ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ
ΥΠΟΛΟΓΙΣΤΩΝ

Υλοποίηση ελεγκτή κλειστού συστήματος ελέγχου γλυκόζης με χρήση αλγορίθμων βαθιάς ενισχυτικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τσολίσου Δάφνη

Επιβλέπων: Ανδρέας Σταφυλοπάτης
Καθηγητής Ε.Μ.Π

Αθήνα, Νοέμβρης 2019

(Η σελίδα αυτή είναι σκόπιμα λευκή.)

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ
ΥΠΟΛΟΓΙΣΤΩΝ

Υλοποίηση ελεγκτή κλειστού συστήματος ελέγχου γλυκόζης με χρήση αλγορίθμων βαθιάς ενισχυτικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τσολίσου Δάφνη

Επιβλέπων: Ανδρέας Σταφυλοπάτης
Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή, 18 Νοεμβρίου 2019.

.....	.....	.....
Ανδρέας Σταφυλοπάτης	Γεώργιος Στάμου	Κωνσταντίνα Σπ. Νικήτα
Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Καθηγήτρια Ε.Μ.Π

Αθήνα, Νοέμβρης 2019

.....

Τσολίσου Δάφνη

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Ο διαβήτης τύπου ένα είναι μία αυτοάνοση ασθένεια η οποία καταστρέφει τα β-κύτταρα του παγκρέατος που παράγουν ινσουλίνη, την ορμόνη υπεύθυνη για την απορρόφηση της γλυκόζης. Οι διαβητικοί αυτορυθμίζουν τη γλυκόζη τους καθημερινά με ενέσεις ινσουλίνης ή μέσω αντλίας. Η δημιουργία ενός κλειστού συστήματος ελέγχου γλυκόζης, το ονομαζόμενο τεχνητό πάγκρεας, με συνδυασμό συσκευής συνεχής μέτρησης γλυκόζης, αντλίας ινσουλίνης και ενός αλγορίθμου ελέγχου θα βελτιώσει την ποιότητα ζωής τους. Ο αλγόριθμος είναι το βασικότερο συστατικό για την κατασκευή πλήρως αυτοματοποιημένου συστήματος κάτι που μέχρι σήμερα δεν υπάρχει λόγω μη γραμμικών παραμέτρων στο μοντέλου γλυκόζης-ινσουλίνης. Χρησιμοποιώντας τον αλγορίθμο ενισχυτικής μάθησης DQN εκπαιδεύσαμε τρεις πράκτορες με ίδιο σύνολο ενεργειών ινσουλίνης τριών δόσεων και συνάρτηση ανταμοιβής με βάση τον δείκτη συνολίκου κινδύνου στο περιβάλλον του προσομοιωτή του μοντέλου γλυκόζης-ινσουλίνης T1D UVA/Padova S2008 σε python3. Ο πρώτος πράκτορας εκπαιδεύτηκε σε έναν ασθενή, δεδομένο σενάριο γεύματος και διάνυσμα καταστάσης που περιελάμβανε την προηγούμενη δόση ινσουλίνης. Ο δεύτερος ήταν γενικού σκοπού, ο τρίτος εκπαιδεύτηκε σε ασθενή με τυχαία γεύματα και οι δύο είχαν διάνυσμα καταστάσεων που περιελάμβανε τη χρονική στιγμή. Η αξιολόγηση έδειξε ότι ο πρώτος κατάφερε καλύτερο γλυκαιμικό έλεγχο σε οποιοδήποτε σενάριο και ότι κάνενας πράκτορας δεν πέτυχε πλήρη έλεγχο λόγω της μερικής παρατηρησιμότητας του περιβάλλοντος, τα χαρακτηριστικά φαρμακοκινητικής της ινσουλίνης, της ακρίβειας των αισθητήρων και του περιορισμένου συνόλου ενεργειών. Συνολικά η διπλωματική αυτή δείχνει ότι η ενισχυτική μάθηση μπορεί να διαχειριστεί το πρόβλημα του διαβήτη τύπου ένα αλλά χρειάζεται περαιτέρω πειραματισμός στο είδος του αλγορίθμου εκπαίδευσης και των παραμέτρων του.

Λέξεις κλειδιά: *τεχνητό πάγκρεας, διαβήτης τύπου 1, βαθιά ενισχυτική μάθηση, προσομοιωτής UVA/Padova 2008, DQN*

(Η σελίδα αυτή είναι σκόπιμα λευκή.)

Abstract

Type 1 diabetes is an autoimmune disease that attacks pancreatic beta cells, responsible for insulin production and glycemic control. Diabetics require external insulin delivery with shots or an insulin pump to self-regulate their glucose levels. The creation of a closed loop glucose control system, also known as an artificial pancreas, from the combination of a continuous glucose monitor, an insulin pump and a control algorithm will achieve better glycemic control and improve their lifestyle. The algorithm is the most important component for a fully automatic insulin delivery system something that does not yet exist due to the existence of nonlinear parameters in the glucose-insulin model. By using the DQN algorithm from reinforcement learning we trained three agents with the same action set of three insulin doses and a reward function based on blood glucose risk index to act on the environment of the T1D UVA/PADOVA simulator 2008 implemented in python3. The first agent was trained for one patient, a fixed meal scenario and a state vector containing previous insulin dose. The second one was trained for all patients and random meals, the third one was trained for one patient and random meals and both with state vectors containing time. Evaluating them concluded that the first one achieved better glucemic control in any meal scenario and that none of them controlled the system successfully due to the environment's partial observabillity, insulin pharmacokinetics, sensor noise and reduced action set. In conclution reinforcement learning can be used to solve the type one diabetes but further experiments in control algorithm and parameters are required.

Keywords: *artificial pancreas, Type 1 Diabetes, Deep Reinforcement Learning, UVA/Padova simulator 2008, DQN*

Ευχαριστώ τους κοντινούς μου ανθρώπους για την υπομονή και την υποστήριξη τους.

Ευχαριστώ τον καθηγητή μου Ανδρέα Σταφυλοπάτη χάρη στον οποίο μου δώθηκε η ευκαρία να γνωρίσω τον τομέα της μηχανικής μάθησης και να τη συνδυάσω με τον τομέα της βιοϊατρικής για την εκπόνηση αυτής της διπλωματικής.

Ευχαριστώ ακόμη τον υποψήφιο διδάκτορα Αθανάσιο Τασάκο για την καθοδήγηση και βοήθεια στην πορεία εξέλιξης αυτής της διπλωματικής.

Περιεχόμενα

Ευρετήριο εικόνων

Ευρετήριο πινάκων

Μέρος Ι

Επισκόπηση - Βασικά Στοιχεία Μελέτης

Κεφάλαιο 1

Εισαγωγή

Σε αυτό το εισαγωγικό κεφάλαιο παρουσιάζεται η ασθένεια του διαβήτη τύπου 1 (Type 1 Diabetes (T1D)) και η σημασία του σωστού ελέγχου των επιπέδων της γλυκόζης στο αίμα (blood glucose - BG). Γίνεται αναφορά στην ανακάλυψη της ινσουλίνης και τον δρόμο που άνοιξε στην προσπάθεια αντιμετώπισης της ασθένειας. Στη συνέχεια περιγράφονται οι μεταβολικές διεργασίες του σώματος, η φυσιολογική λειτουργία παγκρέατος κατά τη διάρκεια τους και θεραπευτικά σχήματα που καλούνται να την αντικαταστήσουν στην περίπτωση του T1D. Παρουσιάζεται το τεχνητό πάγκρεας (Artificial Pancreas - AP), δηλαδή ένα σύστημα ελέγχου γλυκόζης κλειστού βρόχου, ως μία λύση σε αυτή την πρόκληση και η δυνατότητα ανάπτυξης του με χρήση αλγορίθμων ενισχυτικής μάθησης.

1.1 Διαβήτης τύπου 1

Ο διαβήτης τύπου 1 (T1D) είναι μία αυτοάνοση ασθενεία όπου το ανοσοποιητικό συστήμα καταστρέφει τα β-κύτταρα που βρίσκονται στις νησίδες Langerhans του παγκρέατος, τα οποία είναι υπεύθυνα για την παραγωγή ινσουλίνης. Η ινσουλίνη είναι μια πεπτιδική ορμόνη η οποία παίζει ρυθμιστικό ρόλο στο μεταβολισμό των τροφών προκαλώντας την απορρόφηση της γλυκόζης από τα λιποκύτταρα, το συκώτι και τους σκελετικούς μυς. Η γλυκόζη είναι η κυριότερη πηγή ενέργειας του σώματος.

Ο T1D εμφανίζεται συνήθως σε παιδιά και εφήβους αλλά μπορεί να ξεκινήσει σε οποιαδήποτε ηλικία. Το 5% των ανθρώπων με διαβήτη έχουν τον τύπο 1, ένα γεγονός που τον κάνει μία σπάνια ασθένεια. Τα τελευταία δημογραφικά στοιχεία δείχνουν ότι υπάρχει άνοδος στα περιστατικά T1D. [1]

Παρόλο που ο διαβήτης ήταν γνωστή ασθένεια ήδη από την αρχαία Αίγυπτο, η ινσουλίνη ανακαλύφθηκε στις αρχές του 20ου αιώνα από τους Banting και Best. Χωρίς την παρουσία της ινσουλίνης αυξάνεται επικίνδυνα η γλυκόζη στο αίμα αφού τα κύτταρα δεν μπορούν να την χρησιμοποιήσουν, προκαλώντας υπεργλυκαιμία. Η παρατεταμένη παραμονή σε κατάσταση υπεργλυκαιμίας έχει μακροπρόθεσμα καταστροφικές συνέπειες για τον οργανισμό όπως καταστροφή νευρώνων και τριχοειδών αγγείων με αποτέλεσμα τη νέκρωση περιοχών του σώματος που δεν γίνεται καλή

αιμάτωση.

Δεν υπάρχει οριστική θεραπεία και γι'αυτό είναι απαραίτητη η λήψη ινσουλίνης εξωτερικά, με ενέσεις ή μέσω αντλίας ινσουλίνης. Οι διαβητικοί με τη βοήθεια ειδικών μαθαίνουν να χειρίζονται μόνοι τους τη γλυκόζη τους σαν κομμάτι της καθημερινότητά τους. Πρέπει να ελέγχουν τακτικά μέσα στη μέρα τα επίπεδα της με κάποια συσκευή μέτρησης και να υπολογίζουν τις δόσεις ινσουλίνης που χρειάζονται ώστε να διατηρούν τη συγκέντρωσή της σε φυσιολογικά επίπεδα.

Η εντατική θεραπεία ινσουλίνης μπορεί να προκαλέσει υπογλυκαιμία, δηλαδή να ρίξει τη γλυκόζη κάτω από τα φυσιολογικά όρια. Αν η υπογλυκαιμία δεν γίνει άμεσα αντιληπτή ο ασθενής μπορεί να εισέλθει σε κόμα. Κίνδυνος υπογλυκαιμίας υπάρχει και κατά τη διάρκεια του ύπνου αφού αποτελεί παρατεταμένη περίοδο νηστείας. Βλέπουμε λοιπόν ότι η σωστή διαχείριση της γλυκόζης είναι ένα πολύ σημαντικό ζήτημα.

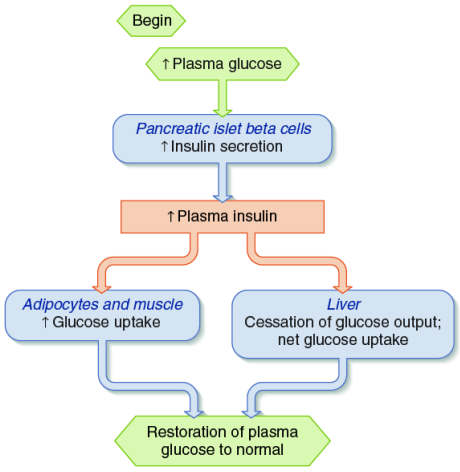
1.2 Ο κύκλος γεύματος-ινσουλίνης

Μετά την κατανάλωση κάθε γεύματος ακολουθεί ένας κύκλος βιοχημικών λειτουργιών για την απορρόφηση και αποθήκευση των θρεπτικών συστατικών που περιέχει όπως περιγράφεται στο (Vander [2]).

Σε ένα γεύμα διακρίνουμε 3 βασικά συστατικά: υδατάνθρακες, πρωτεΐνες και λίπη. Οι υδατάνθρακες κατά την απορρόφηση τους μετατρέπονται σε γλυκόζη και γι'αυτό παρέχουν το μεγαλύτερο μέρος της ενέργειας που χρησιμοποιεί το σώμα. Η γλυκόζη που λαμβάνεται κατά τη διάρκεια ενός γεύματος χρησιμοποιείται άμεσα σε ένα ποσοστό για να καλύψει τις ανάγκες του σώματος και η υπόλοιπη αποθηκεύεται από το συκώτι και τους σκελετικούς μύες σε μορφή γλυκογόνου (μακρομόρια γλυκόζης) και από τα λιπώδη κύτταρα ως λίπος.

Κατά τη λήψη ενός γεύματος τα θρεπτικά του συστατικά εισέρχονται από το πεπτικό σύστημα στο αίμα πράγμα που προκαλεί την αύξηση της συγκέντρωσης της γλυκόζης στο πλάσμα. Η αύξηση αυτή προκαλεί την έκκριση ινσουλίνης από τα β-κύτταρα η οποία ενεργοποιεί την απορρόφηση της γλυκόζης από τα κύτταρα. Αυτό τελικά μειώνει τη συγκέντρωση της γλυκόζης στο αίμα και έτσι το πάγκρεας μειώνει και την παραγωγή ινσουλίνης.

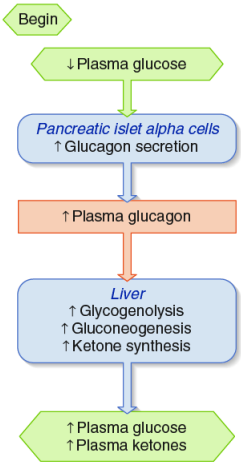
Η διαδικασία που περιγράφηκε ονομάζεται απορροφητική περίοδος. Στο διάγραμμα ροής της εικόνας 1.1 φαίνεται ο έλεγχος της ινσουλίνης από την αύξηση της συγκέντρωσης της γλυκόζης.



Εικόνα 1.1: Ο έλεγχος της έκκρισης ινσουλίνης από την αύξηση της συγκέντρωσης γλυκόζης στο πλάσμα. [Πηγή, Vander's Human Physiology, The Mechanisms of Body Function, 10th Edition, chapter 16, pg 623]

Όταν δεν υπάρχει φαγητό στο πεπτικό σύστημα το σώμα χρησιμοποιεί τις αποθήκες του για να καλύψει τις ενεργειακές του ανάγκες. Το πάγκρεας σε απόκριση της μείωσης της γλυκόζης ξεκινά τη έκκριση γλυκαγόνης από τα άλφα κύτταρα (Α-κύτταρα) των νησίδων Langerhans η οποία αναγκάζει το συκώτι να ελευθερώσει γλυκόζη στο αίμα σπάζοντας το γλυκαγόνο που έχει αποθηκευμένο. Αυτή η γλυκόζη καταναλώνεται χάρη στην παρουσία της ινσουλίνης της οποίας η συγκέντρωση είναι σημαντικά λιγότερη από ότι κατά την απορροφητική περίοδο και ονομάζεται ινσουλίνη υποβάθρου.

Η διαδικασία αυτή ονομάζεται μεταπορροφητική περίοδος. Στο διάγραμμα ροής της εικόνας 1.2 φαίνεται ο έλεγχος της γλυκαγόνης από τη μείωση της συγκέντρωσης της γλυκόζης.



Εικόνα 1.2: Ο έλεγχος της έκκρισης γλυκαγόνη από τη μείωση της συγκέντρωσης γλυκόζης στο πλάσμα. [Πηγή, Vander's Human Physiology, The Mechanisms of Body Function, 10th Edition, chapter 16, pg 624]

Ο πλήρης μεταβολισμός ενός γεύματος χρειάζεται περίπου 4 ώρες. Αν θεωρήσουμε ότι ο μέσος άνθρωπος καταναλώνει 3 γεύματα την ημέρα και εώς 3 ενδιάμεσα σνακ τότε το σώμα χρησιμοποιεί τις αποθήκες του συνήθως κατά τη διάρκεια του ύπνου και σε περιόδους δίαιτας και νηστείας. Η

διατήρηση της συγκέντρωση της γλυκόζης εντός φυσιολογικών ορίων καθ'όλη τη διάρκεια της ημέρας διασφαλίζει συνεχές απόθεμα ενέργειας και την ομαλή λειτουργία του οργανισμού.

1.3 Φαρμακευτική ινσουλίνη και θεραπείες

Η πρώτη ινσουλίνη που χορηγήθηκε σε ανθρώπους προερχόταν από ζώα. Η εξελίξεις που έγιναν τον 20ο αιώνα στους τομείς της ιατρικής, της βιολογίας και της τεχνολογίας οδήγησαν στην επίτευξη της σύνθεσης «ανθρώπινης» ινσουλίνης από βακτήρια, χρησιμοποιώντας την τεχνική ανασυνδυασμένου DNA. [3]

Παρακάτω αναφέρονται οι τέσσερις βασικοί τύποι συνθετικής ινσουλίνης.

- **Γρήγορης δράσης:** Έναρξη 15 λεπτά μετά τη λήψη, γίνεται μέγιστη μετά από 1 ώρα περίπου και συνέχιζει να δρα για 2 με 4 ώρες.
- **Κανονική ή σύντομης δράσης:** Έναρξη 30 λεπτά μετά τη λήψη, γίνεται μέγιστη μετά από 2-3 ώρες και έχει συνολική δράση για 3 με 6 ώρες.
- **Μέσης δράσης:** Έναρξη 2 με 4 ώρες μετά τη λήψη, γίνεται μέγιστη σε 4 με 12 ώρες και έχει συνολική δράση για 12 με 18 ώρες.
- **Μακράς δράσης:** Εμφανίζεται στο αίμα μέσα σε κάποιες ώρες από τη λήψη και έχει συνολική δράση εώς και 24 ώρες.

Οι θεραπείες του T1D περιλαμβάνουν συνδυασμούς των διαφόρων τύπων ινσουλίνης, ανάλογα με τον χρόνο δράσης τους, με στόχο να αντικαταστήσουν τη λειτουργία που θα έκανε φυσιολογικά το πάγκρεας. Το ποσό ινσουλίνης που χρειάζονται οι διαβητικοί εξαρτάται άμεσα από τα γεύματα, δηλαδή τις ποσότητες που καταναλώνουν. Η γλυκόζη όμως επηρεάζεται και από άλλους παράγοντες όπως είναι οι ασθένειες, η άθληση, το άγχος. Επίσης κάθε άνθρωπος εμφανίζει διαφορετική ευαισθησία στην απορρόφηση της ινσουλίνης. Όλα αυτά πρέπει να λαμβάνονται υπόψιν στην αναζήτηση μιας θεραπείας.

Για κάθε άνθρωπο κατασκευάζονται διαφορετικά πλάνα δόσεων ινσουλίνης ανάλογα με τις ανάγκες του. Ένα συνηθισμένο μοντέλο δόσεων το οποίο προσεγγίζει περισσότερο τη φυσιολογική λειτουργία του παγκρέατος είναι το Basal-Bolus το οποίο περιλαμβάνει πολλές ενέσεις ινσουλίνης κάθε ημέρα.

Χορηγείται μια δόση ινσουλίνης μακράς διάρκειας (basal ινσουλίνη) η οποία έχει το ρόλο της ινσουλίνης υποβάθρου. Λαμβάνονται επιπλέον δόσεις γρήγορης δράσης (bolus ινσουλίνη) πριν ή κατά τη διάρκεια των γευμάτων, σε ποσότητα η οποία υπολογίζεται με βάση το χρόνο δράσης της ινσουλίνης, τους υδατάνθρακες, τη γλυκόζη πριν το γεύμα και την ινσουλίνη που είναι ήδη ενεργή. Αυτός ο τρόπος διαχείρισης επιτρέπει κάποια ευελιξία στον ασθενή γιατί του δίνει ελευθερία στα γεύματα ως προς την ποσότητα των υδατανθρακών και την ώρα που θα πάρει μια επιπλέον δόση.

