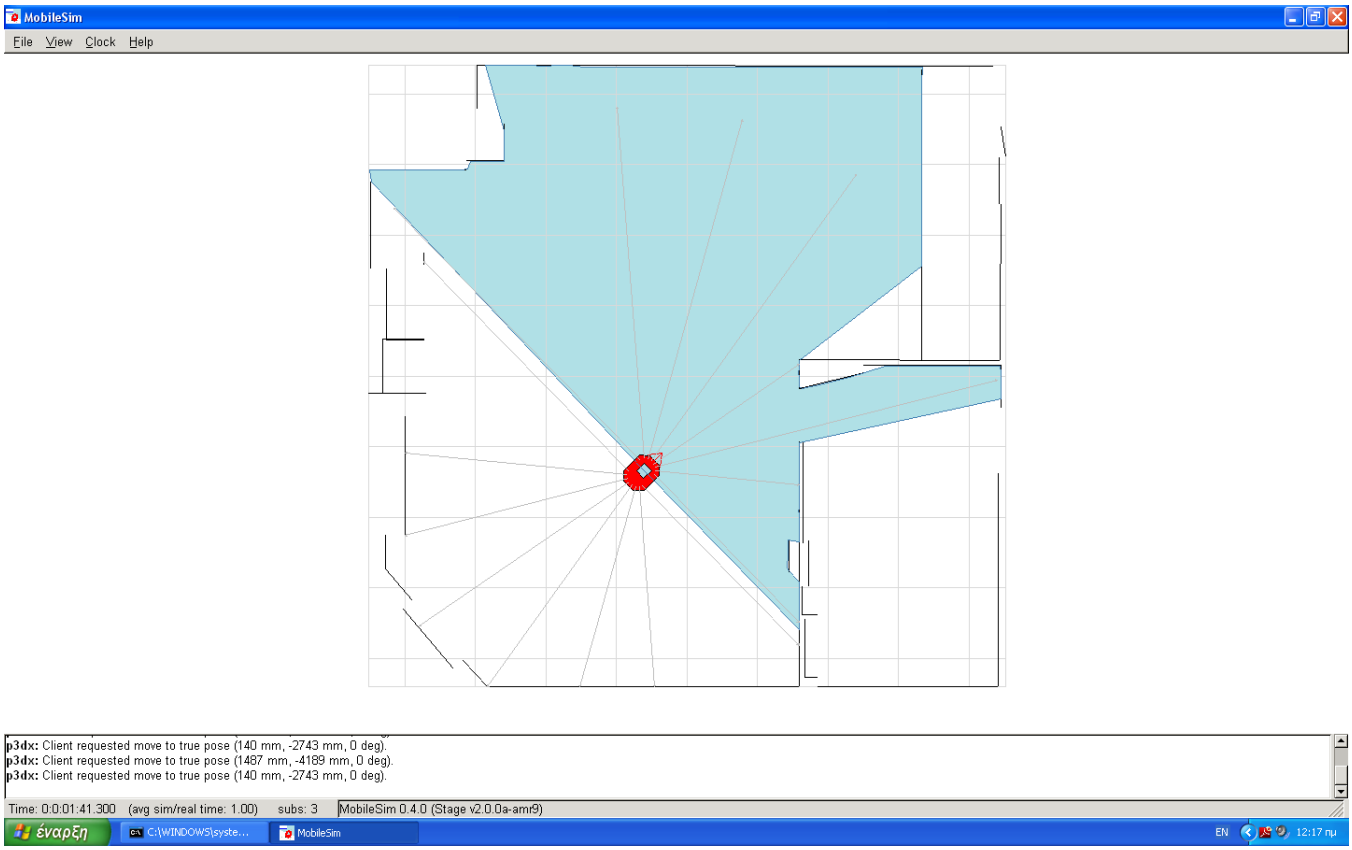
Η επικοινωνία με το συγκεκριμένο ρομπότ γίνεται χάρη σε μια συλλογή βιβλιοθηκών και εφαρμογών (Software Development Kit (SDK) Pioneer) (Σχήμα 5.6), που ελέγχουν τις λειτουργίες της συγκεκριμένης πλατφόρμας. Συγκεκριμένα, οι βιβλιοθήκες μέσω τον οποίον μπορούμε να επικοινωνήσουμε το ρομπότ είναι: η Aria, η Arnl και η ArNetworking. Στα πειράματα μας χρησιμοποιήθηκε κυρίως η βιβλιοθήκη Aria. Η Aria παρέχει μια

διασύνδεση και ένα πλαίσιο για τον έλεγχο και την λήψη δεδομένων από το ρομπότ και είναι γραμμένη σε γλώσσα προγραμματισμού C++. Για το λόγο αυτό, οι αλγόριθμοι που υλοποιήσαμε για το συγκεκριμένο περιβάλλον έχουν υλοποιηθεί σε C++. Επίσης, για την υλοποίηση των εφαρμογών μας χρησιμοποιήθηκαν και οι εξής εφαρμογές: MobileEyes, Mapper3, MobileSim.



