Νέοι Αλγόριθμοι Εκπαίδευσης Τεχνητών Νευρωνικών Δικτύων και Εφαρμογές

Αριστοτέλης Ε. Κωστόπουλος

Διδακτορική Διατριβή

Πανεπιστήμιο Πατρών
Σχολή Θετικών Επιστημών
Τμήμα Μαθηματικών
Πάτρα 2011

Στους γονείς μου και τη
σύζυγό μου που με
πιστεύουν και στηρίζουν

Περίληψη

Η παρούσα διδακτορική διατριβή πραγματεύεται το θέμα της εκπαίδευσης εμπρόσθιων τροφοδοτούμενων τεχνητών νευρωνικών δικτύων και τις εφαρμογές τους. Η παρουσίαση των θεμάτων και των αποτελεσμάτων της διατριβής οργανώνεται ως εξής:

Στο Κεφάλαιο 1 παρουσιάζονται τα τεχνητά νευρωνικά δίκτυα, τα οφέλη της χρήσης τους, η δομή και η λειτουργία τους. Πιο συγκεκριμένα, παρουσιάζεται πως από τους βιολογικούς νευρώνες μοντελοποιούνται οι τεχνητοί νευρώνες, που αποτελούν το θεμελιώδες στοιχείο των τεχνητών νευρωνικών δικτύων. Στη συνέχεια αναφέρονται οι βασικές αρχιτεκτονικές των εμπρόσθιων τροφοδοτούμενων τεχνητών νευρωνικών δικτύων. Το κεφάλαιο ολοκληρώνεται με μια ιστορική αναδρομή για τα τεχνητά νευρωνικά δίκτυα και με την παρουσίαση κάποιων εφαρμογών τους.

Στο Κεφάλαιο 2 παρουσιάζονται μερικοί από τους υπάρχοντες αλγόριθμους εκπαίδευσης τεχνητών νευρωνικών δικτύων. Γίνεται μια περιληπτική αναφορά του προβλήματος της εκπαίδευσης των τεχνητών νευρωνικών δικτύων με επίβλεψη και δίνεται η μαθηματική του μοντελοποίηση που αντιστοιχεί στην ελαχιστοποίηση μιας συνάρτησης κόστους. Στην συνέχεια γίνεται μια περιληπτική αναφορά στις μεθόδους που βασίζονται στην κατεύθυνση της πιο απότομης καθόδου, στις μεθόδους δευτέρας τάξεως όπου απαιτείται ο υπολογισμός του Εσσιανού πίνακα της συνάρτησης κόστους, στις μεθόδους μεταβλητής μετρικής, και στις μεθόδους συζυγών κλίσεων. Κατόπιν, παρουσιάζεται ο χώρος των βαρών, η επιφάνεια σφάλματος και οι διάφορες τεχνικές αρχικοποίησης των βαρών των τεχνητών νευρωνικών δικτύων και περιγράφονται οι επιπτώσεις που έχουν στην εκπαίδευση τους.

Στο κεφάλαιο 3 παρουσιάζεται ένας νέος αλγόριθμος εκπαίδευσης τεχνητών νευρωνικών δικτύων βασισμένος στον αλγόριθμο της οπισθοδρομικής διάδοσης του σφάλματος και στην αυτόματη προσαρμογή του ρυθμού εκπαίδευσης χρησιμοποιώντας πληροφορία δύο σημείων [81]. Η κατεύθυνση αναζήτησης του νέου αλγορίθμου είναι η κατεύθυνση της πιο απότομης καθόδου, αλλά για τον προσδιορισμό του ρυθμού εκπαίδευσης χρησιμοποιούνται προσεγγίσεις δύο σημείων της εξίσωσης χορδής των μεθόδων ψευδό-Newton [5]. Επιπλέον, παράγεται ένας νέος ρυθμός εκπαίδευσης προσεγγίζοντας την νέα εξίσωση χορδής, που προτάθηκε από τον Zhang [90], η οποία χρησιμοποιεί πληροφορία παραγώγων και συναρτησιακών τιμών. Στην συνέχεια, ένας κατάλληλος μηχανισμός

επιλογής του ρυθμού εκπαίδευσης ενσωματώνεται στον αλγόριθμο εκπαίδευσης ώστε να επιλέγεται κάθε φορά ο κατάλληλος ρυθμός εκπαίδευσης. Τέλος, γίνεται μελέτη της σύγκλισης του αλγορίθμου εκπαίδευσης και παρουσιάζονται τα πειραματικά αποτελέσματα για διάφορα προβλήματα εκπαίδευσης.