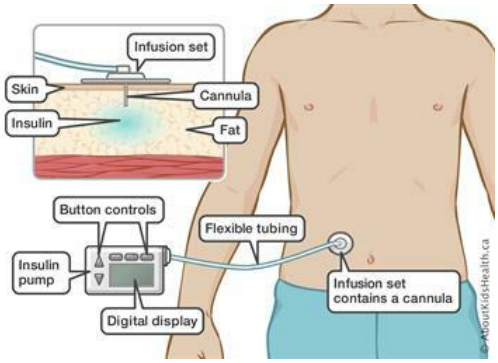
Πολλοί άνθρωποι όμως νιώθουν άβολα με τις ενέσεις και κυρίως όταν χρειάζεται να κάνουν μία σε εξωτερικό χώρο. Γι' αυτό το λόγο κατασκευάστηκαν αντλίες ινσουλίνης ώστε να προσφέρουν μεγαλύτερη ευελιξία στην καθημερινότητα. Επίσης για να αντικατασταθεί η μέτρηση της

γλυκόζης με αίμα από το δάχτυλο, πράγμα το οποίο σημαίνει πολλά τρυπήματα μέσα στη μέρα για τους διαβητικούς κατασκευάστηκαν αισθήτηρες συνεχούς μέτρησης

1.3.1 Αντλία ινσουλίνης

Μία αντλία ινσουλίνης είναι μία συσκευή συνεχούς έγχυσης ινσουλίνης η οποία χρησιμοποιείται εναλλακτικά από τις ενέσεις ινσουλίνης. [4] Είναι μια μικρή, προγραμματιζόμενη εξωτερική συσκευή η οποία μιμείται τη λειτουργία του παγκρέατος εγχύοντας συνεχώς μικρές δόσεις ινσουλίνης γρήγορης δράσης (basal ρυθμός). Δίνει τη δυνατότητα λήψης μεγαλύτερης ποσότητας ινσουλίνης για να καλύψει την αυξημένη ανάγκη στην περίπτωση ενός γεύματος (bolus).

Αποτελείτε από έναν καθετήρα ο οποίος εισέρχεται στον υποδόριο ιστό του ασθενή, συνήθως στην περιοχή κοντά στο στομάχι, ένα ρεζερβουάρ ινσουλίνης και ένα σύστημα έγχυσης. Μπορεί να τοποθετηθεί σε ζώνη γύρω από τη μέση του ασθενή, σε τσέπη ή ειδικό βραχιόλι. Αυτό την κάνει πρακτική στη μεταφορά και διακριτική. Τα μέρη της αντλίας φαίνονται στην εικόνα 1.3



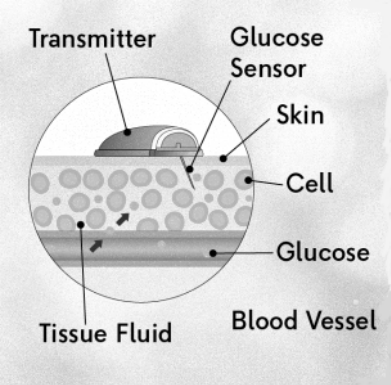
Εικόνα 1.3: Τα βασικά συστατικά και η σύνδεση μιας αντλίας ινσουλίνης

Ο basal ρυθμός καθορίζεται με τη βοήθεια του γιατρού, ενώ η bolus δόση υπολογίζεται πριν από κάθε γεύμα. Οι αντλίες επιτρέπουν την ύπαρξη πολλών διαφορετικών ρυθμών basal για τον ίδιο ασθενή, ανάλογα τις ανάγκες μέσα στη μέρα, αλλά τυπικά ένας ασθενής χρειάζεται 4 με 6 διαφορετικούς ρυθμούς. Η αντία διαθέτει διεπαφή για τον καθορισμό του basal ρυθμού και τον υπολογισμό της bolus δόσης.

1.3.2 Συνεχής μέτρηση γλυκόζης

Η παρακολούθησης της γλυκόζης έγινε πιο εύκολη χάρη στην ανάπτυξη συσκευών συνεχούς μέτρησης (Continuous Glucose Monitor - CGM). Αυτές είναι μικρές συσκευές οι οποίες τοποθετούνται κάτω από το δέρμα (εικόνα 1.4), μετράνε τη συγκέντρωση της γλυκόζης στον υποδόριο ιστό με συχνότητα 1-5 λεπτά και υπολογίζουν με έναν αλγόριθμο την αντίστοιχη στο αίμα.

Συνδέονται ασύρματα με έναν δέκτη για τακτικό έλεγχο των επιπέδων γλυκόζης. Καταγράφουν τις τάσεις της γλυκόζης (ρυθμό και κατεύθυνση) και διαθέτουν σύστημα ειδοποιήσης όταν προβλέπουν κίνδυνο υπογλυκαιμίας ή υπεργλυκαιμίας. Σε αντίθεση με άλλους τρόπους μέτρησης που παίρνουν λίγα δείγματα μέσα στη μέρα, ένα CGM παρέχει σχεδόν συνεχείς μετρήσεις σε πραγματικό χρόνο. [5]



Εικόνα 1.4: Η σύνδεση ενός αισθητήρα συνεχούς μέτρησης γλυκόζης στον υποδόριο χώρο.

Οι συσκευές CGM χρειάζονται μία-δύο φορές την ημέρα ρύθμιση με βάση την πραγματική μέτρηση γλυκόζης με αίμα από το δάχτυλο, για να μπορούν να κάνουν ακριβείς μετρήσεις στη συνέχεια. Επειδή οι μετρήσεις που κάνουν γίνονται στον υποδόριο ιστό και όχι απευθείας στο πλάσμα του αίματος υπάρχει πάντα ένα ποσοστό θορύβου στη μέτρηση σε σχέση με την πραγματική τιμή. Αυτό συμβαίνει διότι υπάρχει καθυστέρηση στη συγκέντρωση της γλυκόζης μεταξύ των δύο χώρων. Αυτό σε συνδυασμό με τυχαίο θόρυβο που εισέρχεται στις μετρήσεις προκαλεί απόκλιση από την πραγματική τιμή. Όμως η ακρίβεια των αισθητήρων CGM αυξάνεται με την πρόοδο της τεχνολογίας.

1.4 Τεχνητό πάγκρεας

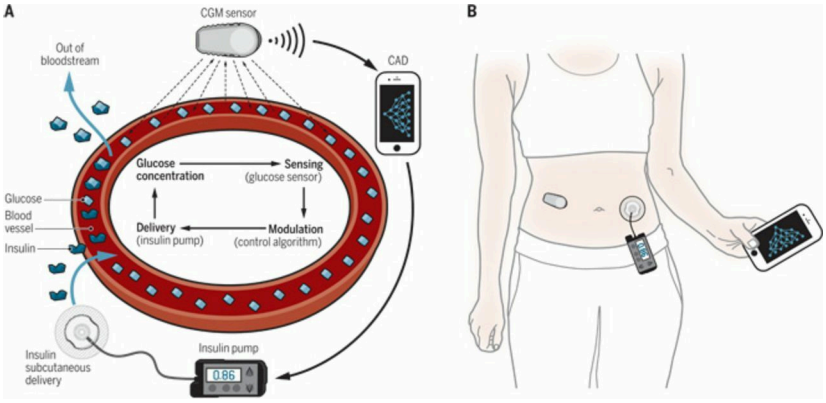
Υπήρξαν προσπάθειες για μεταμοσχεύσεις παγκρέατος και νησίδων Langerhans αλλά χωρίς θετικά αποτελέσματα. Η προσοχή των επιστημόνων στράφηκε στην κατασκευή μιας συσκευής που συνεχώς θα ελέγχει τη συγκέντρωση της γλυκόζης και αυτόματα θα προσαρμόζει την ινσουλίνη που παρέχει, δηλαδή ένα τεχνητό πάγκρεας. [6]

Οι προσπάθειες ανάπτυξης ενός τέτοιου συστήματος ξεκίνησαν από τη δεκαετία του 70 αλλά οι συσκευές που το αποτελούσαν ήταν ογκόδεις και μη πρακτικές. Η εξέλιξη της τεχνολογίας και οι νέες αντλίες και συσκευές συνεχούς ελέγχου γλυκόζης, έχουν κάνει τη δημιουργία τεχνητού παγκρέατος πλέον ένα ρεαλιστικό στόχο ο οποίος θα βελτιώσει την ποιότητα ζωής των ασθενών. Στόχος αυτής της τεχνολογίας είναι να ελαφρύνει το βάρος της διαχείρισης της γλυκόζης και να παρέχει εξατομικευμένη θεραπεία μειώνοντας τον κίνδυνο υπογλυκαιμίας και υπεργλυκαιμίας.

Το τεχνητό πάγκρεας είναι ένα σύστημα ελέγχου κλειστού βρόχου το οποίο μιμείται τη φυσιολογική λειτουργία ενός υγιούς παγκρέατος. Αποτελείται από το συνδυασμό ενός CGM, μίας αντλίας ινσουλίνης και ενός αλγόριθμου ελέγχου. Ο τελευταίος μπορεί να είναι ενσωματωμένος στην αντλία ή να βρίσκεται σε κάποια εξωτερική συσκευή. Οι συσκευές επικοινωνούν ασύρματα μεταξύ τους.

Ο αλγόριθμος ελέγχου συνδέει την αντλία με το CGM και είναι το πιο σημαντικό συστατικό στο τεχνητό πάγκρεας. Παρακολουθεί τις διακυμάνσεις της γλυκόζης από τις μετρήσεις του CGM και προσαρμόζει σε πραγματικό χρόνο το ρυθμό έγχυσης της αντλίας. Επίσης θα πρέπει να είναι

προσαρμοσμένος στα ιδιαίτερα χαρακτηριστικά του εκάστοτε ασθενή. Στην εικόνα 1.5 βλέπουμε ένα κλειστό σύστημα ελέγχου γλυκόζης όπου ο αλγόριθμος ελέγχου βρίσκεται σε μία άλλη συσκευή διαφορετική από την αντλία.



Εικόνα 1.5: Σύστημα τεχνητού παγκρέατος κλειστού βρόχου με ασύρματη επικοινωνία. [Πηγή, A. KITTERMAN/SCIENCE TRANSLATIONAL MEDICINE]

1.5 Αλγόριθμοι ελέγχου

Οι αλγόριθμοι οι οποίοι αναφέρονται στη βιβλιογραφία [7] και έχουν δοκιμαστεί πιο πολύ στην έρευνα είναι ο αλγόριθμος Model-Predictive Control (MPC) και Proportional-Integral-Derivative (PID).

- **MPC:** Ένας αλγόριθμος ελέγχου ο οποίος λειτουργεί σε δυναμικά μοντέλα διεργασιών, εφαρμόζει βελτιστοποίηση χρησιμοποιώντας συναρτήσεις κόστους σε πεπερασμένα χρονικά διαστήματα της διεργασίας και περιλαμβάνει ιστορία παρελθοντικών ενεργειών. Αυτό τον κάνει έναν αλγόριθμο ο οποίος αναζητεί την βέλτιστη έγχυση ινσουλίνης με βάση προβλέψεις που κάνει για το μέλλον. Πιο συγκεκριμένα θεωρεί όλες τις σειρές ενεργειών ελέγχου που μπορεί να εφαρμόσει στο μέλλον και επιλέγει αυτή που προβλέπεται από τη συνάρτηση κόστους ότι μπορεί να επιφέρει καλύτερα αποτελέσματα, δηλαδή να κάνει καλύτερο γλυκαιμικό έλεγχο. Αυτή τη διαδικασία την κάνει σε κάθε βήμα. Η πρόβλεψη γίνεται με χρήση του μοντέλου ή με βάση την ιστορία.

- **PID:** Ένας αλγόριθμος από τα συστήματα αυτομάτου ελέγχου ο οποίος χρησιμοποιείται σε συστήματα που απαιτούν συνεχή έλεγχο. Προσπαθεί να ρυθμίσει ένα σύστημα ώστε να κινείται γύρω από έναν σταθερό στόχο, υπολογίζοντας συνεχώς το σφάλμα από τον στόχο (proportional), το συνολικό σφάλμα για να το μηδενίσει στη μόνιμη κατάσταση (integral) και το ρυθμό αλλαγής της μεταβλητής ελέγχου (derivative). Έτσι ανάλογα με τις αλλαγές που παρατηρεί στην συγκέντρωση της γλυκόζης αλλάζει τον ρυθμό έγχυσης της ινσουλίνης. Το integral κομμάτι μπορεί να θεωρηθεί ως αργή μεταβολή στη basal ινσουλίνη ενώ τα proportional και derivative οι αποκρίσεις σε γεύματα. Για να λάβει υπόψιν τους την καθυστέρηση δράσης της ινσουλίνης ο αλγόριθμος προσαρμόστηκε με μία ανάδραση ινσουλίνης.

Και οι δύο αλγόριθμοι ελέγχουν αποτελεσματικά τη γλυκόζη, μειώνοντας τους κινδύνους υπογλυκαιμίας. Ένα αρνητικό του PID σε σχέση με τον MPC είναι ότι η λειτουργία του βασίζεται στις αλλαγές της γλυκόζης κάθε στιγμή ενώ ο MPC αν ρυθμιστεί σωστά επιτρέπει την πρόβλεψη των δυναμικών του μοντέλου της γλυκόζης. Επίσης ο MPC μπορεί να λάβει υπόψιν του παραμέτρους του μοντέλου προσαρμοσμένες στον ασθενή και να παρέχει εξειδικευμένο έλεγχο.

1.5.1 Εμπορικές συσκευές

Μέχρι σήμερα υπάρχει μόνο μία εμπορική συσκευή που η λειτουργία της είναι κοντά σε αυτό που ονομάζουμε τεχνητό πάγκρεας και έχει εγκριθεί από τον Αμερικάνικο Οργανισμό Τροφίμων και Φαρμάκων (U.S. Food and Drug Administration - FDA), η Minimed 670G της Medtronic. Αυτή η συσκευή είναι ένα υβριδικό σύστημα ελέγχου, με τον αλγόριθμο ενσωματωμένο στην αντλία, το οποίο ελέγχει τη γλυκόζη και προσαρμόζει αυτόματα τον basal ρυθμό. Επίσης διακόπτει την έγχυση ινσουλίνης όταν η γλυκόζη πέσει κάτω από κάποιο προκαθορισμένο όριο και την αυξάνει όταν παρατηρήσει αύξηση της γλυκόζης χωρίς αναγγελία γεύματος.

Το πλήρως αυτοματοποιημένο τεχνητό πάγκρεας δεν έχει ακόμα δημιουργηθεί και γίνεται πολύ έρευνα πάνω σε αυτό το θέμα.

1.6 Αντικείμενο και δομή της διπλωματικής

Το πρόβλημα του T1D είναι ένα πρόβλημα βελτιστοποίησης όπου πρέπει να γίνει αυστηρός γλυκαιμικός έλεγχος, δηλαδή διατήρηση της γλυκόζης εντός φυσιολογικών ορίων, μέσω έγχυσης ινσουλίνης χωρίς την αύξηση του κινδύνου υπογλυκαιμίας. Η βέλτιστη διαχείριση του διαβήτη τύπου 1 είναι στην πραγματικότητα ένας συνδυασμός σύγχρονων τεχνολογιών οι οποίες στοχεύουν να προσεγγίσουν τον κάθε άνθρωπο μεμονωμένα λαμβάνοντας υπόψιν τους τα μοναδικά του φυσιολογικά χαρακτηριστικά και συμπεριφορές. Το ιδανικό τεχνητό πάγκρεας θα πρέπει να προσομειώνει τη λειτουργία του παγκρέατος ενός υγιή ανθρώπου και γι'αυτό ο αλγόριθμος ελέγχου παίζει σημαντικό ρόλο. Οι αλγόριθμοι που έχουν δοκιμαστεί μέχρι τώρα αποτυγχάνουν σε ένα βαθμό λόγω της αδυναμίας τους πρόβλεψης γεύματος και της καθυστερημένης δράσης της φαρμακευτικής ινσουλίνης. Υπάρχει ανάγκη κατασκευής ενός αυτόνομου συστήματος με την ελάχιστη εμπλοκή του ασθενή.

1.6.1 Αντικείμενο

Αντικείμενο αυτής της εργασίας είναι η προσέγγιση του προβλήματος του διαβήτη τύπου 1 με μεθόδους Ενισχυτικής Μάθησης (Reinforcement Learning (RL)). Τα συστήματα ελέγχου που υπάρχουν δεν είναι πλήρως αυτοματοποιημένα αφού οι διαβητικοί χρειάζεται να μετράνε τους υδατάνθρακες που καταναλώνουν και να ενημερώνουν την αντλία τους.

Ο έλεγχος της γλυκόζης στην περίπτωση του διαβήτη τύπου 1 είναι ένα δύσκολο πρόβλημα από τη φύση του. Αυτό συμβαίνει γιατί το σύστημα κινητικότητας της γλυκόζης δεν είναι γραμμικό και

επηρεάζεται από εξωτερικούς και άγνωστους παράγοντες όπως το είδος του φαγητού, η άθληση, το άγχος, οι ορμόνες. Το να μοντελοποιηθούν όλα αυτά σε ένα σύστημα είναι πολύ δύσκολο και γι'αυτό η πρόβλεψη πάνω σε αυτά δεν είναι ακριβής. Η ενισχυτική μάθηση όμως δεν χρειάζεται ένα ακριβές μοντέλο του περιβάλλοντος της γιατί το ανακαλύπτει μέσω της αλληλεπίδρασης με αυτό, τη συλλογή εμπειριών από τις πράξεις που έκανε, τις καταστάσεις που βρέθηκε και την ανταμοιβή που μάζεψε.

Οι τελευταίες εξελίξεις στην ενισχυτική μάθηση οδήγησαν στη δημιουργία αλγορίθμων γενικού σκοπού που μπορούν να αποκτήσουν ένα ευρύ φάσμα δυνατοτήτων μέσα από εμπειρίες. Η ικανότητα των πρακτόρων ενισχυτικής μάθησης να δρουν σε αβέβαια περιβάλλοντα, όπου δεν υπάρχει πλήρες μοντέλο, αλλά να παίρνουν αποφάσεις με βάση τις παρατηρήσεις και να κάνουν προβλέψεις για άγνωστες καταστάσεις τους κάνει ιδανικούς υποψήφιους για την υλοποίηση τεχνητού παγκρέατος.

Η ύπαρξη προσομοιωτή των δυναμικών γλυκόζης-ινσουλίνης κάνει την έρευνα για την ανάπτυξη τεχνητού παγκρέατος πιο εύκολη από ποτέ. Αυτό επιχειρεί να κάνει και η εργασία αυτή με χρήση του αλγορίθμου Deep Q-Network (DQN). Το κλείσιμο του βρόχου είναι μία πρόκληση στην επίτευξη της οποίας πιστεύουμε ότι η μηχανική μάθηση μπορεί να συνεισφέρει και αυτό επιχειρούμε να δείξουμε.

1.6.2 Δομή

Στο μέρος II υπάρχει για λόγους πληρότητας η θεωρία της ενισχυτικής μάθησης και οι αλγόριθμοι Q-learning και Deep Q-learning. Στη συνέχεια στο μέρος III υπάρχει περιγραφή του προσομοιωτή του μοντέλου των δυναμικών μεταξύ γλυκόζης-ινσουλίνης, ο οποίος έχει εγκριθεί για να αντικαταστήσει τα πειράματα σε ζώα, στην αναζήτηση νέων θεραπειών. Επίσης εκεί περιγράφονται το αυστηρό πρόβλημα ενισχυτικής μάθησης που διαχειριστήκαμε, οι λεπτομέρειες υλοποιήσης, τα εργαλεία εκπαίδευσης και αξιολόγησης. Στο μέρος IV παραθέτουμε τα αποτελέσματα των πειραμάτων και σχολιάζουμε τα ευρήματα μας και τέλος στο μέρος V συνοψίζουμε την πορεία της διπλωματικής, τα αποτελεσματά της και αναφέρουμε μελλοντικές προοπτικές.

Μέρος ΙΙ

Θεωρητικό Υπόβαθρο

Κεφάλαιο 2

Ενισχυτική Μάθηση

Η ενισχυτική μάθηση (Reinforcement Learning - RL), όπως περιγράφεται στο βιβλίο των Sutton και Barto [8], είναι μία μέθοδος Μηχανικής Μάθησης στην οποία ο αλγόριθμος μαθαίνει να κάνει βέλτιστες ενέργειες μέσω αλληλεπίδρασης με τον κόσμο στον οποίο δρα και τη βοήθεια ενός συστήματος ανταμοιβής. Είναι εμπνευσμένη από τον τομέα της συμπεριφοριστικής ψυχολογίας η οποία περιγράφει τον ανθρώπινο τρόπο μάθησης ως συλλογή εμπειριών μέσα από αλληλεπίδραση. Οι εμπειρίες αυτές αξιολογούνται από το εσωτερικό σύστημα επιβράβευσης του εγκεφάλου και αποθηκεύονται στο βιολογικό νευρωνικό του σύστημα. Αν ήταν θετικές επιβραβεύονται με την έκκριση ντοπαμίνης (αίσθημα ικανοποίησης) διαφορετικά υπάρχει κάποιο είδος τιμωρίας π.χ. πόνος. Όποτε βρεθεί σε παρόμοιες κατάστασεις χρησιμοποιεί την εμπειρία του για να αποφασίσει τι να κάνει. Αντίστοιχα ένας αλγόριθμος ενισχυτικής μάθησης ανακαλύπτει τι πρέπει να κάνει, δηλαδή πως να αντιστοιχίσει κατάστασεις σε ενέργειες, με χρήση της εμπειρίας του και στόχο την απόκτηση της μέγιστης δυνατής ανταμοιβής.