Σχήμα 5.6: Pioneer Software Development Kit (SDK)

Εξαιτίας διάφορων φυσικών περιορισμών (π.χ μπαταρία, μεγάλη διάρκεια πειραμάτων), δεν είχαμε την δυνατότητα να εφαρμόσουμε τους αλγορίθμους μας απευθείας στο πραγματικό ρομπότ. Έτσι η εφαρμογή των αλγορίθμων μας και κατά συνέπεια η εκπαίδευση του ρομπότ, έγινε στο προσομοιωτή που είναι διαθέσιμος (MobileSim). Ο προσομοιωτής προσομοιώνει πλήρως το πραγματικό περιβάλλον μέσα στο οποίο κινείται και αλληλεπιδρά το ρομπότ μας, δίνοντάς μας τη δυνατότητα να εκπαιδεύσουμε το ρομπότ σε πραγματικές συνθήκες. Αρχικά, με τη βοήθεια του ρομπότ (χρησιμοποιήσαμε το λογισμικό MobileEyes, Mapper καθώς επίσης και τη βιβλιοθήκη Arnl,Aria,ArNetworking) κατασκευάζουμε το χάρτη του περιβάλλοντος (Σχήμα 5.9) μέσα στο οποίο θέλουμε να μάθει να περιπλανείται αυτόνομα το ρομπότ μας. Έπειτα, εκπαιδεύουμε το ρομπότ στο προσομοιωτή (Σχήμα 5.7), έτσι ώστε να μάθει από μόνο του το χώρο μέσα στον οποίο θα κινείται. Αφού εκπαιδεύσουμε το ρομπότ με τη βοήθεια του προσομοιωτή, εφαρμόζουμε τη μέθοδο απευθείας στο πραγματικό ρομπότ για να αξιολογήσουμε τη συμπεριφορά του σε συνθήκες του πραγματικού κόσμου.



Σχήμα 5.7: Προσομοιωτής Πειραματικού Περιβάλλοντος

Ανάλυση προβλήματος

Στο παρών πρόβλημα, ο πράκτορας θα πρέπει να επιλύσει ένα πρόβλημα πλοήγησης στο χώρο του εργαστηριού ρομποτικής του τμήματος Πληροφορικής που παρουσιάζεται στο Σχήμα 5.8. Συγκεκριμένα θα πρέπει να εντοπίσει το συντομότερο μονοπάτι από οποιοδήποτε σημείο $\mathcal{X}$ σε κάποιο σημείο στόχο (GOAL) (Σχήμα 5.9), ενώ ταυτόχρονα αποφεύγει πιθανά εμπόδια που εμφανίζονται στη πορεία του. Ο πράκτορας θεωρούμε ότι έχει την δυνατότητα να επιλέξει ανάμεσα από 8 δυνατές ενέργειες. Δηλαδή μπορεί να στρίψει κατά $0, 45, 90, 135, 180, 225, 270, 315$ μοίρες και στη συνέχεια να κάνει ένα βήμα. Στις εφαρμογές θεωρούμε ότι κάνει ένα βήμα ενός μέτρου. Οπότε μπορούμε να αναπαραστήσουμε την ενέργεια ως ένα μοναδιαίο διάνυσμα $\boldsymbol{a} = (\cos(\theta), \sin(\theta))$ με κατεύθυνση τη κατεύθυνση της αντίστοιχης κίνησης, κατ'αυτό το τρόπο το διάνυσμα $\boldsymbol{a}$ ανήκει στο μοναδιαίο κύκλο. Η θέση του ρομπότ στο χώρο προσδιορίζει την κατάσταση στην οποία βρίσκεται εκείνη τη στιγμή. Έτσι οι καταστάσεις του περιβάλλοντος ορίζονται ως διανύσματα δύο διαστάσεων, $\mathbf{x} = (x_1, x_2)$, όπου η x_1 προσδιορίζει τη συντεταγμένη x και η x_2 προσδιορίζει τη συντεταγμένη y στο χώρο. Η συγκεκριμένη εργασία εκτελείται σε επεισόδια. Η ανταμοιβή που λαμβάνει ο πράκτορας σε κάθε βήμα είναι -1 έκτος από την περίπτωση που φτάνει στη κατάσταση-στόχο, όπου λαμβάνει 0. Στη περίπτωση που βρίσκετε σε απόσταση μικρότερη από 0.5 μέτρα από κάποιο εμπόδιο ο πράκτορας παίρνει ανταμοιβή -100, το επεισόδιο τερματίζει και ξεκινά ένα καινούριο επεισόδιο από μια τυχαία αρχική θέση $\mathbf{x}'$. Στόχος του πράκτορα (ρομπότ) είναι

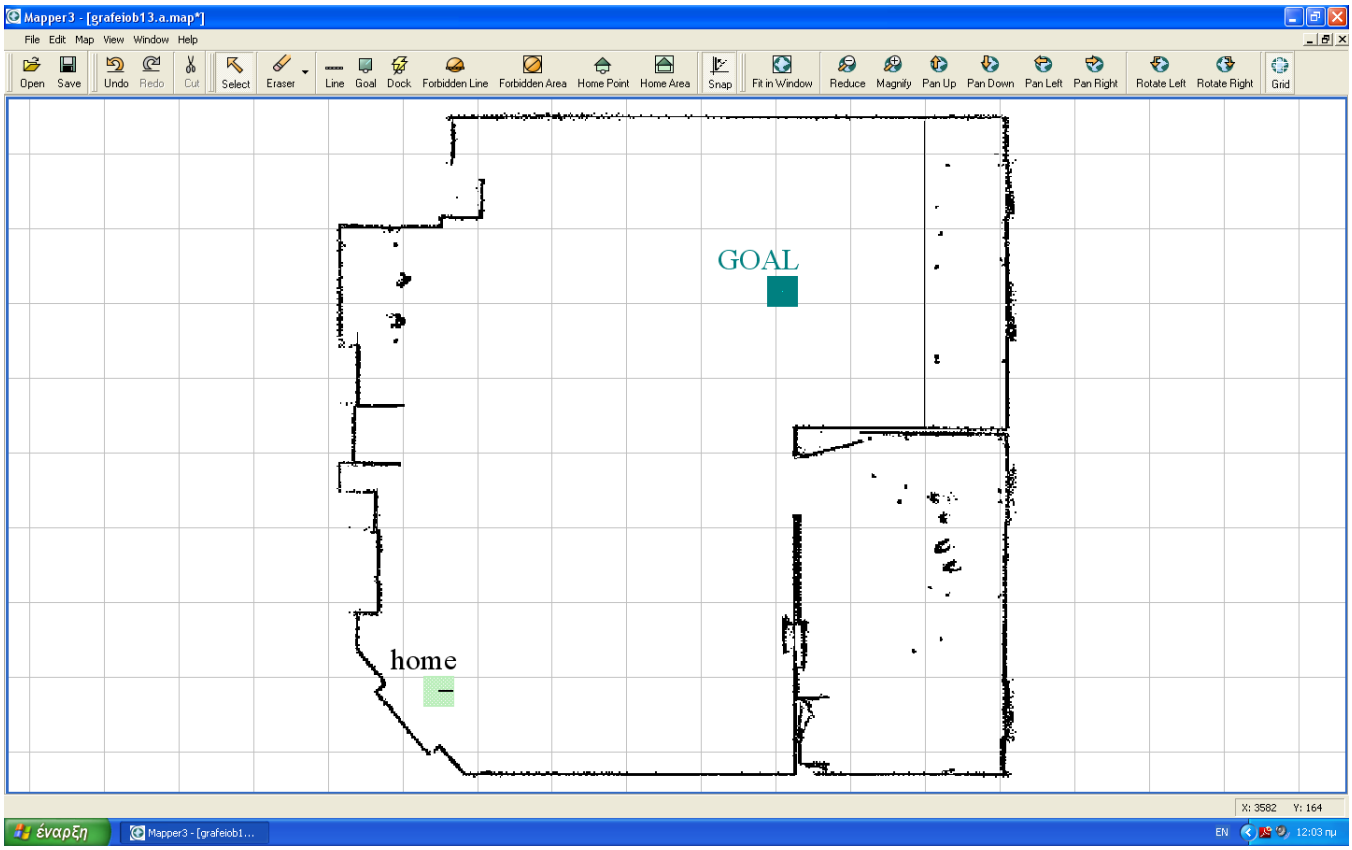
Σχήμα 5.8: Πραγματικό Περιβάλλον Εκπαίδευσης

να φθάσει στη κατάσταση-στόχο σε λιγότερο από 100 βήματα βρίσκοντας ταυτόχρονα τη συντομότερη διαδρομή (βέλτιστο μονοπάτι). Τότε ξεκινάει ένα καινούριο επεισόδιο και το ρομπότ τοποθετείται σε μια τυχαία θέση στο χώρο. Στο Σχήμα 5.7 μπορούμε να δούμε την εκπαίδευση του ρομπότ στο χώρο του εργαστηρίου με τη βοήθεια του προσομοιωτή.

Επιπλέον, μπορούμε να θεωρήσουμε ότι η κατάσταση στην οποία βρίσκεται το ρομπότ ορίζεται από ένα διάνυσμα N συντεταγμένων, $\mathbf{x} = (x_1, x_2, \ldots, x_N)$. Ουσιαστικά, χωρίζουμε το χώρο γύρω από το ρομπότ σε N "φέτες" και κάθε συντεταγμένη x_i του διανύσματος ορίζει την κανονικοποιημένη απόσταση που απέχει το ρομπότ από το κοντινότερο εμπόδιο στη συγκεκριμένη περιοχή. Μετά από πειράματα διαπιστώσαμε ότι η συγκεκριμένη αναπαράσταση δεν βοηθά ικανοποιητικά για να λύσουμε πλήρως το πρόβλημα της πλοήγησης. Αυτό οφείλεται στο γεγονός ότι το ρομπότ μπορεί να αναπαραστήσει δύο διαφορετικές καταστάσεις με το ίδιο διάνυσμα. Παρόλα αυτά παρατηρήσαμε πως αν αναπαραστήσουμε τις καταστάσεις με τη συγκεκριμένη μορφή μπορούμε να λύσουμε πλήρως το πρόβλημα της αποφυγής εμποδίων.