2.1 Θεωρία της ενισχυτικής μάθησης

2.1.1 Χαρακτηριστικά

Ένα βασικό χαρακτηριστικό της ενισχυτικής μάθησης είναι η απουσία ενός παντογνώστη επιβλέποντα που να υποδυκνύει ποιες είναι οι σωστές ενέργειες. Στη θέση αυτού υπάρχει το σύστημα ανταμοιβής. Οι ενέργειες που γίνονται είναι στην κατεύθυνση της μεγιστοποίησης της συνολικής αριθμητικής ανταμοιβής. Παρόλο που θυμίζει μάθηση χωρίς επίβλεψη δεν ανήκει ούτε σε αυτή την κατηγορία. Στη μάθηση χωρίς επίβλεψη ο αλγόριθμος προσπαθεί να βρει κρυφές δομές σε ένα σύνολο δεδομένων. Κάτι τέτοιο δεν είναι στόχος της ενισχυτικής μάθησης και γι'αυτό θεωρείται ξεχωριστή κατήγορία.

Ένα άλλο σημαντικό χαρακτηριστικό είναι η πιθανόν καθυστερημένη ανταμοιβή. Αυτό σημαίνει ότι μία ενέργεια δεν έχει μόνο άμεση επίδραση στην παρούσα κατάσταση του περιβάλλοντος αλλά μπορεί να έχει μακροπρόθεσμες συνέπειες σε μελλοντικές καταστάσεις και τις ανταμοιβές που θα επιφέρουν αυτές. Μία πρόκληση της ενισχυτικής μάθησης είναι ο συμβιβασμός ανάμεσα

στην επιλογή της άμεσης ανταμοιβής για την μακροπρόθεσμη. Γενικότερα ο χρόνος είναι πολύ σημαντικός στο πρόβλημα της ενισχυτικής μάθησης. Η διαδικασία αποφάσεων είναι ακολουθιακή και το σύστημα είναι δυναμικό, δηλαδή κάθε ενέργεια θα αλλάξει τα δεδομένα που θα έρθουν στη συνέχεια.

Ένας πράκτορας πρέπει να συμβιβάσει την εξερεύνηση του περιβάλλοντος με την εκμετάλλευση της ήδη αποκτημένης γνώσης (exploration-exploitation). Πρέπει να προτιμά πράξεις που δοκιμάστηκαν στο παρελθόν και επέφεραν μεγάλη επιβράβευση αλλά για να ανακαλύψει τέτοιες πράξεις πρέπει να δοκιμάσει και καινούργιες. Άρα πρέπει να εκμεταλευτεί όσα ξέρει αλλά και να εξερευνήσει το περιβάλλον ώστε να πάρει καλύτερες αποφάσεις μελλοντικά. Αυτά τα δύο δεν μπορούν να γίνουν ξεχωριστά το ένα από το άλλο γιατί τότε είναι καταδικασμένα να αποτύχουν. Πρέπει να γίνονται συμβιβασμοί ανάμεσα στα δύο σε κάθε βήμα και καθώς περνάει ο χρόνος να προτιμάτε η τακτική που επιφέρει καλύτερα αποτελέσματα.

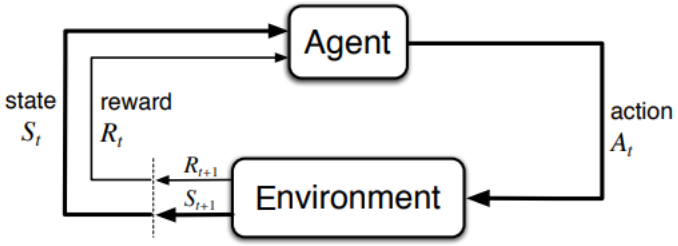
2.1.2 Βασικά στοιχεία ενισχυτικής μάθησης

Σε ένα πρόβλημα ενισχυτικής μάθησης η οντότητα που μαθαίνει και παίρνει αποφάσεις ονομάζεται *πράκτορας* (agent). Ο κόσμος με τον οποίο αλληλεπιδρά ο πράκτορας, δηλαδή οτιδήποτε είναι πέρα από τον έλεγχο του πράκτορα, ονομάζεται *περιβάλλον* (environment). Ο πράκτορας και το περιβάλλον αλληλεπιδρούν μεταξύ τους για διακριτό και πεπερασμένο αριθμό βημάτων $t = 0, 1, 2, \ldots$. Ο πράκτορας επιλέγει ενέργειες και το περιβάλλον απαντά με μία νέα κατάσταση και μία ανταμοιβή. Στόχος του πράκτορα είναι να διαλέξει ενέργειες ώστε να μεγιστοποιήσει τη συνολική ανταμοιβή που θα λάβει.

Σε κάθε βήμα t ο πράκτορας:

- πρέπει να επιλέξει μια ενέργεια A_t
- θα λάβει μία αριθμητική τιμή ανταμοιβής R_t
- θα λάβει μια παρατήρηση από το περιβάλλον O_t

Μια σειρά από τέτοια βήματα ονομάζεται *ιστορία*: $H_t = A_1 O_1 R_1 A_2 O_2 R_2 \ldots A_t O_t R_t$ και το τι θα συμβεί μετά εξαρτάται από αυτή. Στην εικόνα 2.1 φαίνεται ο βρόχος αλληλεπίδρασης του πράκτορα με το περιβάλλον. Η διαδικασία αυτή είναι μία Μαρκοβιανή διαδικασία απόφασης (Markov decision process - MDP).



Εικόνα 2.1: Η αλληλεπίδραση μεταξύ πράκτορα-περιβάλλοντος [Πηγή, Reinforcement Learning An Introduction, Sutton and Barto]

Η αλληλεπίδραση πράκτορα-περιβάλλοντος είναι μία συνεχής διαδικασία στην προσπάθεια επίτευξης του στόχου. Κάθε τέτοια αλληλεπίδραση λέγεται *επεισόδιο* (episode). Ενα επεισόδιο στη γενική περίπτωση έχει μια κατανομή από αρχικές καταστάσεις από όπου μπορεί να ξεκινήσει. Οι αποφάσεις όμως που παίρνει ο πράκτορας σε κάθε επεισόδιο τον οδηγουν σε διαφορετική ακολουθία καταστάσεων που επισκέπτεται και σε διαφορετική συνολική ανταμοιβή. Μετά από αρκετά επεισόδια ο πράκτορας μαθαίνει να κάνει καλύτερες επιλογές.

Η *ανταμοιβή* (reward) R_t είναι ένα αριθμητικό σήμα ανάδρασης, δηλαδή μια εκτίμηση του πόσο καλά τα πάει ο πράκτορας σε βήμα t. Σε κάθε πρόβλημα της ενισχυτικής μάθησης ισχύει η υπόθεση της ανταμοιβής: όλοι οι *στόχοι* μπορούν να περιγραφούν ως η μεγιστοποίηση της μέσης συσσωρευμένης ανταμοιβής.

Το άθροισμα των ανταμοιβών ξεκινώντας από βήμα t λέγεται *επιστροφή* (return) $G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + ... = R_{t+1} + \gamma G_{t+1}, 0 < \gamma < 1$.

Το γ ονομάζεται *συντελεστής μείωσης* (discount factor) και καθορίζει πόση αξία δίνουμε στην άμεση ανταμοιβή σε σχέση με την μακροπρόθεσμη. Αν $\gamma = 0$ τότε ο πράκτορας συμπεριφέρεται άπληστα και διαλέγει ενέργειες που έχουν άμεση βέλτιστη ανταμοιβή χωρίς να λαμβάνει υπόψιν του το μέλλον. Χρησιμοποιούμε τιμές $0 < \gamma < 1$ γιατί υπάρχει αβεβαιότητα στο μέλλον και επειδή θέλουμε να αποφύγουμε το άπειρο. Ανάλογα την τιμή του δείχνουμε την προτιμησή μας στις άμεσες ανταμοιβές.

Η *κατάσταση* (state) S_t είναι η πληροφορία που χρησιμοποιείται σε κάθε βήμα για να καθορίστει τι θα συμβεί μετά. Η κατάσταση του περιβάλλοντος S_t^e είναι η εσωτερική του αναπαράσταση του τι συμβαίνει και συνήθως είναι διαφορετική από του πράκτορα. Η κατάσταση του πράκτορα S_t^a καθορίζεται από τον προγραμματιστή και είναι η πληροφορία του περιβάλλοντος που μας ενδιαφέρει. Είναι ένα διάνυσμα με στοιχεία $S \subseteq O$ ανάλογα την παρατηρησιμότητα του περιβάλλοντος και τι θεωρούμε χρήσιμη πληροφορία.

Διακρίνουμε το περιβάλλον με βάση την παρατηρησιμότητα του και καθορίζουμε την κατάσταση σε κάθε περίπτωση:

- *Πλήρως Παρατηρησιμο* όπου ισχύει $O_t = S_t^a = S_t^e$. Ο πράκτορας γνωρίζει πλήρως την κατάσταση του περιβάλλοντος και η κατάσταση του είναι ίδια αυτού.
- *Μερικώς Παρατηρήσιμο* όπου $S_t^e \neq S_t^a$. Ο πράκτορας παρατηρεί έμμεσα το περιβάλλον και πρέπει μόνος του να κατασκευάσει τη δική του κατάσταση.

Η *πολιτική* (policy) καθορίζει τη συμπεριφορά του πράκτορα σε δεδομένη στιγμή, είναι δηλαδή μια απεικόνιση από καταστάσεις σε ενέργειες. Μια πολιτική μπορεί να είναι:

1. *Ντετερμινιστική*: πχ μια συνάρτηση $\alpha = \pi(s)$ ή ένας πίνακας αντιστοίχησης καταστάσεων σε ενέργειες.

2. *Στοχαστική*: πχ συνάρτηση πυκνότητας πιθανότητας $\pi(a|s) = \Pr(A_t = a|S_t = s)$

Η *συνάρτηση αξίας* (value function) $v(s)$ είναι μια πρόβλεψη της μελλοντικής ανταμοιβής και χρησιμοποιείται για την εκτίμηση των καταστάσεων ως καλές ή κακές και την επιλογή ενέργειων. Οι ανταμοιβές είναι άμεσες ενώ οι αξίες δείχνουν την μακροπρόθεσμη επίδραση της κατάστασης. Η αξία μιας κατάστασης είναι η μέση τιμή του συνόλου των ανταμοιβών που περιμένουμε να πάρουμε ξεκινώντας από τη δεδομένη κατάσταση όπως φαίνεται στην εξίσωση 2.1 :

$$v(s) = E[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + ...|S_t = s] = E[G_t|S_t = s], 0 < \gamma < 1 \tag{2.1}$$

όπου γ ο συντελεστής μείωσης.

Υπάρχει πάντα τουλάχιστον μία πολιτική η οποία είναι *βέλτιστη (optimal policy)* $\pi_\star$, γιατί είναι καλύτερη ή ίση με όλες τις άλλες. Μία πολιτική π είναι καλύτερη ή ίση με μία πολιτική π' αν και μόνο αν η μέση τιμή της αναμενόμενης ανταμοιβής τής είναι μεγαλύτερη ή ίση από της π' για κάθε κατάσταση, δηλαδή ανν $v_\pi(s) \geq v_{\pi'}(s)$. Η *βέλτιστη συνάρτηση αξίας* (optimal value function) είναι η συνάρτηση η οποία επιφέρει τη μέγιστη μέση επιστροφή από όλες τις πολιτικές:

$$v_\star(s) = \max_\pi v_\pi(s) \tag{2.2}$$

Ανάλογα ορίζονται η *συνάρτηση ενέργειας-αξίας* (action-value function) $q(s,a)$ (2.3) και η βέλτιστη συνάρτηση αξίας $q_\star(s,a)$ (2.4). Η πρώτη είναι μία συνάρτηση αξίας που εκφράζει το πόσο καλή είναι η ενέργεια a στην κατάσταση s. Η δεύτερη είναι η συνάρτηση η οποία επιφέρει τη μέγιστη μέση επιστροφή ακολουθώντας οποιαδήποτε πολιτική και κάνοντας ενέργεια a.

$$q(s,a) = E[G_t|S_t = s, A_t = a] \tag{2.3}$$

$$q_\star(s,a) = \max_\pi q_\pi(s,a) \tag{2.4}$$

Το *μοντέλο* (model) προβλέπει τι θα κάνει το περιβάλλον στη συνέχεια. Για παράδειγμα δεδομένων μιας κατάστασης και μιας ενέργειας ένα μοντέλο μπορεί να προβλέψει την ανταμοιβή και την επόμενη κατάσταση. Τα μοντέλα χρησιμοποιούνται στον σχεδιασμό επίλυσης του περιβάλλοντος και οι πράκτορες που τα χρησιμοποιούν ονομάζονται model-based.

Ένα μοντέλο δεν είναι πάντα δυνατό να υπάρχει και οι περισσότεροι πράκτορες λειτουργούν σε αβέβαια περιβάλλοντα. Η ενισχυτική μάθηση όπως έχει περιγραφεί μέχρι τώρα δεν χρειάζεται

απαραίτητα κάποιο μοντέλο αφού μπορεί να μάθει με αλληλεπίδραση ακολουθώντας την στρατηγική της δοκιμής και σφάλματος (trial and error).

2.1.3 Μαρκοβιανή διαδικασία απόφασης

Οι διακριτές *Μαρκοβιανές Διαδισίες Απόφασης (Markov Decision Process - MDP)* είναι το μαθηματικό πλαίσιο που χρησιμοποιείται για το φορμαλισμό των προβλημάτων ενισχυτικής μάθησης, δηλαδή τη διατύπωση εξισώσεων που τα περιγράφουν και την επίλυση τους. Προέρχονται από τη θεωρία των διακριτών δυναμικών συστημάτων και είναι ο κλασσικός τρόπος αναπαράστασης προβλημάτων βελτιστοποίησης που περιλαμβάνουν ακολουθιακή λήψη αποφάσεων. Περιέχουν την έννοια της κατάστασης, των ενεργειών, της ανταμοιβής με καθυστέρηση και του συμβιβασμού ανάμεσα στην άμεση και την μακροπρόθεσμη ανταμοιβή.

Είναι επέκταση των διακριτών *Μαρκοβιανών Αλυσίδών (Markov Chains)* οι οποίες είναι στοχαστικές ανελίξεις που χαρακτηρίζονται από καταστάσεις και πιθανότητες μετάβασης μεταξύ αυτών. Οι καταστάσεις σε μία μαρκοβιανή αλυσίδα πρέπει να τηρούν την *ιδιότητα Markov* (εξίσωση 2.5).

$$Pr(S_{t+1}|S_t) = Pr(S_{t+1}|S_1, S_2, ..., S_t) \tag{2.5}$$

Με λόγια η ιδιότητα Markov συνοψίζεται ως εξής:

«το μέλλον είναι ανεξάρτητο από το παρελθόν δεδομένου του παρόντος».

Στη θεωρία συστημάτων η ιδιότητα αυτή είναι η έλλειψη μνήμης. Δηλαδή μία διαδικασία Markov είναι μία τυχαία (στοχαστική) διαδικασία χωρίς μνήμη.

Μια διακριτή MDP είναι μία τούπλα <*S,A,P,R,γ*> όπου:

- S ένα διακριτό σύνολο καταστάσεων
- A ένα διακριτό σύνολο ενεργειών
- P ένας *πίνακας καταστάσεων-μεταβάσεων (state transition matrix)*, $P^a_{ss'} = Pr[S_{t+1} = s'|S_t = s], s, s' \in S$ που δείχνει πόσο πιθανή είναι η μετάβαση από την κατάσταση s στην s '
- R μια συνάρτηση ανταμοιβής, $R^a_s = E[R_{t+1}|S_t = s, A_t = a], s \in S, a \in A$
- γ ένας συντελεστής μείωσης $\gamma \in [0,1]$

Στην ενισχυτική μάθηση οι καταστάσεις του περιβάλλοντος S^e_t και η ιστορία $H_t = S_0A_0R_1S_1A_1R_2S_2A_2...$ είναι Markov γιατί η επόμενη κατάσταση εξαρτάται μόνο από την παρούσα κατάσταση και την ενέργεια που θα κάνουμε, όχι από όλη την ιστορία μεταβάσεων. Αυτό σημαίνει ότι κάθε κατάσταση S_t περιλαμβάνει όλες τις χρήσιμες πληροφορίες της ιστορίας.

Στο πλάισιο των MDPs μία πολιτική π ορίζεται ως μία απεικόνιση από καταστάσεις σε πιθανότητες επιλογής μίας ενέργειας. Είναι δηλαδή μία συνάρτηση πυκνότητας πιθανότητας ενεργειών δεδομένων των καταστάσεων: $\pi(a|s) = \Pr(A_t = a|S_t = s)$. Κατ'επέκταση ορίζονται η συνάρτηση

κατάστασης-αξίας (state-value function): $v_\pi = E_\pi[G_t|S_t = s]$ και η συνάρτηση ενέργειας-αξίας: $q_\pi(s,a) = E_\pi[G_t|S_t = s, A_t = a]$, ακολουθώντας πολιτική π.

Τα παρατηρήσιμα περιβάλλοντα εξ'ορισμού αποτελούν μία MDP. Οι πιθανότητες p του πίνακα καταστάσεων-μεταβάσεων χαρακτηρίζουν πλήρως τις δυναμικές του περιβάλλοντος. Τα μη παρατηρήσιμα περιβάλλοντα αποτελούν *Μερικώς Παρατηρήσιμη Μαρκοβιανή Διαδικασία Απόφασης (Partially Observable Markov Decision Process - POMDP)*. Ακόμα και τέτοια προβλήματα όμως μπορούν να μετατραπούν σε MDPs. Σχεδόν όλα τα προβλήματα ενισχυτικής μάθησης περιγράφονται από το πλαίσιο των MDPs.

2.1.4 Η εξίσωση Bellman

Η εξίσωση Bellman (2.6) εκφράζει μία θεμελίωδη ιδιότητα των συναρτήσεων αξίας η οποία χρησιμοποιείται από τους περισσότερους αλγορίθμους ενισχυτικής μάθησης. Κάθε εξίσωση αξίας εκφράζεται ως το άθροισμα της άμεσης ανταμοιβής και της μειωμένης αντάμοιβής της επόμενης κατάστασης.

$$v(s) = [R_{t+1} + \gamma v(S_{t+1})|S_t = s] \tag{2.6}$$

Οι βέλτιστες συναρτήσεις αξίας $(v_\star, q_\star)$ υπακούν επίσης την εξίσωση Bellman. Έτσι ορίζεται η *εξίσωση βελτιστότητας Bellman (Bellman optimality equation)* (2.7). Διαισθητικά η αξία μιας κατάστασης υπό τη βέλτιστη πολιτική είναι ίση με τη μέση αναμενόμενη επιστροφή της καλύτερης ενέργειας από αυτή την κατάσταση.[8]

$$v_\star(s) = \max_\alpha q_\star(s,\alpha) \tag{2.7}$$

2.2 Τεχνητά νευρωνικά δίκτυα και ενισχυτική μάθηση

Τα τεχνητά νευρωνικά δίκτυα(Artificial Neural Networks - ANNs) είναι δίκτυα τεχνητών νευρώνων τα οποία χρησιμοποιούνται ευρέως για την προσέγγιση μη γραμμικών συναρτήσεων. Στην ενισχυτική μάθηση χρησιμοποιούνται για να προσέγγισουν συναρτήσεις αξίας όταν ο χώρος καταστάσεων-ενεργειών είναι πολύ μεγάλος.

Διαθέτουν ένα επίπεδο εισόδου (input layer), ένα επίπεδο εξόδου (output layer) και ενδιάμεσα ένα ή περισσότερα κρυφά επίπεδα (hidden layers). Δίκτυα με περισσότερα από ένα κρυφά επίπεδα ονομάζονται βαθιά νευρωνικά δίκτυα (deeply-layered ANNs). Οι εξελίξεις στην εκπαίδευση αυτών (deep learning) είναι υπεύθυνες για την πρόσφατη ραγδαία πρόοδο των συστημάτων μηχανικής μάθηση. Τα επίπεδα συνδέονται μεταξύ τους και κάθε σύνδεση περιλαμβάνει ένα βάρος (weight) το οποίο μπορεί να συσχετιστεί με την αποτελεσματικότητα της συναπτικής σύνδεσης στα βιολογικά νευρωνικά δίκτυα.