Εφαρμογή και πειραματικά αποτελέσματα

Αρχικά, για την επίλυση του παραπάνω προβλήματος υλοποιήσαμε τον αλγόριθμο ΧΔ, Sarsa. Οι βέλτιστες τιμές των παραμέτρων του αλγορίθμου Sarsa βρέθηκε πως είναι οι εξής: ρυθμός μάθησης $\alpha = 1$, ρυθμός έκπτωσης $\gamma = 1$ και πιθανότητα επιλογής τυχαίας ενέργειας $\epsilon = 0.3$. Έπειτα εφαρμόσαμε τον αλγόριθμο ΧΔ με ίχνη επιλεξιμότητας, Sarsa(λ). Διαπιστώσαμε πως οι βέλτιστες τιμές των παραμέτρων του αλγορίθμου Sarsa(λ) είναι: ρυθμός



Σχήμα 5.9: Χάρτης Περιβάλλοντος Εκπαίδευσης

μάθησης $\alpha = 1$, ρυθμός έκπτωσης $\gamma = 1$, παράμετρος μείωσης του ίχνους $\lambda = 0.5$ και πιθανότητα επιλογής τυχαίας ενέργειας $\epsilon = 0.3$. Θα πρέπει να σημειωθεί ότι σε κάθε επεισόδιο των πιο πάνω αλγορίθμων, ο ρυθμός μάθησης μειώνεται κατά ένα παράγοντα 0.98. Στη συνέχεια, για το πρόβλημα πλοήγησης του ρομποτικού συστήματος υλοποιήσαμε τον αραιό αλγόριθμο GPTD. Βρήκαμε ότι οι βέλτιστες τιμές των παραμέτρων του αλγορίθμου GPTD για το συγκεκριμένο πρόβλημα είναι οι εξής: ρυθμός έκπτωσης $\gamma = 0.999$, διακύμανση του θορύβου $\sigma^2 = 1$ και πιθανότητα επιλογής τυχαίας ενέργειας $\epsilon = 0.3$. Ως συνάρτηση πυρήνα της κατάστασης k_x χρησιμοποιούμε μια Γκαουσιανή συνάρτηση της μορφής

$$k_x(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma_x^2}\right),$$

όπου σ_x^2 είναι το πλάτος των καταστάσεων. Θεωρούμε ότι δύο καταστάσεις είναι όμοιες μεταξύ τους όταν η μεταξύ τους απόσταση είναι μικρή. Για το λόγο αυτό, θεωρούμε τη διακύμανση των καταστάσεων $\sigma_x^2 = 1$. Δηλαδή, δυο καταστάσεις είναι αρκετά όμοιες μεταξύ τους όταν η μια βρίσκεται σε απόσταση μικρότερη του ενός μέτρου από την άλλη. Ορίζουμε τη συνάρτηση πυρήνα ενέργειας k_a ως ακολούθως

$$k_a(\boldsymbol{a}, \boldsymbol{a}') = 1 + \frac{(1-b)}{2}(\boldsymbol{a}^\top \boldsymbol{a}' - 1),$$

όπου b είναι μια σταθερά στο διάστημα $[0, 1]$. Εφόσον το εσωτερικό γινόμενο $\boldsymbol{a}^\top \boldsymbol{a}'$ είναι το συνημίτονο της γωνίας ανάμεσα στην $\boldsymbol{a}$ και στην $\boldsymbol{a}'$, η $k_a(\boldsymbol{a}, \boldsymbol{a}')$ επιτυγχάνει τη μέγιστη

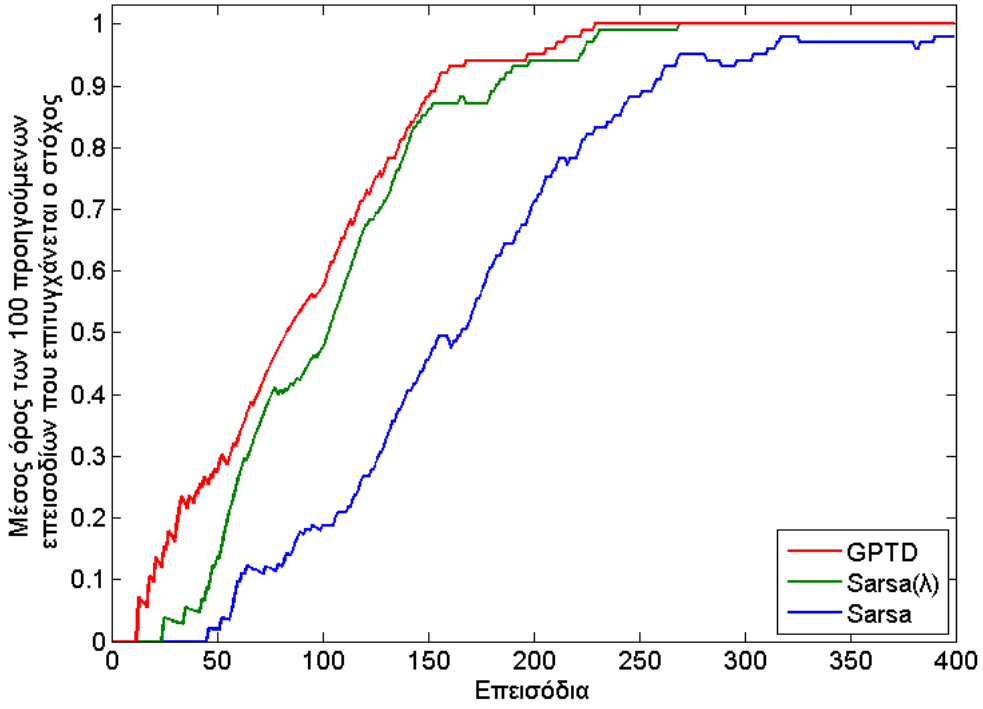
τιμή 1 όταν οι δύο ενέργειες είναι οι ίδιες και την ελάχιστη τιμή b όταν οι ενέργειες είναι κατά 180° αντίθετες.

Συγκριτικά αποτελέσματα των παραπάνω μεθόδων για το πρόβλημα της αυτόνομης πλοήγησης ενός ρομπότ, παρουσιάζονται στα Σχήματα 5.10(a) και 5.10(b). Στο Σχήμα 5.10(a), περιγράφεται το ποσοστό επιτυχίας εύρεσης στόχου των 100 προηγούμενων επεισοδίων μετά την εφαρμογή των μεθόδων GPTD, SARSA, SARSA(λ) στο πρόβλημα. Γίνεται εύκολα αντιληπτό ότι ο αλγόριθμος GPTD συγκλίνει οριακά γρηγορότερα σε σχέση με τον αλγόριθμο SARSA(λ). Αντίθετα, τόσο ο GPTD όσο και ο SARSA(λ) συγκλίνουν εμφανώς γρηγορότερα σε σχέση με τον αλγόριθμο SARSA. Στο Σχήμα 5.10(b), βλέπουμε το μέσο όρο των απολαβών που λαμβάνει ο πράκτορας εφαρμόζοντας τις μεθόδους GPTD, SARSA και SARSA(λ) στο πρόβλημα. Μπορούμε εύκολα να διαπιστώσουμε ότι οι αλγόριθμοι GPTD και SARSA(λ) συγκλίνουν σχεδόν ταυτόχρονα στη βέλτιστη πολιτική με αποτέλεσμα να οδηγούν στις ίδιες σχεδόν ανταμοιβές. Βλέποντας το Σχήμα 5.10 διαπιστώνουμε ότι και οι τρείς αλγόριθμοι συγκλίνουν στην ίδια σχεδόν πολιτική, με μόνη διαφορά τους το ρυθμό σύγκλισης στη συγκεκριμένη πολιτική.

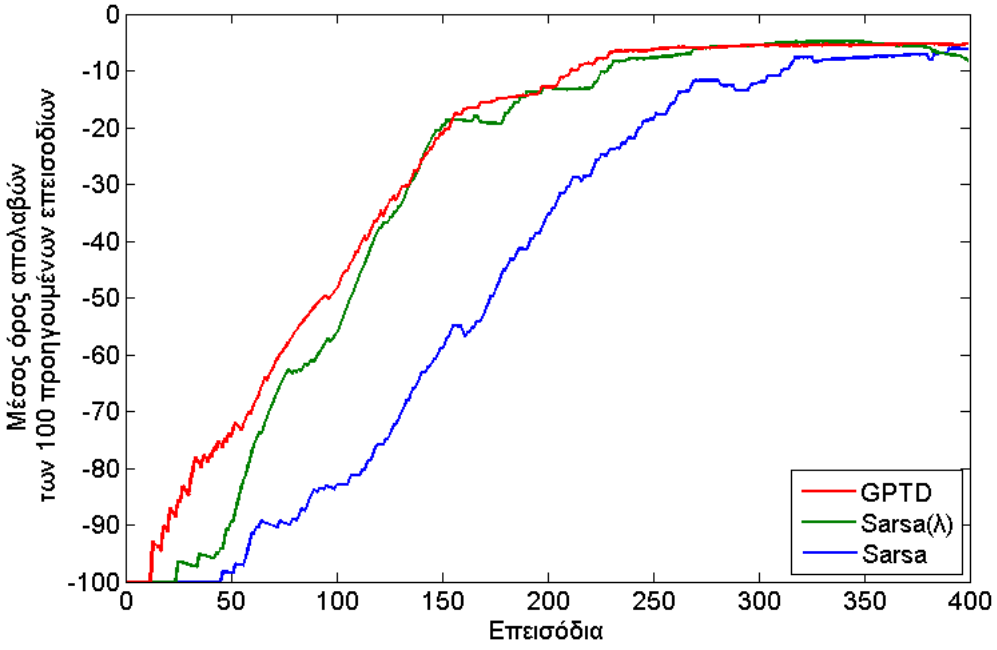
Στη συνέχεια, τροποποιούμε τον αλγόριθμο GPTD έτσι ώστε να λαμβάνουμε υπόψη τις ομοιότητες ανάμεσα στις καταστάσεις, σε ένα μεγαλύτερο εύρος. Αυτό επιτυγχάνεται αν αντί για μια Γκαουσιανή συνάρτηση πυρήνα, επιλέξουμε η συνάρτηση πυρήνα κατάστασης να είναι ένας γραμμικός συνδυασμός Γκαουσιανών συναρτήσεων πυρήνα με διαφορετικές τιμές πλάτους. Στο συγκεκριμένο πείραμα χρησιμοποιήθηκαν $N = 10$ διαφορετικές τιμές πλάτους. Με το τρόπο αυτό, λαμβάνουμε υπόψη ένα μεγαλύτερο εύρος τιμών των καταστάσεων και πετυχαίνουμε έναν ορθότερο υπολογισμό ομοιότητας μεταξύ δύο καταστάσεων. Η συνάρτηση πυρήνα της κατάστασης παίρνει τη μορφή:

$$k_x(\mathbf{x}, \mathbf{x}') = \sum_{i=0}^{9} w_i \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma_i^2}\right), \text{ ισχύει ότι } \sum_{i=0}^{9} w_i = 1.$$

Στό συγκεκριμένο πείραμα θεωρήσαμε ότι $w_i = 0.1$ και $\sigma_i^2 = 0.2 + 0.2i, \quad \forall i$. Για να διακρίνουμε τις δύο μεθόδους, ορίζουμε τη μέθοδο που προκύπτει από την GPTD ως GPTDL. Στα Σχήματα 5.11(a) και 5.11(b) παρουσιάζονται κάποια συγκριτικά αποτελέσματα των δύο αυτών αλγορίθμων για το πρόβλημα της αυτόνομης πλοήγησης ενός ρομπότ. Από το Σχήμα 5.11(a) γίνεται εμφανές ότι ο αλγόριθμος GPTDL συγκλίνει ταχύτερα σε σχέση με τον αλγόριθμο GPTD, ενώ σε συνδυασμό με το Σχήμα 5.11(a) γίνεται ξεκάθαρο ότι ο αλγόριθμος GPTDL συγκλίνει γρηγορότερα στη βέλτιστη πολιτική.

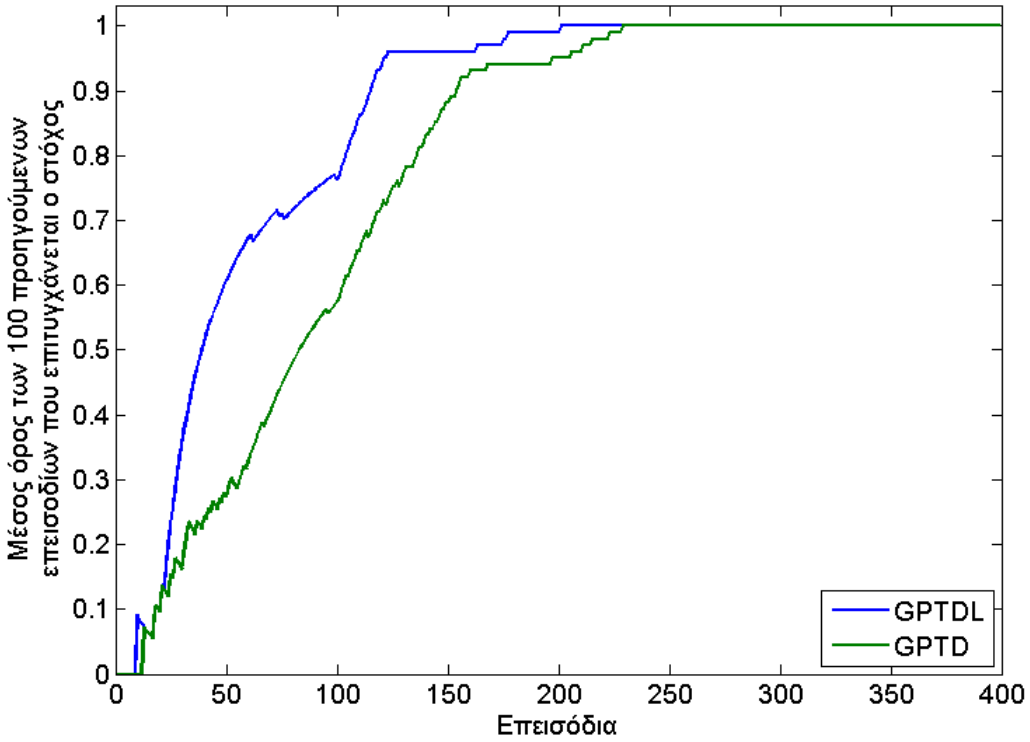


(a) Ποσοστό επιτυχίας ευρεσης στόχου των 100 προηγούμενων επεισοδίων, εφαρμόζοντας τις μεθόδους GPTD, SARSA και SARSA(λ) στο πρόβλημα πλοήγησης ενός ρομποτικού συστήματος