Οι θεμελιώδεις μονάδες σε κάθε ANN, δηλαδή οι νευρώνες, είναι ημι-γραμμικές με την έννοια ότι υπολογίζουν τον τρέχων μέσο της εισόδου τους με βάρη και του εφαρμόζουν μία μη γραμμική συνάρτηση ενεργοποίησης, συνήθως σιγμοειδή, για να παράξουν την έξοδο τους. Η ενεργοποίηση των εξόδων είναι μία μη γραμμική συνάρτηση των μορφών ενεργοποίησης των εισόδων η οποία

παραμετροποιείται ως προς τα βάρη του δικτύου. Έτσι ένα βαθύ ANN προσεγγίζει μη γραμμικές συναρτήσεις.

2.3 Αλγόριθμοι ενισχυτικής μάθησης

Σχεδόν όλοι οι αλγόριθμοι ενισχυτικής μάθησης προσεγγίζουν συναρτήσεις αξίας. Οι συναρτήσεις αυτές ορίζονται σε συνδυασμό με πολιτικές που ακολουθεί ο πράκτορας. Το πώς θα αλλάξει η πολιτική που ακολουθεί ο πράκτορας ως αποτέλεσμα της εμπειρίας που αποκτά καθορίζεται από τον εκάστοτε αλγόριθμο. Οι παρακάτω αλγόριθμοι προσεγγίζουν τη συνάρτηση $q_\star$.

2.3.1 Q-learning

Ο αλγόριθμος Q-learning είναι ένας αλγόριθμος *χωρίς μοντέλο (model-free)* ο οποίος προσπαθεί να μάθει τη συνάρτηση action-value $Q(s,a)$ μιας άγνωστης MDP προσεγγίζοντας τη $q_\star$ απευθείας. Αυτό το κάνει κοιτώντας ένα βήμα μπροστά τη φορά (one step look-ahead) και επιλέγοντας την ενέργεια που θα μεγιστοποιήσει την q τιμή της επόμενης κατάστασης ανεξάρτητα από την πολιτική που υπάρχει. Οι αλγόριθμοι που το κάνουν αυτό ονομάζονται *off-policy*. Τέτοιοι αλγόριθμοι μαθαίνουν τη βέλτιστη πολιτική από την εμπειρία που αποκτούν εξερευνώντας το περιβάλλον μέσα από επεισόδια.

Χρησιμοποιεί έναν $Q[s,a]$ πίνακα για να αποθηκεύει τις q τιμές όλων των δυνατών ζευγαριών καταστάσεων-ενεργειών. Αρχικά ο πίνακας έχει τυχαίες τιμές. Σε κάθε επεισόδιο ο αλγόριθμος επιλέγει την ενέργεια με τη μεγαλύτερη q τιμή για την κατάσταση που εξετάζει. Με την ανταμοιβή που επιστρέφεται ανανεώνει την αντίστοιχη τιμή του πίνακα σύμφωνα με τη σχέση 2.8. Μετά από πολλά επεισόδια ο πίνακας προσεγγίζει τη βέλτιστη συνάρτηση $q_\star$.

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha[R_{t+1} + \gamma \max_{a'} Q(S_{t+1},a') - Q(S_t,A_t)] \tag{2.8}$$

Η σχέση 2.8 ονομάζεται *κανόνας ενημέρωσης (update rule)* και με αυτόν γίνεται η προσέγγιση της συνάρτησης action-value. Η γενική του μορφή είναι η 2.9 και χρησιμοποιείται από πολλούς αλγόριθμους ενισχυτικής μάθησης.

$$NewEstimate \leftarrow OldEstimate + StepSize[Target - OldEstimate] \tag{2.9}$$

Ο όρος $R_{t+1} + \gamma \max_{a'} Q(S_{t+1},a')$ είναι ο στόχος (target) του Q-learning, δηλαδή το άθροισμα της ανταμοιβής που επιστρέφει το περιβάλλον και της μέγιστης q τιμής που μπορεί να φέρει η επόμενη κατάσταση πολλαπλασιασμένη με το συντελεστή μείωσης γ που έχουμε επιλέξει. Λέμε ότι έχει «θόρυβο» γιατί δεν είναι εξ'αρχής σωστό και αλλάζει μετά από ενημέρωση των καταστάσεων.

Το *Target - OldEstimate* ονομάζεται *σφάλμα (error)* της εκτίμησης. Μειώνεται καθώς πλησιάζουμε το target, το οποίο μας δείχνει την επιθυμητή κατεύθυνση παρόλο το θόρυβο που περιέχει. Στην περίπτωση του Q-learning το σφάλμα είναι το $R_{t+1} + \gamma \max_{a'} Q(S_{t+1},a') - Q(S_t,A_t)$.

Ο συντελεστής α είναι το *βήμα (StepSize)* ή αλλιώς *ρυθμός μάθησης (learning rate)* του κανόνα ενημέρωσης και χρησιμοποιείται ως παράθυρο για να κόβει παρελθοντικές εκτιμήσεις οι οποίες πλέον δεν χρειάζονται. Γενικά στη μηχανική μάθηση ο ρυθμός μάθησης βοηθάει να μην εμφανίσει ο αλγόριθμος ταλαντώσεις μεγάλου πλάτους που μπορεί να μην του επιτρέπουν να συγκλίνει.

Εχει αποδειχτεί ότι ο Q-learning συγκλίνει με πιθανότητα 1 αν συναντήσει όλα τα ζευγάρια (s,a). Για να βεβαιωθούμε ότι θα συμβεί αυτό χρησιμοποιούμε πολιτική $\varepsilon - greedy$ (2.10) η οποία εξασφαλίζει ότι όλες οι m ενέργειες δοκιμάζονται με μη-μηδενική πιθανότητα. Με πιθανότητα $1-\varepsilon$ επιλέγεται η άπληστη ενέργεια και με πιθανότητα ε επιλέγεται μια τυχαία ενέργεια. Έτσι γίνεται συμβιβασμός ανάμεσα στην αναζήτηση και την εκμετάλλευση.

$$\pi(\alpha|s) = \begin{cases} \frac{\varepsilon}{m} + 1 - \varepsilon, & \text{if } a^\star = \arg\max_{a \in A} Q(s,a). \\ \frac{\varepsilon}{m}, & \text{αλλιώς.} \end{cases} \tag{2.10}$$

Παρακάτω βλέπουμε τον ψευδοκώδικα του αλγορίθμου:

Algorithm 1 Q-learning

Require: learning rate $\alpha \in (0,1] \vee small\varepsilon > 0$
  Initialize array $Q[s,a] \forall s \in S, a \in A$
  $Q(terminal, \cdot) \leftarrow 0$
  for all *episodes* **do**
    Reset s
    for $t = 1$ to end of episode **do**
      Choose a_t from s_t using $\varepsilon - greedy$
      Take action a_t, observe r_{t+1}, s_{t+1}
      $Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t)]$
      $s_t \leftarrow s_{t+1}$
    end for
  end for

Ένα πρόβλημα του αλγορίθμου προκύπτει όταν οι συνδυασμοί καταστάσεων-ενεργειών (s,a) είναι μεγάλοι σε αριθμό. Τότε η μνήμη και ο υπολογιστικός χρόνος που χρειάζεται για να συγκλίνει ο πίνακας Q κάνουν τη μέθοδο μη αποδοτική. Επίσης δεν έχει την ικανότητα γενίκευσης αφού δεν μπορεί να κάνει εκτίμηση για άγνωστες καταστάσεις. Αυτό τον καθιστά προβληματικό σε εφαρμογές στον τομέα της τεχνητής νοημοσύνης και σε μηχανικά συστήματα. Λύση σε αυτά δίνουν μέθοδοι που προσεγγίζουν τη συνάρτηση $Q(s,a)$ με κάποια μέθοδο προσέγγισης συνάρτησης πχ: νευρωνικό δίκτυο.

2.3.2 Deep Q-Network (DQN)

Ο αλγόριθμος αυτός παρουσιάστηκε το 2015 από την DeepMind [9] ως ένας καινοτόμος αλγόριθμος ο οποίος μπορεί να αποκτήσει ένα μεγάλο εύρος ικανοτήτων σε διαφόρων ειδών απαιτητικά έργα. Συνδυάζει την ενισχυτική μάθηση, συγκεκριμένα το Q-learning, με πολυεπίπεδα (βαθειά)

ANNs (deep neural networks), για να προσεγγίσει τη βέλτιστη συνάρτηση $Q^{\star}(s,a)$ (2.11). Είναι ένας αλγόριθμος *model free* και *off-policy*.

$$Q^{\star}(s,a) = \max_{\pi} E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots | s_t = s, a_t = a, \pi] \tag{2.11}$$

Η αρχιτεκτονική του δικτύου που χρησιμοποιεί ο αλγόριθμος έχει ξεχωριστή έξοδο για κάθε πιθανή ενέργεια a και για είσοδο το διάνυσμα κατάστασης s. Για να προσεγγίσει την 2.11 την παραμετροποιεί ως προς τα βάρη του δικτύου θ_i: $Q(s,a;\theta_i)$. Το δίκτυο εκπαιδεύεται προσαρμόζοντας τις παραμέτρους θ_i σε κάθε πέρασμα i με βάση τις ενημερώσεις του q-learning ώστε να μειωθεί το μέσο τετραγωνικό σφάλμα (2.12) της εξίσωσης Bellman. Οι ενημερώσεις έχουν για targets $y = r + \gamma \max_{a'} Q(s',a';\theta_i^-)$, όπου θ_i^- είναι κάποια προηγούμενα βάρη.

$$L_i(\theta_i) = E_{s,a,r,s'}[(r + \gamma \max_{a'} Q(s',a';\theta_i^-) - Q(s,a;\theta_i))^2] \tag{2.12}$$

Οι αλγόριθμοι ενισχυτικής μάθησης είναι ασταθείς ή αποκλίνουν όταν επιχειρούν να προσεγγίσουν μη γραμμικές συναρτήσεις Q. Ο DQN όμως είναι ένας σταθερός αλγόριθμος γιατί χρησιμοποιεί τις παρακάτω τεχνικές:

- **Επανάληψη της Εμπειρίας (Experience replay):** Σε κάθε βήμα t εμπειρία $e_t = (s_t, a_t, r_t, s_{t+1})$ αποθηκεύεται σε μία δομή δεδομένων $D_t = e_1, \ldots, e_t$ που ονομάζεται *replay memory*. Κατά τη διάρκεια της εκπαίδευσης ο αλγόριθμος επιλέγει τυχαία και με ομοιόμορφη κατανομή ένα δείγμα μικρού μεγέθους (minibatches) από τη D στο οποίο εφαρμόζει τον κανόνα ενημέρωσης του $Q-learning$. Έτσι αφαιρούνται συσχετίσεις μεταξύ των δεδομένων της ακολουθίας παρατηρήσεων, μειώνεται η διακύμανση των ενημερώσεων κι ο αλγόριθμος αποφεύγει τοπικά ελάχιστα της συνάρτησης $L_i(\theta_i)$. Στην πράξη κρατώνται στη μνήμη οι N τελευταίες εμπειρίες.

- **Δίκτυο Στόχων (Target Network):** Οι q τιμές ενημερώνονται με κατεύθυνση κάποιες τιμές στόχους (target values). Για να μένουν σταθεροί οι στόχοι κατά τη διάρκεια της εκπαίδευσης χρησιμοποιούνται δύο νευρωνικά δίκτυα. Το ένα είναι αυτό που εκπαιδεύεται σε κάθε βήμα (action network) και παράγει τις q τιμές και το δεύτερο δίκτυο παρέχει του στόχους με τα βάρη θ_i^- (target network). Το target network ανανεώνει τα βάρη του πιο αργά από το action network. Αρχικά ξεκινάνε και τα δύο με τις ίδιες τυχαίες τιμές και περιοδικά, κάθε C βήματα, τα βάρη του action network αντιγράφονται στο target network. Έτσι μειώνεται η συσχέτιση με τον στόχο. Έτσι μειώνεται η συσχέτιση με τον στόχο, αποφεύγονται ταλαντώσεις και γίνεται πιο πιθανή η σύγκλιση.

- ***ε-greedy*:** Ο αλγόριθμος επιλέγει ενέργειες με βάση την $\varepsilon - greedy$ πολιτική, με ε το οποίο μειώνεται σταθερά και τελικά διατηρείται μικρό για να συμβιβάσουμε την εξερεύνηση με την εκμετάλλευση.

Ο αλγόριθμος DQN της DeepMind φαίνεται παρακάτω:

Algorithm 2 Deep Q-learning with experience replay

Initialize replay memory D to capacity N
Initialize action-value function Q with random weights θ
Initialize target action-value function $\hat{Q}$ with weights $\theta^{-} = \theta$
for *episode* $= 1$ to M **do**
  Initialize sequence $s_1 = o_1$
  for $t = 1$ to T **do**
    With probability ε select a random action a_t otherwise select $a_t = \arg\max_a Q(s_t, a; \theta)$
    Execute action a_t in emulator and observe reward r_t and next observation o_{t+1}
    Set $s_{t+1} = s_t, a_t, o_{t+1}$
    Store transition (s_t, a_t, r_t, s_{t+1}) in D
    Sample random minibatch of transitions (s_j, a_j, r_j, s_{j+1}) from D
    if episode terminates at step $J+1$ **then**
      $y_j \leftarrow_j$
    else
      $y_j \leftarrow r_j + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^{-})$
    end if
    Perform a gradient descent step on $(y_j - Q(s_j, a_j; \theta))^2$ with respect to the network parameters θ
    Every C steps reset $\hat{Q} = Q$
  end for
end for

Μέρος ΙΙΙ

Μεθοδολογία

Κεφάλαιο 3

Ο Προσομειωτής UVA/PADOVA T1D

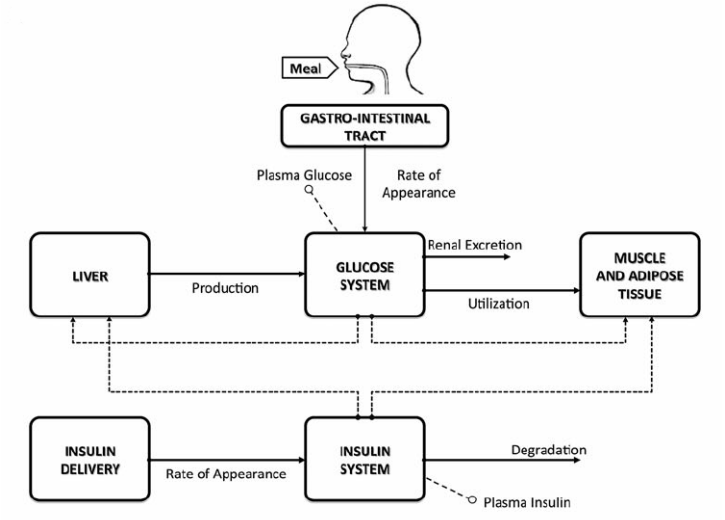
Η εξέλιξη των υπολογιστικών συστημάτων και η βελτίωση των μοντέλων κινητικότητας της γλυκόζης σε περιπτώσεις υπογλυκαιμίας οδήγησαν στην ανάπτυξη του πρώτου in silico μοντέλου διαβήτη τύπου 1 (T1D) από τα πανεπιστήμια της Virginia (UVA) και Padova. Το 2008 έγινε η παρουσίαση και εγκρίθηκε από τον Αμερικάνικο Οργανισμό Τροφίμων και Φαρμάκων (Food and Drug Administration - FDA) για να αντικαταστήσει τα πειράματα σε ζώα κάποιων θεραπειών ινσουλίνης, όπως του αλγορίθμου ελέχου τεχνητού παγκρέατος(AP). Τα in silico πειράματα υπερέχουν λόγω της ταχύτητας διεκπεραίωσης τους και του χαμηλότερου κόστους σε σχέση με τα in vivo και τελικά έχουν ανοίξει το δρόμο στην ανάπτυξη ενός AP και των δοκιμών του σε ανθρώπους.

Η πρώτη έκδοση του προσομοιωτή (S2008) [10] αποτελείται από ένα μοντέλο των δυναμικών της γλυκόζης και της ινσουλίνης κατά τη διάρκεια ενός γεύματος και από έναν in silico πληθυσμό 300 ασθενών που περιλαμβάνει ισόποσο αριθμό παιδιών, εφήβων και ενηλίκων. Τα χαρακτηριστικά των ασθενών παρήχθηκαν με τέτοιο τρόπο ώστε να αντιπροσωπεύουν τη μεταβλητότητα που εμφανίζει ο T1D όπως περιέγραφε η βιβλιογραφία που υπήρχε μέχρι τότε.

3.1 Έκδοση 2008

3.1.1 Το μοντέλο

Το μοντέλο κινητικότητας γλυκόζης-ινσουλίνης του προσομοιωτή (εικόνα 3.1) θεωρεί ότι τα υποσυστήματα της γλυκόζης (G) και της ινσουλίνης (I) συνδέονται μεταξύ τους μέσω του ελέγχου που ασκεί η ινσουλίνη στην χρήση και την ενδογενή παραγωγή γλυκόζης. Το υποσύστημα της G αποτελείται από δύο διαμερίσματα κινητικότητας της γλυκοζης και το υποσύστημα της I επίσης από δύο, το συκώτι και το πλάσμα.



Εικόνα 3.1: Το σχήμα του μοντέλου των δυναμικών μεταξύ γλυκόζης και ινσουλίνης που περιλαμβάνει ο προσομοιωτής UVA/PADOVA T1D S2008.

Με βάση το σχήμα περιγράφονται οι κύριες μεταβολές της G:

1. η *ενδογενής παραγωγή γλυκόζης (endogenous glucose production - EGP)*, η οποία όπως έχουμε πει στο κεφάλαιο 1 γίνεται από το συκώτι. Η καταστολή της EGP θεωρείται γραμμικά εξαρτώμενη από τις συγκέντρωση των G και I στο πλάσμα και από ένα καθυστερημένο σήμα ινσουλίνης το οποίο έχει να κάνει με τον χρόνο εμφάνισης της υποδόρια χορηγούμενης ινσουλίνης στο αίμα.

2. ο *ρυθμός εμφάνισης (rate of appearance - Ra)* ο οποίος περιγράφεται από ένα μοντέλο μετάβασης της γλυκόζης από το στομάχι στο λεπτό έντερο. Το στομάχι αναπαρίσταται με δύο διαμερίσματα, ένα για το στερεό στάδιο της τροφή και ένα για το υγρό, ενώ το έντερο με ένα διαμέρισμα. Ο ρυθμός αδειάσματος του γαστρεντερικού συστήματος είναι μη γραμμική συνάρτηση της ποσότητας υδατανθράκων που υπάρχουν στο στομάχι.

3. η *χρήση (utilization - U)* κατά τη διάρκεια ενός γεύματος η οποία είναι το άθροισμα δύο όρων:

   (i) της ανεξάρτητης από την I χρήσης η οποία αντιπροσωπεύει το πλάσμα και την κατανάλωση G από τον εγκέφαλο και τα ερυθρά αιμοσφαίρια. Συμβαίνει στο πρώτο διαμέρισμα του υποσυστήματος της G και είναι σταθερή.

   (ii) της εξαρτημένης από την I χρήσης η οποία αντιπροσωπεύει τους περιφερειακούς ιστούς που απορροφάνε αργά τη G. Συμβαίνει στο δεύτερο διαμέρισμα του υποσυστήματος της G το οποίο είναι απομακρυσμένο και εξαρτάται μη γραμμικά από τη G στους ιστούς.

4. η *νεφρική απέκκριση (renal extraction - E)* η οποία συμβαίνει εφόσον η G ξεπεράσει ένα όριο.

Επίσης οι κύριες μεταβολές της I είναι:

1. ο *ρυθμός εμφάνισης* της από τον υποδόριο χώρο στην κυκλοφορία του αίματος.
2. η *αποδόμηση (degradation - D)* της.

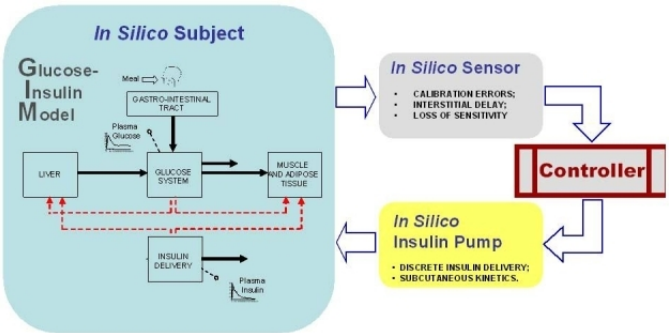
Τα παραπάνω υποσυστήματα και οι μεταξύ τους αλληλεπιδράσεις περιγράφονται από ένα σύνολο εξισώσεων που συνθέτουν το ολοκληρωμένο σύστημα. Το μοντέλο έχει 26 ελεύθερες μεταβλητές με πιο σημαντική την ευαισθησία στην ινσουλίνη, δηλαδή την ικανότητα της I του πλάσματος να αναστείλει την EGP και να ενισχύσει την E. Όλες οι μεταβλητές του μοντέλου θα έπρεπε να αλλάζουν μέσα στη μέρα, ειδικά η ευαισθησία στην ινσουλίνη, όπως συμβαίνει στην πραγματικότητα, αλλά κάτι τέτοιο δεν έχει υλοποιηθεί στον συγκεκριμένο προσομοιωτή λόγω έλλειψης γνώσης κατά την ανάπτυξη του.