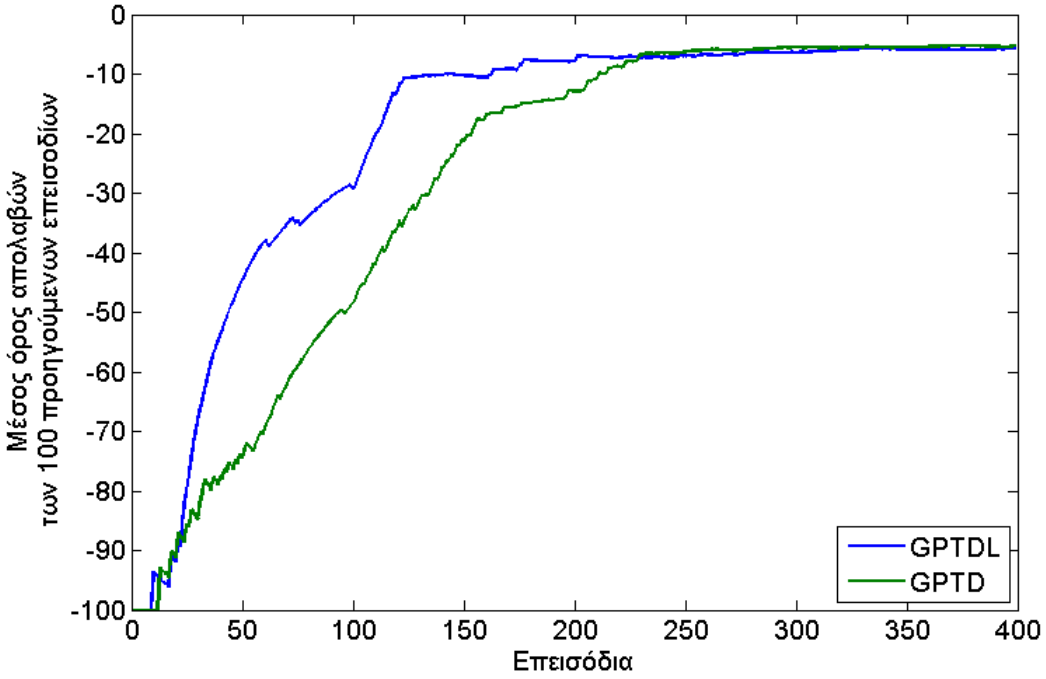


(b) Μέσος όρος απολαβών που λαμβάνει ο πράκτορας, εφαρμόζοντας τις μεθόδους GPTD, SARSA και SARSA(λ)στο πρόβλημα αυτόνομης πλοήγησης ενός ρομποτικού συστήματος

Σχήμα 5.10: Σύγκριση των μεθόδων GPTD, SARSA και SARSA(λ) στο πρόβλημα αυτόνομης πλοήγησης ενός ρομποτικού συστήματος



(a) Ποσοστό επιτυχίας εύρεσης στόχου των 100 προηγούμενων επεισοδίων, εφαρμόζοντας τις μεθόδους GPTD και GPTDL στο πρόβλημα πλοήγησης ενός ρομποτικού συστήματος



(b) Μέσος όρος απολαβών που λαμβάνει ο πράκτορας, εφαρμόζοντας τις μεθόδους GPTD και GPTDL στο πρόβλημα αυτόνομης πλοήγησης ενός ρομποτικού συστήματος

Σχήμα 5.11: Σύγκριση των μεθόδων GPTD και GPTDL στο πρόβλημα αυτόνομης πλοήγησης ενός ρομποτικού συστήματος

Κεφαλαιο 6

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στη παρούσα διατριβή ασχοληθήκαμε με το πρόβλημα της αυτόνομης πλοήγησης ρομποτικών συστημάτων με τεχνικές ενισχυτικής μάθησης. Η ενισχυτική μάθηση αντιμετωπίζει το πρόβλημα της εκπαίδευσης μιας βέλτιστης συμπεριφοράς ενός πράκτορα μέσω της αλληλεπίδρασης του με το περιβάλλον. Αρχικά μελετήσαμε τις βασικές κλάσεις μεθόδων για την επίλυση του προβλήματος της ενισχυτικής μάθησης. Στη συνέχεια περιγράψαμε μια Μπεϋσιανή προσέγγιση εκτίμησης πολιτικής σε γενικούς χώρους καταστάσεων και ενεργειών, η οποία χρησιμοποιεί στατιστικά γεννητικά μοντέλα μέσω Γκαουσιανών διαδικασιών για τις συναρτήσεις αξίας. Η επέκταση που προτάθηκε στη συγκεκριμένη διατριβή εισάγει τη χρήση RVM στη Μπεϋσιανή προσέγγιση εκτίμησης πολιτικής και οδηγεί σε ακόμη αραιότερα μοντέλα εκτίμησης της συνάρτησης αξίας. Οι παραπάνω μέθοδοι εφαρμόστηκαν και αξιολογήθηκαν σε δύο από τα πιο γνωστά πειραματικά περιβάλλοντα: το Mountain Car και το ανεστραμμένο εκκρεμές (Cart Pole). Και στα δύο προβλήματα πλοήγησης παρατηρήσαμε ότι η Μπεϋσιανή προσέγγιση (GPTD) οδηγεί σε πολύ καλύτερη συμπεριφορά σε σχέση με τις μεθόδους Sarsa και Sarsa(λ), τόσο στο χρόνο σύγκλισης όσο και στην ανταμοιβή που λαμβάνει ο πράκτορας για να φθάσει στο στόχο. Επιπλέον, μελετήσαμε το πρόβλημα ενός πραγματικού ρομπότ (PeopleBot) στο χώρο του εργαστηρίου ρομποτικής του τμήματος. Σε αυτό το πείραμα παρατηρήσαμε ότι η Μπεϋσιανή προσέγγιση (GPTD) είναι οριακά καλύτερη σε σχέση με την μέθοδο Sarsa(λ) τόσο στο χρόνο σύγκλισης όσο και στην ανταμοιβή που λαμβάνει ο πράκτορας. Απεναντίας εξακολουθεί να είναι αισθητά καλύτερη σε σχέση με τη μέθοδο Sarsa. Στο συγκεκριμένο πείραμα τροποποιήσαμε τη Μπεϋσιανή προσέγγιση (GPTDL) επιλέγοντας τη συνάρτηση πυρήνα κατάστασης να είναι ένας γραμμικός συνδυασμός $N = 10$ Γκαουσιανών συναρτήσεων πυρήνα με διαφορετικές τιμές πλάτους και παρατηρήσαμε ότι οδηγεί αρκετά γρηγορότερα στη εύρεση καλύτερων πολιτικών σε σχέση με τις υπόλοιπες μεθόδους. Η υλοποίηση και η εφαρμογή των παραπάνω μεθόδων στο πρόβλημα πλοήγησης ενός πραγματικού ρομποτικού συστήματος μας έδωσε τη δυνατότητα να αξιολογήσουμε τις παραπάνω μεθόδους σε πραγματικές συνθήκες.

Βιβλιογραφια

[1] R. Bellman. Dynamic Programming. *Princeton University Press*, Princeton, NJ, 1957.

[2] D. P. Bertsekas. Dynamic Programming: Deterministic and Stochastic Models. *Prentice-Hall*, Englewood Cliffs, NJ, 1987.

[3] D. P. Bertsekas, J. N. Tsitsiklis. Neuro-Dynamic Programming. *Athena Scientific*, 1996.

[4] C. Bishop. *Pattern Recognition and Machine Learning*. Springel, 2006.

[5] Y. Engel. Algorithms and Representations for Reinforcement Learning. *Phd Thesis, Herbrew University*, 2005.

[6] Y. Engel, S. Mannor, R. Meir. *Bayesian Reinforcement Learning with Gaussian Process Temporal Difference Methods*, 2007.

[7] Y. Engel, S. Mannor, R. Meir. Reinforcement Learning with Gaussian Processes. 22^{nd} *International Conference on Machine Learnign*, 2005.

[8] J. Gao, P. W. Kwan, D. Shi. Sparse kernel learning with LASSO and Bayesian inference algorithm. *Neural Networks 23*, 257–264, 2010.

[9] L. P. Kaelbling, M. L. Littman, A. W. Moore. Reinforcement Learning: A Survey, *Journal of Artificial Intelligence Research 4*, 237–285, 1996.

[10] C. E. Rasmussen, C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[11] G. A. Rummery, M. Niranjan. On-line Q-Learning using connectionist systems. CUED/F-INFENG/TR 166, Cambridge University, 1994.

[12] L. L. Scharf. Statistical Signal Processing. *Addison-Wesley*, 1991.

[13] S. P. Singh, R. S. Sutton. Reinforcement learning with replacing eligibility traces. *Machine Learning*, 1996.

[14] R. S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3(1), 9–44, 1988.

[15] R. S. Sutton, A. G. Barto. *An Introduction to Reinforcement Learning*. The MIT Press, 1998.

[16] G. Tesauro. Temporal difference learning and TD-Gammon. *Comm. ACM*, 38:58-68, 1995

[17] M. E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Reasearch 1*, 211–244, 2001.

[18] C. J. Watkins. Learning from Delayed Reward. *Phd Thesis, King's College, Cambridge, UK*, 1989.

[19] C. J. Watkins, P. Dayan. Q-learning. *Machine Learning*, 8(3), 279–292, 1991.

[20] R. J. Williams, L. C. Baird. Tight performance bounds on greedy policies based on imperfect value functions. *Tech. rep. NU-CCS-93-42*, Northeastern University, College of Computer Science, Boston, MA, 1993b.

Βιογραφικο

Ο Νικόλαος Τζιωρτζιώτης γεννήθηκε στα Τρίκαλα το 1984. Το 2002 αποφοίτησε από το 2ο Ενιαίο Λύκειο Τρικάλων και εισήχθη στο Τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων από το οποίο αποφοίτησε το 2007. Το 2008 συνέχισε τις σπουδές του στο Πρόγραμμα Μεταπτυχιακών Σπούδων του ίδιου τμήματος από όπου αποφοίτησε το Σεπτέμβριο του 2010 αποκτώντας ειδίκευση στις "Τεχνολογίες-Εφαρμογές". Το 2010 πήρε υποτροφία απο το Τμήμα Πληροφορίκης του Πανεπιστημίου Ιωαννίνων για τη διεξαγωγή φροντιστηριακών και εργαστηριακών ασκήσεων. Τα ερευνητικά του ενδιαφέροντα εστιάζονται κυρίως στους τομείς της Ρομποτικής, της Μηχανικής Μάθησης, της Αναγνώρισης Προτύπων και της Εξόρυξης Δεδομένων.