3.1.2 Ο πληθυσμός

Ο in silico πληθυσμός δημιουργήθηκε από τις εξισώσεις που περιγράφουν το μοντέλο και ένα διάνυσμα παραμέτρων για κάθε ψηφιακό ασθενή με τιμές οι οποίες παράχθηκαν τυχαία με βάση κοινές κατανομές παραμέτρων. Όταν αναπτύχθηκε η έκδοση S2008 αξιόπιστες κατανομές παραμέτρων υπήρχαν για πληθυσμό χωρίς διαβήτη. Χρησιμοποιήθηκε κοινός πίνακας συνδιακύμανσης με κάποιες κλινικά σχετικές διορθώσεις στο μέσο διάνυσμα παραμέτρων. Για παράδειγμα θεωρήθηκε ότι οι διαβητικοί ασθενείς έχουν υψηλότερη EGP και ότι η ευαισθησία στην ινσουλίνη είναι διαφορετική σε κάθε ηλικιακή ομάδα.[11]

3.2 Ολοκληρωμένο περιβάλλον ανάπτυξης

Το ολοκληρωμένο περιβάλλον ανάπτυξης (Integrated Development Environment - IDE) του προσομοιωτή περιλαμβάνει, πέρα από το μοντέλο του in silico πληθυσμού, έναν in silico αισθητήρα CGM και μία in silico αντλία ινσουλίνης. Ο βρόχος κλείνει με κάποιον αλγόριθμο ελέγχου που επιλέγει ο προγραμματιστής. Το συνολικό σύστημα ελέγχου φαίνεται στην εικόνα 3.2



Εικόνα 3.2: Το in silico σύστημα ελέγχου κλειστού βρόχου με τα βασικά συστατικά που υλοποιεί το υπολογιστικό περιβάλλον του προσομοιωτή.

3.2.1 In silico αισθητήρας

Οι διακυμάνσεις της γλυκόζη στο αίμα είναι μια συνεχής διαδικασία $BG(t)$. Ένα CGM παρουσιάζει την διαδικασία αυτή ως μια διακριτή χρονική σειρά παρατηρήσεων $BG(t_n), n = 1, 2 \ldots$ η οποία την προσεγγίζει σε βήματα που καθορίζει η ανάλυση της κάθε συσκευής.

Η κατασκευή του in silico αισθητήρα CGM βασίζεται στην ανάλυση σφαλμάτων των αισθητήρων. Τέτοια σφάλματα προκύπτουν λόγω της φυσιολογικής καθυστέρησης στη μεταφορά της γλυκόζης από το αίμα στον υποδόριο χώρο και λόγω της ευαισθησίας, της σταθερότητας και της βαθμονόμησης κάθε οργάνου.

Τα δεδομένα ερευνών της ακρίβειας των CGMs οδήγησαν στην αποδόμηση των σφαλμάτων τους σε τρείς όρους λόγω της βαθμονόμησης, της καθυστέρησης μεταφοράς και τυχαίου θορύβου. Ο προσομοιωτής μοντελοποιεί το σφάλμα του αισθητήρα παράγοντας ένα τυχαίο σφάλμα βαθμονόμησης, συνδυάζοντάς το με το μοντέλο κινητικότητας γλυκόζης για το χρόνο καθυστέρησης, έναν μη λευκό θόρυβο και εφαρμόζοντάς το στις μετρήσεις $BG(t_n)$. Τα σφάλματα που παράγει είναι σε σενάρια χειρότερης περίπτωσης και στην πραγματικότητα είναι μικρότερα.

Το IDE υλοποιεί τα χαρακτηριστικά τριών διαφορετικών εμπορικών συσκευών:

1. Freestyle Navigator™ (Abbott Diabetes Care, Alameda, CA)
2. Guardian RT (Medtronic, Northridge, CA)
3. Dexcom™ STS™, 7-day sensor (Dexcom, Inc., San Diego, CA)

3.2.2 In silico αντλία

Η in silico αντλία προσομειώνει τον τρόπο χορήγησης υποδόριας ινσουλίνης. Για να το κάνει αυτό λαμβάνει υπόψιν το χρόνο εμφάνισης της ινσουλίνης στο πλάσμα μετά από την χορήγηση της στον υποδόριο χώρο και ότι οι δόσεις ινσουλίνης χορηγούνται σε διακριτές ποσότητες και χρόνους με βάση τον basal ρυθμό και τις bolus ποσότητες. Άρα το μοντέλο της αντλίας χρησιμοποιεί το μοντέλο κινητικότητας της ινσουλίνης δύο διαμερισμάτων (συκώτι και πλάσμα) και τα χαρακτηριστικά μιας εμπορικής συσκευής.

Το IDE υλοποιεί τα χαρακτηριστικά δύο διαφορετικών εμπορικών αντλιών:

1. OmniPod Insulin Management System (Insulet Corp., Bedford, MA)
2. Deltec Cozmo® (Smiths Medical MD, Inc., St. Paul, MN)

3.3 Νέες εκδόσεις

Η ανανεωμένη έκδοση του 2013 (S2013) [11] ενσωμάτωσε μη γραμμική ανταπόκριση της γλυκόζης σε καταστάσεις υπογλυκαιμίας και ένα μοντέλο του γλυκαγόνου για ρύθμιση αυτής. Επίσης ανανέωσε τα χαρακτηριστικά του in silico πληθυσμού με βάση την καινούργια βιβλιογραφία.

Και οι δύο εκδόσεις S2008, S2013 είναι σχεδιασμένες να προσομειώνουν την ανταπόκριση σε ένα γεύμα, η οποία είναι ίδια σε όλη τη διάρκεια της ημέρας που προσομειώνεται. Αυτό συμβαίνει γιατί κρατάνε σταθερές τις μεταβλητές του κάθε ασθενή, όπως η ευαισθησία στην ινσουλίνη, ενώ αυτές στην πραγματικότητα αλλάζουν. Η εξέλιξη της τεχνολογίας απαιτεί πιο ρεαλιστικά σενάρια λειτουργίας για την ανάπτυξη ενός AP. Ο ιδανικός προσομοιωτής θα πρέπει να λαμβάνει υπόψιν διακυμάνσεις στη γλυκόζη μέσα στη μέρα λόγω ασθένειας, άθλησης, γεύματα διαφορετικής σύστασης.

Η έκδοση του 2017 [12] είναι πιο ρεαλιστική γιατί μοντελοποιεί την ημερήσια διακύμανση της ευαισθησίας στην ινσουλίνη. Επίσης το μοντέλο της υποδόριας χορήγησης ινσουλίνης ενημερώθηκε με βάση τις νέες εμπορικές ινσουλίνες ταχείας δράσης και προστέθηκαν νέα μοντέλα σφάλματος των αισθητήρων. Οι in silico ασθενείς που παρήχθηκαν με αυτές τις αλλαγές εμφανίζουν ημερήσιες διακυμάνσεις γλυκόζης παρόμοιες με αυτές που έχουν καταγραφεί σε κλινικές δοκιμές, ειδικά ως προς την αύξηση της γλυκόζης κατά τη διάρκεια της νύχτας και την ευαισθησία στην ινσουλίνη.

3.4 Στατιστικά εργαλεία αξιολόγησης

Η αξιολόγηση της αποτελεσματικότητας ενός ελεγκτή έχει να κάνει με τη διακύμανση της γλυκόζης του αίματος (BG) κατά τη διάρκεια μιας προσομείωσης. Στη βιβλιογραφία υπάρχουν συγκεκριμένοι δείκτες και στατιστικά εργαλεία ανάλυσης δεδομένων από αισθητήρες CGM [13]. Το IDE χρησιμοποιεί τον δείκτη κινδύνου ο οποίος σχετίζεται με τους κινδύνους υπεργλυκαιμίας και υπογλυκαιμίας. Παρακάτω παρουσιάζουμε τα εργαλεία που θα χρησιμοποιήσουμε αργότερα στην αξιολόγηση των πειραμάτων.

3.4.1 Γλυκαιμικές ζώνες και δείκτης κινδύνου

Οι τρεις γλυκαιμικές ζώνες που έχουν οριστεί είναι:

1. η *υπογλυκαιμία (hypoglycemia)* ($< 70 mg/dl$)
2. η φυσιολογική περιοχή ή *στόχος (target range)* ($70 - 180 mg/dl$)
3. η *υπεργλυκαιμία* ($> 180 mg/dl$).

Σε περιόδους νηστείας, πχ κατά τη διάρκεια του ύπνου, τα αποδεκτά όρια είναι $70 - 140 mg/dl$. Μέσα στη μέρα ο ελεγκτής πρέπει να διατηρήσει τη γλυκόζη εντός της ζώνης στόχου. Το ποσοστό παραμονής σε κάθε περιοχή χρησιμοποιείται σαν δείκτης για την εκτίμηση της συχνότητας ακραίων τιμών.

Ο δείκτης κινδύνου κάθε χρονική στιγμή υπολογίζεται εφαρμόζοντας πρώτα την εξίσωση συμμετρικοποίησης 3.1 σε κάθε παρατήρηση της γλυκόζης.

$$f(BG) = 1.509 \times [(\ln BG)^1.084 - 5.381] \tag{3.1}$$

Στη συνέχεια εφαρμόζεται η συνάρτηση κινδύνου $r(BG) = 10 \times (f(BG))^2$ η οποία σπάει σε δύο κλάδους 3.2 που αντιστοιχούν στη χαμηλή (rl) και την υψηλή (rh) γλυκόζη.

$$r(BG) = \begin{cases} rl(BG), & \text{αν } f(BG) < 0 \\ rh(BG), & \text{αν } f(BG) > 0 \\ 0, & \text{αλλιώς} \end{cases} \tag{3.2}$$

Υπολογίζοντας τον μέσο κίνδυνο στο χρόνο για κάθε παρατήρηση παίρνουμε τους *δείκτες χαμηλής και υψηλής γλυκόζης αίματος (Low and High Blood Glucose Indices)*, οι οποίοι συμβολίζονται LBGI και HBGI αντίστοιχα.

$$LBGI = \frac{1}{n}\sum_{t=1}^{n} rl(BG_t) \text{ και } HBGI = \frac{1}{n}\sum_{t=1}^{n} rh(BG_t) \tag{3.3}$$

Οι δείκτες αυτοί χωρίζουν τη συνολική διασπορά της γλυκόζης σε δύο ανεξάρτητα μέρη που το καθένα σχετίζεται με τις μεταβάσεις στην υπέρ- και υπογλυκαιμία. Το άθροισμα τους είναι ο συνολικός δείκτης κινδύνου *Blood Glucose Risk Index - BGRI*:

$$BGRI = LBGI + HBGI \tag{3.4}$$

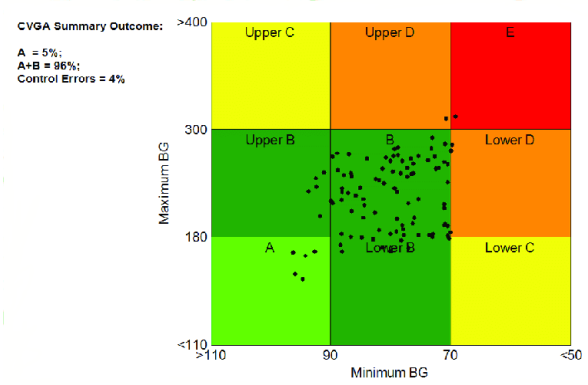
Χαρακτηρίζουμε το είδος κινδύνου που βρίσκεται κάποιος σύμφωνα με τον πίνακα 3.1.

Δείκτης	**Όρια**	**Κίνδυνος**
LBGI	≤ 1.1	μηδαμινός υπογλυκαιμίας
	$(1.1, 2.5]$	χαμηλός υπογλυκαιμίας
	$(2.5, 5]$	μέτριος υπογλυκαιμίας
	> 5	υψηλός υπογλυκαιμίας
HBGI	≤ 4.5	χαμηλός υπεργλυκαιμίας
	$(4.5, 9]$	μέτριος υπεργλυκαιμίας
	> 9	υψηλός υπεργλυκαιμίας

Πίνακας 3.1: Τα είδη κινδύνου σύμφωνα με τις τιμές των δεικτών LBGI, HBGI

3.4.2 Πλέγμα ανάλυσης ελέγχου διακύμανσης

Το *πλέγμα ανάλυσης ελέγχου διακύμανσης (control variability grid analysis - CVGA)* είναι ένας τρόπος ποσοτικοποίησης της ποιότητας του ελέγχου που κάνει ένα σύστημα κλειστού βρόχου. Στο γράφημα CVGA (εικόνα 3.3) ο κάθε ασθενής αναπαρίσταται με ένα σημείο του οποίου οι συντεταγμένες είναι η μικρότερη τιμή γλυκόζης $x \in (50mg/dl, 110mg/dl)$ και η μεγαλύτερη $y \in (110mg/dl, 400mg/dl)$ που παρατηρήθηκαν κατά τη διάρκεια της προσομείωσης.



Εικόνα 3.3: Παράδειγμα CVGA γραφήματος για πολλούς ασθενείς. Καθένας αναπαρίσταται από ένα σημείο με βάση τις ακραίες τιμές γλυκόζης που εμφανίζει σε ένα χρονικό διάστημα.

Οι τιμές στον x εμφανίζονται ανάστροφα με αποτέλεσμα στην κάτω αριστερή γωνία να είναι η βέλτιστη περιπτώση που μπορεί να βρεθεί ένας ασθενής ενώ στην πάνω δεξιά η χειρότερη. Το x-y επίπεδο χωρίζεται συνολικά σε 9 τομείς οι οποίοι επεξηγούνται στον πίνακα 3.2. Για να είναι οι ζώνες τετραγωνικές χρησιμοποιείται ένα απλός γραμμικός μετασχηματισμός. Επίσης στο γράφημα σημειώνονται τα ποσοστά των σημείων που βρίσκονται σε κάθε ζώνη.

Ζώνη	**Σημασία**	**Όρια**
A	Ακριβής έλεγχος	$(x,y) \in (90,110) \times (110,180)$
Lower B	Ασήμαντες αποκλίσεις προς την υπογλυκαιμία	$(x,y) \in (70,90) \times (110,180)$
B	Ασήμαντες αποκλίσεις ελέγχου	$(x,y) \in (70,90) \times (180,300)$
Upper B	Ασήμαντες αποκλίσεις προς την υπεργλυκαιμία	$(x,y) \in (90,110) \times (180,300)$
Lower C	Υπερβολική διόρθωση της υπεργλυκαιμίας	$(x,y) \in (50,70) \times (110,180)$
Upper C	Υπερβολική διόρθωση της υπογλυκαιμίας	$(x,y) \in (90,110) \times (300,400)$
Lower D	Αποτυχία στη διαχείριση της υπογλυκαιμίας	$(x,y) \in (50,70) \times (180,300)$
Upper D	Αποτυχία στη διαχείριση της υπεργλυκαιμίας	$(x,y) \in (70,90) \times (300,400)$
E	Λανθασμένος έλεγχος	$(x,y) \in (50,70) \times (300,400)$

Πίνακας 3.2: Πίνακας επεξήγησης των ζωνών του διαγράμματος CVGA.

Τα παραπάνω όρια τιμών επιλέγχθηκαν με βάση τις τρεις γλυκαιμικές ζωνες που αναφέραμε. Ειδικά το κάτω όριο της ζώνης A επιλέχθηκε να είναι το $90 mg/dl$ για να είναι η ελάχιστη ασφαλής τιμή που πρέπει να διατηρήσει ένας ελεγκτής ώστε να μην υπάρξει κίνδυνος υπογλυκαιμίας. Το $300 mg/dl$ είναι το άνω όριο της αποδεκτής υπογλυκαιμίας που μπορεί να έχει κάποιος σύμφωνα με απόψεις γιατρών.

Το CVGA γράφημα σχεδιάζεται από τις ελάχιστες και μέγιστες τιμές μια παρατήρησης η οποία περιέχει θόρυβο, αφού όπως είπαμε οι τιμές των αισθητήρων CGM περίεχουν σφάλματα. Αυτό κάνει το γράφημα επιρρεπές στο θόρυβο και τις ακραίες τιμές. Γι'αυτό το κάτω όριο τίθεται στο 2,5% εκατοστημόριο (percentile) της κατανομής των δεδομένων και το άνω όριο στο 97,5%.

3.5 Υλοποίηση σε Python

Μια έκδοση ανοιχτού κώδικα του προσομοιωτή S2008 έχει υλοποιηθεί σε Python3 [14] με σκοπό να παρέχει ένα περιβάλλον ανάπτυξης αλγορίθμων ενισχυτικής μάθησης για έλεγχο του συστήματος κλειστού βρόχου. Περιλαμβάνει τους εμπορικούς αισθητήρες CGM και τις αντλίες που υλοποιεί ο S2008 καθώς και 30 in silico ασθενείς: 10 παιδιά, 10 εφήβους και 10 ενήλικες. Επίσης διαθέτει υλοποιημένο έναν basal-bolus ελεγκτή (BBcontroller). Ο προγραμματιστής μπορεί να δημιουργήσει τους δικούς του ελεγκτές πχ RL, PID καθώς και δικές του συναρτήσεις ανταμοιβής.

Παρέχει ενσωματωμένο το gym 0.9.4, ένα εργαλείο το οποίο χρησιμοποιείται για την ανάπτυξη των αλγορίθμων. Μέσω αυτού δημιουργείται το περιβάλλον, παίρνουμε παρατηρήσεις, κάνουμε ενέργειες και λαμβάνουμε ανταμοιβές.

Στο τέλος κάθε προσομείωσης παράγει ένα αρχείο *.csv* με τις μετρήσεις που έκανε σε βήμα του ρυθμού δειγματοληψίας του αισθητήρα CGM που επιλέχθηκε και εμφανίζει μια σειρά διαγραμμάτων. Οι μετρήσεις περιλαμβάνουν την πραγματική γλυκόζη στο αίμα (BG) σε mg/dl , την αντίστοιχη τιμή που μετράει το cgm (CGM), τις μονάδες (U) ινσουλίνης που έδωσε η αντλία, τα γραμμάρια (g) υδατάνθρακα (CHO) που καταναλώθηκαν και τους δείκτες κινδύνου LBGI, HBGI, BGRI. Τα διαγράμματα που παράγονται είναι:

1. η εξέλιξη των τιμών της γλυκόζης BG και της γλυκόζης που μετρά ο αισθητήρας CGM στον άξονα του χρόνου
2. ιστόγραμμα του ποσοστού παραμονής της γλυκόζης του αίματος σε γλυκαιμικές ζώνες
3. ιστόγραμμα των δεικτών κινδύνου
4. CVGA

Κεφάλαιο 4

Λεπτομέρειες Υλοποίησης

Προσεγγίσαμε το πρόβλημα του T1D ως ένα πρόβλημα ενισχυτικής μάθησης και επιχειρήσαμε να το λύσουμε με την χρήση του αλγορίθμου DQN. Δηλαδή ελέγξαμε αν ένας πράκτορας που έχει εκπαιδευτεί με τον τρόπο που περιγράφει ο DQN μπορεί να παίξει το ρόλο του ελεγκτή σε ένα τεχνητό πάγκρεας. Η διατύπωση του προβλήματος, οι λεπτομέρειες υλοποίησης του DQN και τα αποτελέσματα των πειραμάτων παρουσιάζονται παρακάτω.

4.1 Βασικά συστατικά ενισχυτικής μάθησης

Θεωρούμε το κλασσικό πρόβλημα ενισχυτικής μάθησης όπου ένας πράκτορας αλληλεπιδρά με ένα περιβάλλον, στη δεδομένη περίπτωση το περιβάλλον τον προσομοιωτή, δηλαδή το μοντέλο γλυκόζης-ινσουλίνης. Κάθε χρονική στιγμή t, η οποία καθορίζεται από το ρυθμό δειγματοληψίας του CGM, ο πράκτορας επιλέγει μία ενέργεια ινσουλίνης από ένα σύνολο ενεργειών που έχουμε ορίσει εμείς. Η ενέργεια δίνεται στο περιβάλλον, αλλάζει την εσωτερική του κατάσταση και αυτό επιστρέφει μία ανταμοιβή και μία παρατήρηση.

Ορίζουμε στο προβλημά μας:

- **Κατάσταση s_t:** το διάνυσμα $S = [CGM, CHO, insulin, time]$, δηλαδή η γλυκόζη που μετράει ο αισθητήρας, οι υδατάνθρακες που καταναλώθηκαν, η προηγούμενη δόση ινσουλίνης που δώθηκε και η ώρα της ημέρας που έγινε η παρατήρησης. Δοκιμάσαμε διάφορα διανύσματα καταστάσεων $O_i \subseteq S$ για να ελέγξουμε ποιο έχει τα καλύτερα αποτελέσματα.

- **Ενέργεια a_t:** μία ενέργεια από το σύνολο ενεργειών $A = 0, basal, 5 \times basal$. Η basal τιμή είναι διαφορετική για κάθε ασθενή και υπολογίζεται από τη σχέση $basal = (u2_{ss} * bw)/6000$, όπου η $U2_{ss}$ είναι η χρήση U της γλυκόζης στη μόνιμη κατάσταση (steady state) και bw είναι το βάρος του ασθενή. Αυτές οι τιμές βρίσκονται από το διάνυσμα παραμέτρων κάθε in silico ασθενή που έχει ο προσομοιωτής.

- **Ανταμοιβή r_t:** η ανταμοιβή που επιστρέφει ο προσομοιωτής μετά από κάθε ενέργεια η οποία υπολογίζεται από τη συνάρτηση 4.1. Η συνάρτηση αυτή είναι η διαφορά του δείκτη

κινδύνου που βρισκόταν ο ασθενής στο βήμα $t-1$ και του δείκτη στο βήμα t.

$$R(t) = BGRI(t-1) - BGRI(t) \tag{4.1}$$

Το περιβάλλον που δρα ο πράκτορας είναι μερικώς παρατηρήσιμο γιατί δεν μπορεί να μάθει απευθείας την τιμή γλυκόζης στο αίμα (BG) αλλά μόνο την μετρούμενη από έναν αισθητήρα CGM τιμή η οποία περιέχει θόρυβο. Μέσα από πολλά επεισόδια παρατηρήσεων, ενεργειών, ανταμοιβών (s_t, a_t, r_t, s_{t+1}) ο αλγόριθμος εκπαιδεύεται να αναγνωρίζει καταστάσεις και να λαμβάνει ενέργειες έτσι ώστε να μεγιστοποιήσει τη συνολική ανταμοιβή που μαζεύει στο τέλος.

Η κατάσταση που βρίσκεται το περιβάλλον κάθε χρονική στιγμή περιέχει όλη την πληροφορία που χρειάζεται ο πράκτορας για να αποφασίσει τι θα κάνει αργότερα. Αυτό συμβαίνει γιατί το σύστημα χαρακτηρίζεται πλήρως σε κάθε στιγμή από την τιμή που έχει η γλυκόζη, αν λαμβάνεται γεύμα ή όχι, τι ινσουλίνη έχει ήδη δοθεί και τι ώρα είναι. Άρα οι καταστάσεις του προβλήματος έχουν την ιδιότητα Markov. Επίσης το περιβάλλον έχει αρχική κατάσταση ίδια για κάθε επεισόδια και τελική κατάσταση. Συμπεραίνουμε λοιπόν ότι ο παραπάνω φορμαλισμός ορίζει μία μεγαλη αλλά πεπερασμένη MDP στην οποία μπορούν να εφαρμοσθούν οι μέθοδοι επίλυσης που περιγράφηκαν στο κεφάλαιο της ενισχυτικής μάθησης.

4.2 Αρχιτεκτονική και υπερπαράμετροι

Ο αλγόριθμος DQN που υλοποιήσαμε χρησιμοποιεί δύο νερωνικά δίκτυα, το action και το target, με 2 κρυφά επίπεδα 265 νευρώνων, πλήρως συνδεδεμένα (fully connected) τα οποία έχουν είσοδο το διάνυσμα παρατηρήσων και 3 εξόδους, μία για κάθε ενέργεια. Οι ενέργειες επιλέγονται με πολιτική $\varepsilon - greedy$ η οποία ξεκινάει με $\varepsilon = 1$ και μειώνεται σταθερά μέχρι την τιμή $\varepsilon = 0.1$. Η συχνότητα ανανέωσης των βαρών του target δικτύου είναι τα 100 βήματα. Κάθε εμπειρία $< s_{t-1}, a_t, r_t, s_t >$ αποθηκεύεται σε μια replay memory της οποίας το μέγεθος επιλέχθηκε να είναι 10000.

Οι τιμές των υπερπαραμέτρων (hyperparameters):

- **Ρυθμός μάθησης:** $lr = 0.0001$
- **Συντελεστής μείωσης:** $\gamma = 0.99$
- **Μέγεθος minibatch:** 128

4.2.1 Ανάπτυξη και Εκπαίδευση

Χρησιμοποιήσαμε το gym για να υλοποιήσουμε τον αλγόριθμο DQN ώστε να επιλύσουμε το πρόβλημα ενισχυτικής μάθησης που ορίσαμε. Εκπαιδεύσαμε τον DQN πράκτορα με διάφορους τρόπους, διαφορετικό διάνυσμα καταστάσεων για να αποφασίσουμε για το καλύτερο, διαφορετικό σύνολο ενεργειών και συνάρτηση ανταμοιβής. Προτιμήσαμε την ενσωματωμένη συνάρτηση γιατί παίρνει τιμές στο διαστημα $-1 \leq r(t) \leq 1$. Στο τέλος κάθε επεισοδίου, αν η τελική κατάσταση

ήταν διαφορετική από το τέλος της ημέρας, δηλαδή απέτυχε να ελέγξει τη γλυκόζη εντός των φυσιολογικών ορίων και η προσομείωση τελείωσε νωρίτερα, εφαρμόζαμε επιπρόσθετη τιμωρία.

Για τη δημιουργία των νευρωνικών δικτύων που χρησιμοποιεί ο DQN αλγόριθμος χρησιμοποιήσαμε το framework Keras. Μέσω αυτού κατασκευάσαμε τα δίκτυα και τα εκπαιδεύσαμε. Η εκπαίδευση γινόταν πρώτη φορά μόλις υπήρχαν αρκετά δείγματα εμπειρίας ($\geq$ minibatch size) στη replay memory και στη συνέχεια στο τέλος κάθε βήματος t. Για την εκπαίδευση έτρεχε ο αλγόριθμος backpropagation για μία εποχή. Χρησιμοποιήσαμε τον Adam optimizer για τα q-updates ο οποίος είναι κατάλληλος για βελτιστοποίηση στοχαστικών συναρτήσεων.

Πραγματοποιήθηκαν διαφορετικές εκπαιδεύσεις για να ελέγξουμε διάφορα διανύσματα κατάστασης και τη δυνατότητα εξειδίκευσης αλλά και γενίκευσης του πράκτορα. Το σενάριο γευμάτων κάθε ασθενή στις εκπαιδεύσεις που έγιναν με τυχαία γεύματα, δημιουργήθηκαν με τη χρήση μιας γεννήτριας παραγωγής τυχαίων σεναρίων (αλγόριθμος 3) γύρω από συγκεκριμένες τιμές υδατανθράκων και χρονικές στιγμές ο οποίος προήλθε από τη βιβλιογραφία [15]. Κάθε εκπαίδευση είχε ως αποτέλεσμα την παραγωγή ενός μοντέλου το οποίο χρησιμοποιήθηκε στη συνέχεια στα πειράματα. Οι εκπαιδεύσεις έγιναν για 10000 επεισόδια διάρκειας 24 ώρων το καθένα.

1. **Μοντέλο 1:** Εκπαίδευση με τον ασθενή adult4, για ιδίο σενάριο γεύματος σε κάθε επεισόδιο, το οποίο δημιουργήθηκε στην αρχή με τον αλγόριθμο 3 και διάνυσμα κατάστασης $[CGM, CHO, insulin]$. Έγιναν δύο διαφορετικές εκπαιδεύσεις με ρυθμό δειγματοληψίας $1min$ και $3min$ και replay memory μεγέθους 1000000 και 10000. Σκοπός αυτής της εκπαίδευσης ήταν η δημιουργία ενός εξειδικευμένου μοντέλου για κάποιον ασθενή με σταθερή ρουτίνα. Επίσης να ελεγχθεί το διάνυσμα κατάστασης για την καταλληλότητά του, να βρεθεί το καλύτερο μέγεθος μνήμης και ο ρυθμός δειγματοληψίας. Στα πειράματα χρησιμοποιήθηκε το μοντέλο που εκπαιδεύτηκε με ρυθμό 1 και μνήμη μεγέθους 10000 γιατί είχε καλύτερα αποτελέσματα.

2. **Μοντέλο 2:** Εκπαίδευση με τυχαίο ασθενή και τυχαίο σενάριο γεύματος σε κάθε επεισόδιο, διάνυσμα κατάσταστης $[CGM, CHO, time]$, ρυθμό δειγματοληψίας $1min$ και replay memory μεγέθους 10000. Οι ασθενείς προέρχονταν από το σύνολο των ενήλικων ασθενών. Σκοπός αυτής της εκπαίδευσης ήταν η δημιουργία ενός γενικού μοντέλου για όλους τους ασθενείς χωρίς συγκεκριμένες ρουτίνες γευμάτων, ώστε να εξεταστεί η δυνατότητα δημιουργίας ελεγκτή γενικού σκοπού.

3. **Μοντέλο 3:** Εκπαίδευση με τον ασθενή adult4, τυχαία σενάρια γεύματων σε κάθε επεισόδιο, διάνυσμα κατάστασης $[CGM, CHO, time]$, ρυθμό δειγματοληψίας $1min$ και replay memory μεγέθους 10000. Σκοπός αυτής της εκπαίδευσης ήταν η δημιουργία ενός μοντέλου που θα εξειδικεύεται σε συγκεκριμένο ασθενή αλλά χωρίς σταθερή ρουτίνα γευμάτων.

Algorithm 3 Generate Meal Schedule

Require: body weight w, number of days n

$MealOcc = [0.95, 0.3, 0.95, 0.3, 0.95, 0.3]$
$TimeLowerBounds = [5, 9, 10, 14, 16, 20] * 12$
$TimeUpperBounds = [9, 10, 14, 16, 20, 23] * 12$
$TimeMean = [7, 9.5, 12, 15, 18, 21.5] * 12$
$TimeStd = [1, 0.5, 1, 0.5, 1, 0.5] * 12$
$AmountMean = [0.7, 0.15, 1.1, 0.15, 1.25, 0.15] * w$
$AmountStd = AmountMean * 0.15$
$Days = []$
for $i = 1$ to n **do**
  $M = [0]_{j=1}^{288}$
  for $j = 1$ to 6 **do**
    $m \sim Binomial(MealOcc[j])$
    $lb = TimeLowerBounds[j]$
    $ub = TimeUpperBounds[j]$
    $\mu_t = TimeMean[j]$
    $\sigma_t = TimeStd[j]$
    $\mu_a = AmountMean[j]$
    $\sigma_a = AmountStd[j]$
    if m **then**
      $t \sim Round(TruncNormal(\mu_t, \sigma_t, lb, ub))$
      $c \sim Round(max(0, Normal(\mu_a, \sigma_a)))$
      $M[t] = c$
    end if
  end for
  $Days.append(M)$
end for

4.3 Συστατικά Συστήματος Ελέγχου

Συνδυάζοντας λοιπόν τον αλγόριθμο DQN με μία αντλία και έναν αισθητήρα CGM επιχειρήσαμε να κατασκευάσουμε ένα σύστημα ελέγχου γλυκόζης κλειστού βρόχου. Κατασκευάσαμε έναν ελεγκτή ο οποίος χρησιμοποιεί το εκπαιδευμένο μοντέλο, ανάλογα το πείραμα, για να επιλέξει τη βέλτιστη ενέργεια σε κάθε βήμα. Οι ενέργειες ινσουλίνης γίνονται αλλάζοντας μόνο την basal δόση που εγχύει η αντλία.

Στην εκπαίδευση χρησιμοποιήσαμε την αντλία Insulet η οποία έχει δυνατότητα να δώσει μέχρι και 30 μονάδες basal ινσουλίνη και αντίστοιχα bolus. Οι δόσεις που παρέχει αυξάνουν με βήμα 0,05 της μονάδας. Επίσης δοκιμάσαμε δύο αισθητήρες CGM που διαθέτει το IDE, τον αισθητήρα Dexcom με ρυθμό δειγματοληψιας $3min$ και τον Navigator με $1min$. Και οι δύο αισθήτηρες μπορούν να μετρήσουν τιμές μέχρι και $600mg/dl$, ενώ η μικρότερη τιμή που μετράει ο πρώτος είναι $39mg/dl$ και ο δεύτερος $32mg/dl$. Τα πειράματα που περιγράφονται στο επόμενο μέρος έγιναν με χρήση της αντλίας Insulet και του αισθητήρα Navigator.

4.3.1 Αξιολόγηση

Η αξιολόγηση των αποτελεσμάτων έγινε με χρήση των γραφημάτων που παράγει το IDE σε python όπως αναφέρθηκαν παραπάνω. Δημιουργήσαμε συγκεντρωτικές γραφικές παραστάσεις για τα διαφορετικά πειράματα ώστε να συγκρίνουμε την αποδοσή τους και να βρούμε ποιο πραγματοποιεί καλύτερο έλεγχο. Ιδιαίτερη έμφαση στην αξιολόγηση δώθηκε στο αποτέλεσμα του CVGA και στο ιστόγραμμα των δεικτών κινδύνου.

Μέρος IV

Αποτελέσματα

Κεφάλαιο 5

Πειραματική Διαδικασία

Προκειμένου να αξιολογήσουμε τη δυνατότητα τον μοντέλων που εκπαιδεύσαμε να αντιμετωπίσουν το πρόβλημα ελέγχου γλυκόζης μέσω έγχυσης ινσουλίνης, κάναμε μια σειρά από πειράματα αυξανόμενης πολυπλοκότητας. Συγκρίναμε την απόδοσή τους μεταξύ τους για να αποφασίσουμε για τα στοιχεία που φέρνουν καλύτερα αποτελέσματα αλλά και με έναν ελεγκτή τυχαίων ενεργειών και τον Basal-Bolus ελεγκτή που περιελάμβανε το python IDE. Η πειραματική διαδικασία και τα αποτελέσματα περιγράφονται παρακάτω.

5.1 Έλεγχος γλυκόζης για έναν ασθενή και συγκεκριμένο πλάνο γεύματος

Αρχικά εντοπίσαμε τις βέλτιστες παραμέτρους του προβλήματος ενισχυτικής μάθησης, κάτι το οποίο είναι μια σημαντική διαδικασία για όλους τους αλγορίθμους ενισχυτικής μάθησης, αφού επηρεάζουν την ικανότητα μάθησης και τη συμπεριφορά των πρακτόρων. Αυτο σημαίνει ότι έπρεπε να ορίσουμε τι θεωρούμε σύνολο ενεργειών που μπορεί να κάνει ο πράκτορας με τέτοιο τρόπο ώστε να αντιπροσωπεύουν πλήρως το πρόβλημα που έχουμε να αντιμετωπίσουμε.

5.1.1 Ορισμός τιμών ενεργειών

Γενικά στο πρόβλημα του ελέγχου της γλυκόζης, ενέργεια ορίζουμε τις μονάδες (U) της ινσουλίνης που εγχύει η αντλία στον ασθενή σε κάθε χρονική στιγμη t. Δοκιμάστηκαν 2 διαφορετικά σχήματα διακριτών ενεργειών γιατί ο αλγόριθμος DQN χρησιμοποιεί τέτοιες ενέργειες. Έγινε λοιπόν διακριτοποίηση ώστε να ικανοποιηθεί αυτό το χαρακτηριστικό.

Στην πρώτη περίπτωση ορίσαμε n ομοιόμορφα διαστήματα τιμών με βάση τη μέγιστη ποσότητα ινσουλίνης που μπορεί να δώσει μία αντλία στο ποσό της ινσουλίνης. Το ποσό που εγχυόταν κάθε φορά καθοριζοταν από το αποτέλεσμα του μοντέλου ενεργειών και δειγματοληψία στο αντίστοιχο διάστημα με χρήση μιας ομοιόμορφης κατανομή.

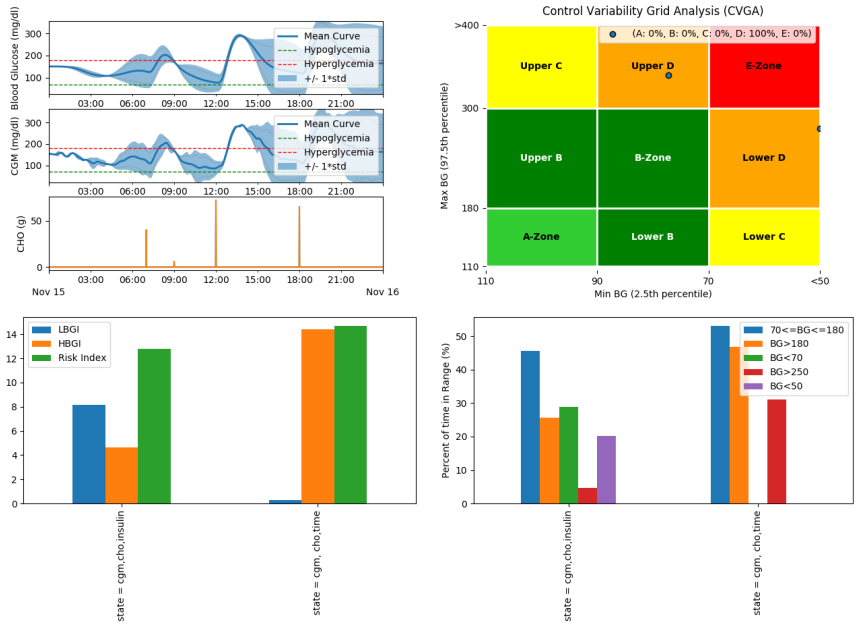
Στην δεύτερη περίπτωση ορίσαμε 3 ενέργειες: $A = 0, basal, 5 \times basal$, όπως προτείνει η βι-

βλιογραφία [15]. Η μηδενική ενέργεια αντιστοιχούσε στο να μην εγχυθεί ινσουλίνη στον ασθενή. Η *basal* ενέργεια αντιστοιχούσε στην έγχυση της σταθερής ποσότητας ινσουλίνης που χρειάζεται ο εκάστοτε ασθενείς για να διατηρήσει σταθερή την γλυκόζη του σε περιόδους νηστείας όπως ορίστικε στο προηγούμενο κεφάλαιο. Η τρίτη ενέργεια ήταν η $5 \times basal$ και αντιστοιχεί στην αναμενόμενη ενέργεια που θα έκανε το πάγκρεας σε περίπτωση λήψης τροφής, όπου χρειάζεται μεγαλύτερη έγχυση ινσουλίνης. Δεν ορίσαμε μεγαλύτερη ποσότητα γλυκόζης για να μειώσουμε την πολυπλοκότητα.

Παρατηρήσαμε ότι ο πράκτορας που ακολουθεί το δεύτερο σχήμα ενεργειών A μπορούσε να ελέγξει το πρόβλημα της γλυκόζης στον T1D ενώ το πρώτο σχήμα ενεργειών αποτύγχανε. Αυτό συνέβαινε γιατί ο πράκτορας δεν μπορούσε να μάθει ποια είναι η σωστή ενέργεια, λόγω της τυχαιότητας που εισάγαμε σε κάθε ενέργεια.

5.1.2 Σταθερό πρόγραμμα γεύματος

Δοκιμάσαμε αρχικά αν ο ελεγκτής του μοντέλου 1 αποδίδει καλύτερα στο γεύμα για το οποίο εκπαιδεύτηκε από ότι ο ελεγκτής του μοντέλου 3 που εκπαιδεύτηκε για τυχαία γεύματα. Τα αποτελέσματα φαίνονται στην εικόνα 5.1. Σκοπός είναι να παρατηρήσουμε τις διαφορές ανάμεσα στους δύο πράκτορες.



Εικόνα 5.1: Τα αποτελέσματα ελέγχου της γλυκόζης στον ασθενή adult4 με χρήση του μοντέλου σταθερής ρουτίνας γευμάτων και του μοντέλου τυχαίων γευμάτων.

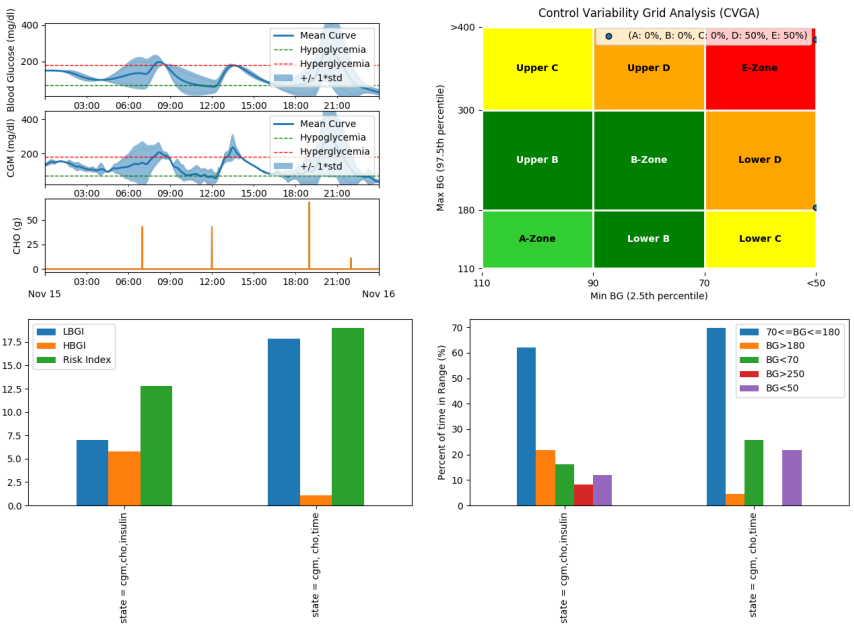
Με βάση τα παραπάνω αποτελέσματα συμπεραίνουμε ότι και οι δύο ελεγκτές δεν διαχειρίζονται σωστά το σύστημα. Ο ελεγκτής του μοντέλου 1, με διάνυσμα κατάστασης $[CGM, CHO, insulin]$, αποτυγχάνει να διαχειριστεί την υπογλυκαιμία, αφού εμφανίζει μέτριο κίνδυνο υπεργλυκαιμίας και υψηλό υπογλυκαιμίας, και γι'αυτό ανήκει στη ζώνη Lower D του CVGA. Ο ελεγκτής του μοντέλου

3, με διάνυσμα κατάστασης $[CGM, CHO, time]$, αποτυγχάνει να διαχειριστεί την υπεργλυκαιμία, αφού εμφανίζει υψηλό υπεργλυκαιμίας και μηδαμινό υπογλυκαιμίας και γι'αυτό ανήκει στη ζώνη Upper D του CVGA. Ο συνολικός κίνδυνος που εμφανίζουν και οι δύο είναι κοντινός, με λίγο χαμηλότερο του πρώτου ελεγκτή.

Βλέπουμε όμως ότι και οι δύο ελεγκτές ξεκινάνε με παρόμοια συμπεριφορά. Περιμένουν την άφιξη του πρώτου γεύματος και γι'αυτό αρχίζουν από νωρίς να ρίχνουν τη γλυκόζη. Ο ελεγκτής που έχει εκπαιδευτεί στο συγκεκριμένο γεύμα προσπαθεί να ρίξει την γλυκόζη και πριν από τα επόμενα γεύματα πιο αποτελεσματικά από ότι ο δεύτερος. Δεν το καταφέρνει όμως πλήρως γιατί δεν μπορεί να ρίξει μεγάλες ποσότητες ινσουλίνης αφού δεν υπάρχει αυτή η επιλογή στο σύνολο ενεργειών και γιατί μπαίνει σε επίπεδα υπογλυκαιμίας.

5.2 Έλεγχος γλυκόζης για έναν ασθενή και τυχαίο πλάνο γεύματος

Πειραματιστήκαμε χρησιμοποιώντας το μοντέλο 3, με διάνυσμα κατάστασης $[CGM, CHO, time]$ και με τυχαία γεύματα σε κάθε επεισόδιο εκπαίδευσης όπως αναφέραμε στο κεφάλαιο 4. Στόχος ήταν να δούμε αν ο πράκτορας του μοντέλου 3 θα αρχίσει να εντοπίζει μοτίβα γευμάτων και αν θα ανταπεξέρχεται σε τυχαίες αφίξεις τους. Τον συγκρίναμε με το αποτέλεσμα ελέγχου του μοντέλου 1 με τυχαίο γεύμα. Το αποτέλεσμα αυτού του πειράματος φαίνεται στις εικόνες του σχήματος 5.2.



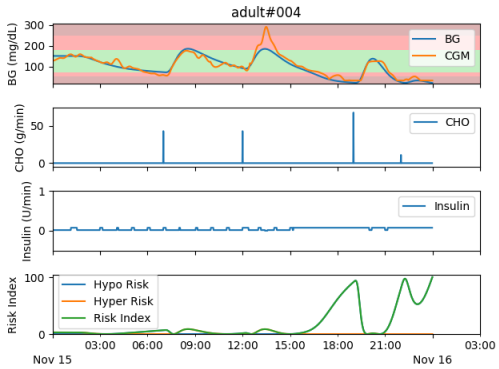
Εικόνα 5.2: Τα αποτελέσματα ελέγχου της γλυκόζης στον ασθενή adult4 για τυχαίο γεύμα με μοντέλο εκπαιδευμένο σε τυχαία γεύματα και διάνυσμα κατάστασης $[CGM, CHO, time]$ και για μοντέλο εκπαιδευμένο σε ένα συγκεκριμένο σενάριο γεύματος και διάνυσμα κατάστασης $[CGM, CHO, insulin]$.

Βλέπουμε ότι ο ελεγκτής που εκπαιδεύτηκε με τυχαία γεύματα διατηρεί τη γλυκόζη κατά 70% στα φυσιολογικά όρια. Εμφανίζει όμως υψηλό κίνδυνο υπογλυκαιμίας και χαμηλό υπεργλυκαι-

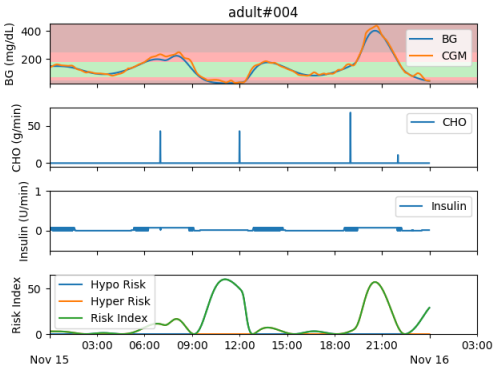
μίας. Γι'αυτό ανήκει στη ζώνη Lower D του CVGA αφού αποτυγχάνει να διαχειριστεί την υπογλυκαιμία.

Παρατηρούμε όμως ότι και ο ελεγκτής που είχε εκπαιδευτεί μόνο με ένα συγκεκριμένο γεύμα διατηρεί τη γλυκόζη στα φυσιολογικά όρια σε ποσοστό λίγο μεγαλύτερο του 60%. Ο συνολικός κίνδυνος που εμφανίζει είναι μικρότερος από ότι του ελεγκτή με το μοντέλο 3 παρουσιάζοντας μέτριο κίνδυνο υπεργλυκαιμίας. Παρ'ολαυτά παρουσιάζει υψηλό κίνδυνο υπογλυκαιμίας και γι'αυτό μπαίνει στη ζώνη E του CVGA αφού αποτυγχάνει να ελέγξει το σύστημα.

Πάλι οι δύο ελεγκτές εμφανίζουν παρόμοια συμπεριφορά στην αρχή, διαχειρίζονται την νηστεία κατά τη διάρκεια της νύχτας και ρίχνουν τη γλυκόζη εν αναμονή για το πρώτο γεύμα. Είναι αξιοσημείωτο ότι ο ελεγκτής που εκπαιδεύτηκε με συγκεκριμένο γεύμα τα καταφέρνει καλά και σε τυχαίο. Ακολουθούν τα διαγράμματα χρόνου για του δύο ελεγκτές (εικόνες 5.3 και 5.4).



Εικόνα 5.3: Οι διακυμάνσεις της γλυκόζης στο χρόνο του ασθενή adult4 για τυχαίο γεύμα με χρήση του μοντέλου 3.



Εικόνα 5.4: Οι διακυμάνσεις της γλυκόζης στο χρόνο του ασθενή adult4 για τυχαίο γεύμα με χρήση του μοντέλου 1.

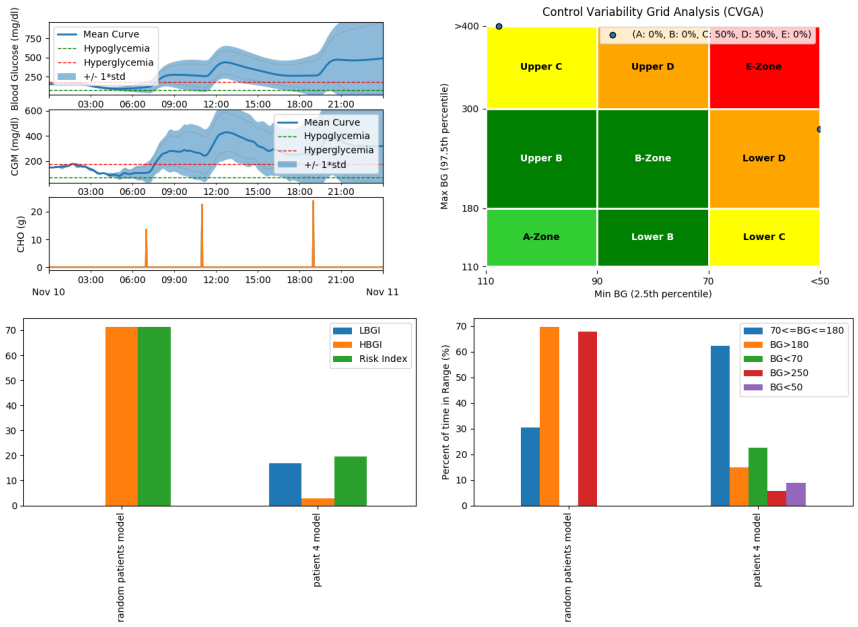
Παρατηρούμε ένα σφάλμα του CGM αισθητήρα που δείχνει τη γλυκόζη στην εικόνα 5.3 να έχει φτάσει σε επίπεδα υπεργλυκαιμίας ενώ η πραγματική τιμή γλυκόζης είναι εντός ορίων. Σε αυτή την περίπτωση ο ελεγκτής έριξε παραπάνω ινσουλίνη και γι'αυτό αργότερα έφτασε η γλυκόζη επίπεδα υπογλυκαιμίας.

5.3 Έλεγχος γλυκόζης για τυχαίο ασθενή και τυχαίο πλάνο γεύματος

Στη συνέχεια προσπαθήσαμε να αντιμετωπίσουμε την γενική περίπτωση του ελέγχου της γλυκόζης στον T1D. Στόχος ήταν η δημιουργία ενός ελεγκτή που να μπορεί να γενικεύει καλά σε διαφορετικούς ασθενείς και τυχαία σενάρια γεύματος, δηλαδή να μπορεί να συνδυάσει την πρόβλεψη με την γενίκευση. Το διάνυσμα κατάστασης που χρησιμοποιήθηκε είναι το $[CGM, CHO, time]$ και σε αυτή την περίπτωση και σε κάθε επεισόδιο το σενάριο γεύματος άλλαζε.

Για την αξιολόγηση του μοντέλου αυτού συγκρίνουμε αρχικά τον έλεγχο που πραγματοποιεί ο πράκτορας που εκπαίδευτηκε για όλους τους ασθενείς (μοντέλο 2) σε σχέση με τον πράκτορα που

εκπαιδεύτηκε σε έναν συγκεκριμένο ασθενή (μοντέλο 1). Τα αποτελέσματα φαίνονται στο σχήμα 5.5.



Εικόνα 5.5: Τα αποτελέσματα ελέγχου της γλυκόζης στον ασθενή adult4 για τυχαίο γεύμα με το μοντέλο 2 που εκπαιδεύτηκε ώστε να είναι γενικού σκοπού και με το εξειδικευμένο μοντέλο 1.

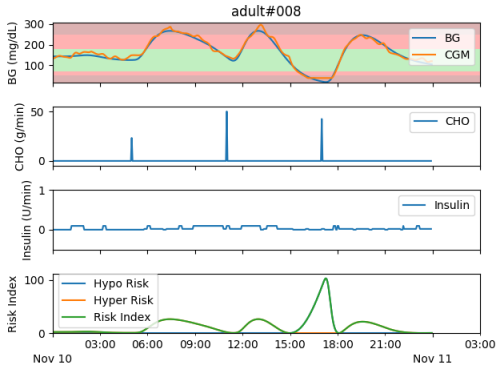
Το μοντέλο γενικού σκοπού (σχήμα 5.5) βλέπουμε ότι διορθώνει υπερβολικά την υπογλυκαιμία (ζώνη Upper C) και γι'αυτό εμφανίζει μεγάλο κίνδυνο υπεργλυκαιμίας και ποσοστό παραμονής σε αυτή. Δεν μπορεί να χρησιμοποιηθεί για να ελέγξει μεμονωμένα ασθενείς γιατί δεν μπορεί να προσαρμοσθεί στα ιδιαίτερα χαρακτηριστικά κάποιου.

Αντιθέτως το εξειδικευμένο μοντέλο κατάφερε να διατηρήσει τη γλυκόζη κατά περίπου 60% στα φυσιολογικά όρια, το οποίο είναι αποδεκτό ποσοστό. Ανήκει όμως στη ζώνη Lower D, πράγμα που σημαίνει ότι αποτυγχάνει να διαχειριστεί την υπογλυκαιμία αφού εμφανίζει και υψηλό κίνδυνο αυτής. Παρολαυτά ο κίνδυνος υπεργλυκαιμίας του είναι χαμηλός (<=4,5).

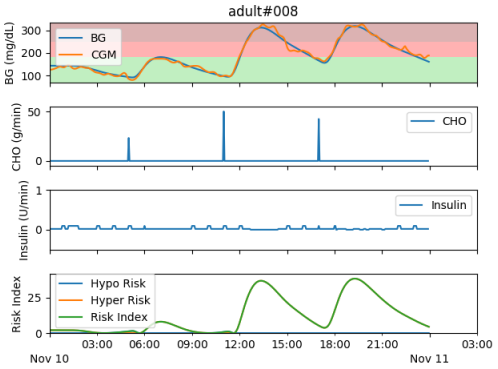
Στη συνέχεια δοκιμάσαμε τα δύο προηγούμενα μοντέλα σε τυχαίο ασθενή, διαφορετικό από τον ασθενή adult4 για τον οποίο εκπαιδεύτηκε το εξειδικευμένο μοντέλο. Επικεντρωθήκαμε στη γραφική παράσταση της διακύμανσης της γλυκόζης στο χρόνο (εικόνες 5.6 και 5.7) για να αξιολογήσουμε το αποτέλεσμα αντί των προηγούμενων γραφικών. Αυτή η τακτική δεν θα χρησιμοποιούταν στην πράξη και γι'αυτό τα γραφήματα αξιολόγησης της απόδοσης του συστήματος ελέγχου σε αυτή την περίπτωση επιλέξαμε να μην τα δείξουμε.

Παρατηρούμε ότι το μοντέλο το οποίο είχε εκπαιδευτεί με τυχαίους ασθενείς έχει καλύτερη συμπεριφορά σε αυτή την περίπτωση. Αυτό είναι κάτι που οφείλεται στο γεγονός ότι δεν έχει υπερεκπαιδευτεί (overfit) σε συγκεκριμένο ασθενή.

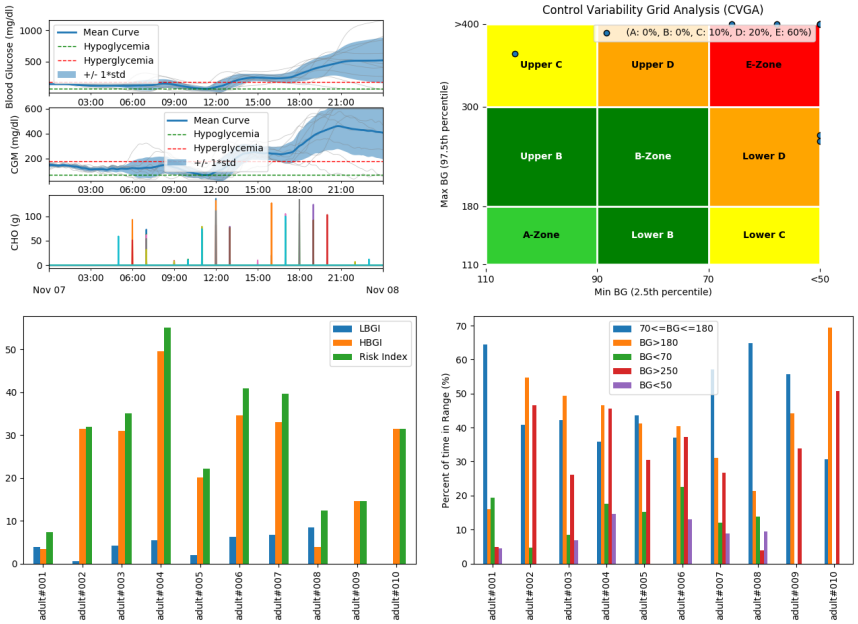
Τέλος είδαμε την απόδοση του γενικού ελεγκτή για όλους τους ενήλικες με τους οποίους εκπαιδεύτηκε (σχήμα 5.8).



Εικόνα 5.6: Οι διακυμάνσεις της γλυκόζης στο χρόνο του ασθενή adult8 για τυχαίο γεύμα με χρήση του μοντέλου 3.



Εικόνα 5.7: Οι διακυμάνσεις της γλυκόζης στο χρόνο του ασθενή adult4 για τυχαίο γεύμα με χρήση του μοντέλου 1.

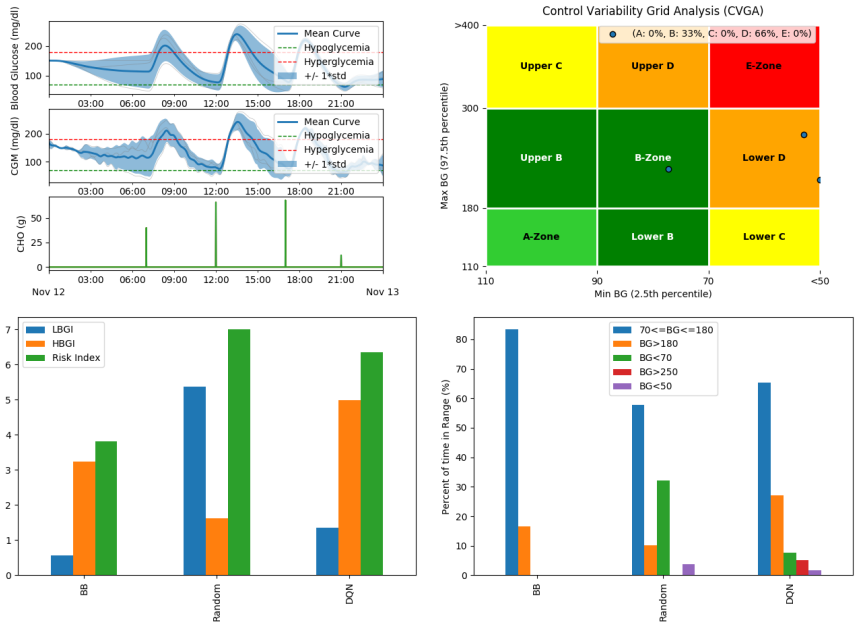


Εικόνα 5.8: Τα αποτελέσματα ελέγχου της γλυκόζης σε όλους τους ενήλικες ασθενείς για τυχαία γεύματα στον καθένα με το μοντέλο γενικού σκοπού.

Βλέπουμε ότι στη γενική περίπτωση δεν έχει σταθερή απόδοση, σε κάποιους ασθενείς τα πηγαίνει σχετικά καλά, σε άλλους αποτυγχάνει. Μόνο σε τέσσερις ασθενείς επιτυγχάνει διατήρηση της γλυκόζης σε φυσιολογικά όρια για ποσοστό μεγαλύτερο του 50% και από αυτούς οι δύο έχουν ποσοστό λίγο μεγαλύτερο τοου 60%. Γενικά παρουσιάζει υψηλούς κινδύνους και δεν καταφέρνει να αντιμετώπισειτη γλυκόζη σαν γενικό πρόβλημα. Αυτό οφείλεται στα διαφορετικά χαρακτηριστικά που παρουσιάζουν οι άνθρωποι μεταξύ τους.

5.4 Έλεγχος γλυκόζης για τυχαίο ασθενή και διάφορους ελεγκτές

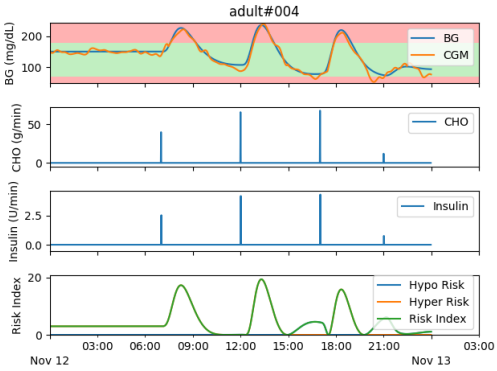
Τελικά συγκρίνουμε το εξειδικευμένα μοντέλο dqn (μοντέλο 1) που κατασκευάσαμε για τον adult4 με τον basal-bolus ελεγκτή του IDE και έναν ελεγκτή ο οποίος κάνει τυχαίες ενέργειες από το σύνολο A. Τα αποτελέσματα φαίνονται στα σχήματα 5.9, 5.10, 5.11, 5.12.



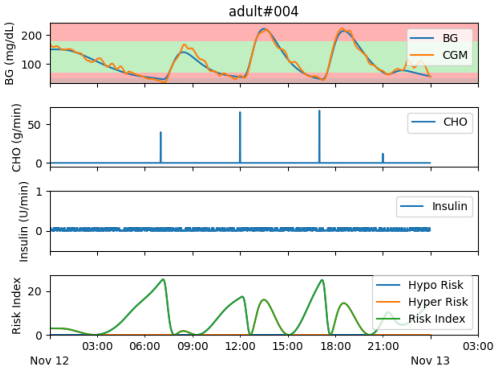
Εικόνα 5.9: Τα αποτελέσματα ελέγχου της γλυκόζης στον ασθενή adult4 για τυχαίο γεύμα με χρήση του basal-bolus ελεγκτή, ενός random ελεγκτή και του εξειδικευμένου dqn ελεγκτή.

Βλέπουμε ότι ο basal-bolus ελεγκτής κάνει καλύτερο γλυκαιμικό έλεγχο στον adult4, κρατώντας σε ποσοστό 80% τη γλυκόζη στη ζώνη στόχο 5.9. Έχει μηδαμινό κίνδυνο υπογλυκαιμίας και χαμηλό υπεργλυκαιμίας. Αλλά ο basal-bolus δεν λαμβάνει υπόψιν την καθυστέρηση δράσης της ινσουλίνης, δεν διαθέτει κάποιο είδος πρόβλεψεις ή εξειδίκευσης σε κάποιον ασθενή.

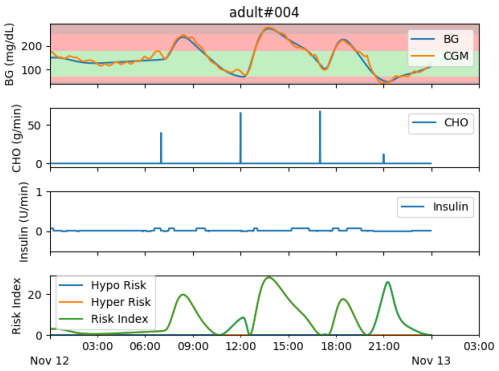
Ο εξεδικευμένος dqn ελεγκτής έχει πάλι παρόμοια συμπεριφορά όπως περιγράφηκε πριν, διατηρεί κατά 65% τη γλυκόζη στα αποδεκτά όρια, χαμηλό κίνδυνο υπογλυκαιμίας, μέτριο κίνδυνο υπεργλυκαιμίας και ανήκει στη ζώνη D. Ο τυχαίος ελεγκτής καταφέρνει επίσης να διατηρήσει τη γλυκόζη στα φυσιολογικά όρια για αρκετά μεγάλο ποσοστό αλλά σε αυτό συνεισφέρει το γεγονός ότι το σύνολο ενεργειών A δεν περιλαμβάνει μεγάλες ποσότητες ινσουλίνης σε ένα βήμα. Αυτό είναι σημαντική παρατήρηση για την ασφάλεια του συστήματος που θα μπορούσε να υλοποιηθεί με dqn ελεγκτή.



Εικόνα 5.10: Οι διακυμάνσεις της γλυκόζης στο χρόνο του ασθενή adult4 για τυχαίο γεύμα με χρήση του basal-bolus ελεγκτή.



Εικόνα 5.11: Οι διακυμάνσεις της γλυκόζης στο χρόνο του ασθενή adult4 για τυχαίο γεύμα με χρήση του random ελεγκτή.



Εικόνα 5.12: Οι διακυμάνσεις της γλυκόζης στο χρόνο του ασθενή adult4 για τυχαίο γεύμα με χρήση του dqn ελεγκτή

Παρατηρούμε στα σχήματα του χρόνου (5.10 και 5.12) ότι η γλυκόζη έχει παρόμοια διακύμανση για τους δύο αυτούς ελεγκτές. Αν ο dqn ελεγκτής είχε λάβει bolus δόσεις από τον ασθενή θα διαχειριζόταν καλύτερα την υπεργλυκαίμια.

5.5 Συζήτηση

Από τα παραπάνω αποτελέσματα συμπεραίνουμε ότι ο ελεγκτής του μοντέλου 1 ο οποίος κατά την εκπαίδευση επικεντρώθηκε πλήρως σε έναν ασθενή και ένα πλάνο γεύματος καταφέρνει συνολικά καλύτερο ελεγχό σε αυτόν ακόμα και για τυχαίο σενάριο γευμάτων μέσα στη μέρα. Επίσης το διάνυσμα καταστάσεων που χρησιμοποιήθηκε σε αυτόν, επειδή περιελάμβανε την τιμή ανάδρασης της ινσουλίνης, υποθέτουμε ότι έπαιξε καθοριστικό ρόλο στην απόδοση του και την ικανότητα του σύγκλισης.

Το μοντέλο 2 που είναι εκπαιδευμένο σε πολλούς ασθενείς με στόχο να χρησιμοποιηθεί σαν μοντέλο γενικού είναι υποδεέστερο του εξειδικευμένου μοντέλου. Αυτό είναι λογικό γιατί κάθε ασθενής έχει διαφορετικά χαρακτηριστικά και είναι δύσκολο να συγκλίνει σε αυτή την περίπτωση. Επίσης σε τυχαίο ασθενή είναι λογικό το γενικό μοντέλο να τα καταφέρνει καλύτερα από το εξει-

δικευμένο που εκπαιδεύτηκε με άλλο ασθενή, αφού ο καθένας έχει διαφορετική ευαισθησία και παραμέτρους. Το γενικό μοντέλο θα μπορούσε να είναι το πρωταρχικό μοντέλο ελέγχου το οποίο αργότερα μέσω εκπαίδευσης θα προσαρμοστεί σε κάποιον ασθενή.

Το μοντέλο 3 ενώ είχε υψηλά ποσοστά διατήρησης σε φυσιολογικά όρια εμφάνισε ταυτόχρονα υψηλούς κινδύνους υπογλυκαιμίας. Η συμπεριφορά του στην αρχή της ημέρας κάθε προσομείωσης ήταν πάρομοια με του μοντέλου 1 κάτι που το κάνει δυνητικά κατάλληλο μοντέλο, αλλά χρειάζεται περισσότερα πειράματα πάνω του και δοκιμή του διανύσματος κατάστασης του μοντέλου 1 για περαιτέρω βελτίωση.

Επίσης συμπεραίνουμε την ασφάλεια του συνόλου ενεργειών που επιλέξαμε. Κάποιο λάθος στις ρυθμίσεις σε σύστημα με αυτό το σύνολο ενεργειών δεν θα άλλαζε γρήγορα τα επίπεδα της γλυκόζης και θα προλάβαινε ο χρήστης να παρέμβει εγκαίρως.

Τέλος συμπεραίνουμε ότι το εξειδικευμένο μοντέλο, με περαιτέρω βελτιώσεις για να διαχειρίζεται την υπογλυκαιμία καλύτερα, θα μπορούσε να χρησιμοποιηθεί σε υβριδικό σύστημα ελέγχου όπως το Minimed που περιγράφηκε στο κεφάλαιο 1. Το basal-bolus μοντέλο έχει καλύτερη απόδοση σε κάποιες περιπτώσεις και αυτό οφείλεται στο ότι μπορεί να λάβει bolus δόσεις μεγαλύτερης ποσότητας σε ένα βήμα.

Μέρος V

Συμπεράσματα και Μελλοντικές Προοπτικές

Κεφάλαιο 6

Επίλογος

6.1 Συνοψη και Συμπεράσματα

Στη συγκεκριμένη διπλωματική εργασία μελετήθηκε το πρόβλημα ελέγχου της γλυκόζης στην περίπτωση του διαβήτη τύπου 1 με τη χρήση του αλγορίθμου ενισχυτικής μάθησης Deep Q-Network. Χρησιμοποιήθηκε ο προσομοιωτής T1D UVA/PADOVA S2008 για την εκπαίδευση και αξιολογήση τεσσάρων πρακτόρων, οι οποίοι διέφεραν ως προς τις λεπτομέρειες εκπαίδευσης όπως το σύνολο καταστάσεων και τα γεύματα.

Μέσω της πειραματικής αξιολόγησης που έγινε φάνηκε ότι ένας πράκτορας ενισχυτικής μάθησης είναι ικανός να διαχειριστεί το πρόβλημα του T1D. Μαθαίνει να διατηρεί τη γλυκόζη σε φυσιολογικά επίπεδα σε περιόδους νηστείας και κάνει ενέργειες για να μην ανεβεί υπερβολικά σε περιπτώσεις γεύματος.

Ο ελεγκτής που εκπαιδεύτηκε για έναν συγκεκριμένο ασθενή, με σταθερό σενάριο γευμάτων μέσα στη μέρα για κάθε επεισόδιο και διάνυσμα καταστάσεων που περιελάμβανε τη μέτρηση γλυκόζης από τον αισθητήρα CGM, τα γραμμάρια υδατανθρακών που καταναλώνονταν σε κάθε γεύμα και την προηγούμενη δόση ινσουλίνης είχε τα καλύτερα αποτελέσματα ελέγχου. Ο κίνδυνος υπογλυκαιμίας που παρατηρήσαμε με αυτόν είναι αρκετά μικρότερος κάτι το οποίο είναι ιδιαίτερα σημαντικό στον έλεγχο κατά τη διάρκεια της νύχτας. Ο συνολικός έλεγχος όμως δεν είναι επιτυχημένος. Αυτό οφείλεται σε εν μέρη στην αβεβαιότητα λήψης των γευμάτων και τον περιορισμό της μέγιστης δόσης ινσουλίνης που επιτρέψαμε να δίνει κάθε στιγμή.

Ένα σημείο το οποίο κρατάει πίσω την υλοποίηση μας όσον αφορά την αντιμετώπιση της αύξησης της γλυκόζης μετά από κάθε γεύμα είναι το σύνολο ενεργειών μας. Ο αλγόριθμος DQN λειτουργεί με διακριτό σύνολο και δεν μπορεί να υπολογίσει με ακρίβεια την δόση ινσουλίνης που χρειάζεται. Το σύνολο που χρησιμοποιήσαμε είναι ασφαλές, όπως μας έδειξε ο ελεγκτής τυχαίων ενεργειών, και γι'αυτό δεν παρουσιάζει μεγάλο κίνδυνο υπογλυκαιμίας το σύστημα. Σε περιπτώσεις όμως που χρειάζονται άμεσα μεγαλύτερες δόσεις (bolus) δεν μπορεί να ανταπεξέλθει. Από τη σύγκριση με τον basal-bolus ελεγκτή συμπεραίνουμε ότι αν υπήρχε η δυνατότητα εξωτερικής παρέμβασης και λήψης μιας bolus δόσης θα γινόταν καλύτερος έλεγχος της υπεργλυκαιμίας.

Επίσης επιλέξαμε διανύσματα καταστάσεων τα οποία να είναι ρεαλιστικά και να περιλαμβά-

νουν μόνο μετρήσεις που μπορούν να παρέχουν οι συσκευές. Αυτό κάνει πιο δύσκολο το φορμαλισμό του προβλήματος αλλά τα αποτελέσματα που παράγονται είναι περισσότερο στατιστικά σημαντικά από ότι σε πειράματα που θα περιελάμβαναν την πραγματική τιμή της γλυκόζης στο σήμα ανάδρασης. Συμπεράναμε ότι το διάνυσμα καταστάσεων που περιλάμβάνει τη γλυκόζη, τους υδατάνθρακες που περιέχονται στο γεύμα και την ινσουλίνη που δόθηκε την προηγούμενη στιγμή είναι ένα καλό διάνυσμα και μπορεί να χρησιμοποιηθεί για την εκπαίδευση πρακτόρων ενισχυτικής μάθησης. Επίσης η συνάρτηση ανταμοιβής που περιλαμβάνει τον δείκτη κινδύνου όπως περιγράφεται από την βιβλιογραφία είναι μια καλή συνάρτηση και συνίσταται η χρήση της.

Δοκιμάσαμε να κατασκευάσουμε ελεγκτή γενικού σκοπού και από τα αποτελέσματα συμπεράναμε ότι αυτό δεν είναι εφικτό αφού κάθε ασθενής εμφανίζει διαφορετική ευαισθησία στην ινσουλινή και έχει διαφορετικές παραμέτρους. Και ο ελεγκτής που εκπαιδεύτηκε με τυχαία γεύματα δεν μάθαινε το ίδιο καλά όσο ο εξειδικευμένος. Σε αυτούς έπαιξε ρόλο και το διάνυσμα καταστάσεων που χρησιμοποιήθηκε κατά την εκπαίδευση.

Από τη συνολική μας μελέτη καθίσταται σαφές ότι το πρόβλημα του T1D είναι ένα απαιτητικό πρόβλημα για την ενισχυτική μάθηση. Ο πρώτος λόγος είναι ότι το περιβάλλον είναι μερικώς παρατηρήσιμο, καθώς ο πράκτορας έχει πρόσβαση σε μια θορυβώδη μέτρηση (CGM) της πραγματικής κατάστασης του περιβάλλοντος (BG). Ο δεύτερος λόγος έχει να κάνει με το ίδιο το μοντέλο γλυκόζης-ινσουλίνης το οποίο είναι περίπλοκο. Το γεγονός ότι περιλαμβάνει μη γραμμικές μεταβλητές καθιστά δύσκολη τη διαδικασία μοντελοποίησης και την πρόβλεψη πάνω του. Αυτό σε συνδυασμό με τα χαρακτηριστικά της κινητικότητα της ινσουλίνης από τον υποδόριο χώρο στο αίμα, τους χρόνους δράσης της φαρμακευτικής ινσουλίνης και εισάγουν μεγαλύτερη πολυπλοκότητα στο σύστημα. Οι ενέργειες δεν έχουν άμεση επίδραση στο περιβάλλον και άρα είναι δύσκολο να αποδιδεται σε κάθε ενέργεια η κατάλληλη ανταμοιβή (credit-assignment problem).

6.2 Μελλοντικές Επεκτάσεις

Οι αλγόριθμοι ενισχυτικής μάθησης μπορεί να είναι η λύση στο πρόβλημα διαχείρισης της γλυκόζης σε ανθρώπους με διαβήτη τύπου 1. Μέχρι τώρα έχουν γίνει και γίνονται πολλές έρευνες πάνω στο θέμα των αλγορίθμων ελέγχου και ο τομέας της ενισχυτικής μάθησης αξίζει να ενταχθεί σε αυτές.

Μια συνεισφορά της συγκεκριμένης διπλωματικής είναι η αναζήτηση και η αποτίμηση των στοιχείων του προβλήματος, (ορισμός κατάλληλων καταστάσεων, ενεργειών και ανταμοιβών), που μπορούν να χρησιμοποιηθούν από τους αλγορίθμους της ενισχυτικη μαθησης για αποτελεσματικό έλεγχο. Χρειάζεται ακόμα έρευνα πάνω σε αυτό το κομμάτι ώστε να μπορέσει να υπάρξει βέλτιστη απόδοση στη διαχείριση του περιβάλλοντος. Άλλα σύνολα ενεργειών, τα οποία να μην περιορίζονται σε τρεις διακριτές δόσεις μπορεί να έχουν καλύτερα αποτελέσματα. Επίσης άλλα διανύσματα κατάστασης που να περιλάμβάνουν και μετρήσεις όπως ο καρδιακός παλμός θα μπορούσαν να βελτιώσουν την απόδοση σε περιπτώσεις άγχους, άθλησης, ασθένειας.

Σημαντική βελτίωση θα μπορούσαν να φέρουν αλγόριθμοι ενισχυτικής μάθησης με συνεχές σύνολο ενεργειών και διαφορετικού τρόπους προσέγγισης των βέλτιστων Q-συναρτήσεων. Ο το-

μέας της ενισχυτικής μάθησης περιλαμβάνει πλήθος μεθόδων και αλγορίθμων που θα μπορούσαν να δοκιμαστούν στο πρόβλημα του T1D και να προσφέρουν νέες πληροφορίες πάνω σε αυτό.

Ιδανικό σενάριο για το σύστημα κλειστού ελέγχου είναι να εξαφανιστεί τελείως η παρέμβαση από τον χρήστη μέσω της αναγγελίας γευμάτων και λήψης bolus δόσης. Αυτό θα αφαιρέσει τον παράγοντα του ανθρώπινου λάθους και τελικά το τεχνητό πάγκρεας θα προσεγγίσει περισσότερο τη λειτουργία του φυσιολογικού. Κάτι τέτοιο όμως χωρίς ακρίβεια στις μετρήσεις και άμεση επίδραση των ενεργειών δεν μπορεί να γίνει εύκολα.

Η βελτίωση του μοντέλου γλυκόζης-ινσουλίνης οδηγεί στην ανάπτυξη καλύτερων προσομοιωτών και την παραγωγή καλύτερα εκπαιδευμένων πρακτόρων. Επίσης οι βελτιωμένοι αισθητήρες, που θα έχουν μεγαλύτερη ακρίβεια και η παραγωγή φαρμακευτικής ινσουλίνης με καλύτερα χαρακτηριστικά κινητικότητας, δηλαδή γρηγορότερη δράση θα επιτρέψουν στους πράκτορες να λαμβάνουν καλύτερες αποφάσεις, με άμεσες επιδράσεις και θα βελτιώσουν την πρόβλεψη που κάνουν περιβάλλον.

Βιβλιογραφία

[1] American Diabetes Association et al. Diagnosis and classification of diabetes mellitus. *Diabetes care*, 33(Supplement 1):S62–S69, 2010.

[2] Eric P. Widmaier, Hershel Raff, and Kevin T. Strang. *Vander's Human Physiology, The Mechanisms of Body Function*. McGraw-Hill, 10th edition, 2010.

[3] Ignazio Vecchio, Cristina Tornali, Nicola Luigi Bragazzi, and Mariano Martini. The discovery of insulin: an important milestone in the history of medicine. *Frontiers in endocrinology*, 9:613, 2018.

[4] Grazia Aleppo. Insulin pump overview: How insulin pumps work, who benefits from them, and different types of pumps. URL `https://www.endocrineweb.com/guides/insulin/insulin-pump-overview`.

[5] Hanna S Mariani, Brian T Layden, and Grazia Aleppo. Continuous glucose monitoring: a perspective on its past, present, and future applications for diabetes management. *Clinical Diabetes*, 35(1):60–65, 2017.

[6] Charlotte K. Boughton and Roman Hovorka. Advances in artificial pancreas systems. *Science Translational Medicine*, 11(484), 2019. ISSN 1946-6234. doi: 10.1126/scitranslmed.aaw4949. URL `https://stm.sciencemag.org/content/11/484/eaaw4949`.

[7] G. Bruttomesso, D.; Grassi. *[Frontiers in Diabetes] Technological Advances in the Treatment of Type 1 Diabetes Volume 24 () || Artificial Pancreas: A Review of Fundamentals and Inpatient and Outpatient Studies*. 2014. ISBN 978-3-318-02336-7,978-3-318-02337-4. doi: 10.1159/000363512. URL `http://gen.lib.rus.ec/scimag/index.php?s=10.1159/000363512`.

[8] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[9] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[10] Boris P Kovatchev, Marc Breton, Chiara Dalla Man, and Claudio Cobelli. In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes, 2009.

[11] Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The uva/padova type 1 diabetes simulator: new features. *Journal of diabetes science and technology*, 8(1):26–34, 2014.

[12] Roberto Visentin, Enrique Campos-Náñez, Michele Schiavon, Dayu Lv, Martina Vettoretti, Marc Breton, Boris P Kovatchev, Chiara Dalla Man, and Claudio Cobelli. The uva/padova type 1 diabetes simulator goes from single meal to single day. *Journal of diabetes science and technology*, 12(2):273–281, 2018.

[13] William Clarke and Boris Kovatchev. Statistical tools to analyze continuous glucose monitor data. *Diabetes technology & therapeutics*, 11(S1):S–45, 2009.

[14] Jinyu Xie. Simglucose v0.2.1 (2018) [online]. URL `https://github.com/jxx123/simglucose`.

[15] Ian Fox and Jenna Wiens. Reinforcement learning for blood glucose control: Challenges and opportunities. 2019.