Στο Κεφάλαιο 4 παρουσιάζονται μερικοί αποτελεσματικοί αλγόριθμοι εκπαίδευσης οι οποίοι βασίζονται στις μεθόδους βελτιστοποίησης συζυγών κλίσεων [43]. Στους υπάρχοντες αλγόριθμους εκπαίδευσης συζυγών κλίσεων προστίθεται ένας νέος αλγόριθμος εκπαίδευσης που βασίζεται στην μέθοδο συζυγών κλίσεων του Perry [60]. Επιπρόσθετα, προτείνονται νέοι αλγόριθμοι συζυγών κλίσεων που προκύπτουν από τις ίδιες αρχές που προέρχονται οι γνωστοί αλγόριθμοι συζυγών κλίσεων των Hestenes-Stiefel, Fletcher-Reeves, Polak-Ribière και Perry, και ονομάζονται κλιμακωτοί αλγόριθμοι συζυγών κλίσεων. Αυτή η κατηγορία αλγορίθμων βασίζεται στη φασματική παράμετρο κλιμακώσης που προτάθηκε από τους Barzilai και Borwein [5]. Επιπλέον, ενσωματώνεται στους αλγόριθμους εκπαίδευσης συζυγών κλίσεων μια αποδοτική τεχνική γραμμικής αναζήτησης, που βασίζεται στις συνθήκες του Wolfe και στην διασφαλισμένη κυβική παρεμβολή [77]. Ακόμη, η παράμετρος του αρχικού ρυθμού εκπαίδευσης προσαρμόζεται αυτόματα σε κάθε επανάληψη σύμφωνα με ένα κλειστό τύπο που προτάθηκε στις εργασίες [77] και [80]. Στη συνέχεια, εφαρμόζεται μια αποτελεσματική διαδικασία επανεκκίνησης, έτσι ώστε να βελτιωθούν περαιτέρω οι αλγόριθμοι εκπαίδευσης συζυγών κλίσεων και να αποδειχθεί η ολική τους σύγκλιση. Τέλος, παρουσιάζονται τα πειραματικά αποτελέσματα για διάφορα προβλήματα εκπαίδευσης.

Στο τελευταίο Κεφάλαιο της παρούσας διδακτορικής διατριβής, απομονώνεται και τροποποιείται ο κλιμακωτός αλγόριθμος του Perry, που παρουσιάστηκε στο προηγούμενο κεφάλαιο. Πιο συγκεκριμένα, ενώ διατηρούνται τα κύρια χαρακτηριστικά του αλγορίθμου εκπαίδευσης, εφαρμόζεται μια διαφορετική τεχνική γραμμικής αναζήτησης η οποία βασίζεται στις μη μονότονες συνθήκες του Wolfe. Επίσης, προτείνεται ένας νέος αρχικός ρυθμός εκπαίδευσης ([41]) για χρήση με τον κλιμακωτό αλγόριθμο εκπαίδευσης συζυγών κλίσεων, ο οποίος φαίνεται να είναι αποδοτικότερος από τον αρχικό ρυθμό εκπαίδευσης που προτάθηκε από τον Shanno [77] όταν χρησιμοποιείται σε συνδυασμό με την μη μονότονη τεχνική γραμμικής αναζήτησης. Στη συνέχεια παρουσιάζονται τα πειραματικά αποτελέσματα για διάφορα προβλήματα εκπαίδευσης. Τέλος, ως εφαρμογή εκπαιδεύεται ένα πολυεπίπεδο εμπρόσθια τροφοδοτούμενο τεχνητό νευρωνικό δίκτυο με τον προτεινόμενο αλγόριθμο για το πρόβλημα της ταξινόμησης καρκινικών κυττάρων του εγκεφάλου και συγκρίνεται η απόδοση του με την απόδοση ενός πιθανοτικού τεχνητού νευρωνικού δικτύου [42].

Η διατριβή ολοκληρώνεται με το Παράρτημα Α', όπου παρουσιάζονται τα προβλήματα εκπαίδευσης τεχνητών νευρωνικών δικτύων που χρησιμοποιήθηκαν για την αξιολόγηση των προτεινόμενων αλγορίθμων εκπαίδευσης.

Synopsis

In this dissertation the problem of the training of feedforward artificial neural networks and its applications are considered. The presentation of the topics and the results are organized as follows:

In the first chapter, the artificial neural networks are introduced. Initially, the benefits of the use of artificial neural networks are presented. In the sequence, the structure and their functionality are presented. More specifically, the derivation of the artificial neurons from the biological ones is presented followed by the presentation of the architecture of the feedforward neural networks. The historical notes and the use of neural networks in real world problems are concluding the first chapter.

In Chapter 2, the existing training algorithms for the feedforward neural networks are considered. First, a summary of the training problem and its mathematical formulation, that corresponds to the unconstrained minimization of a cost function, are given. In the sequence, training algorithms based on the steepest descent, Newton, variable metric and conjugate gradient methods are presented. Furthermore, the weight space, the error surface and the techniques of the initialization of the weights are described. Their influence in the training procedure is discussed.

In Chapter 3, a new training algorithm for feedforward neural networks based on the backpropagation algorithm and the automatic two-point step size (learning rate) is presented [81]. The algorithm uses the steepest descent search direction while the learning rate parameter is calculated by minimizing the standard secant equation [5]. Furthermore, a new learning rate parameter is derived by minimizing the modified secant equation introduced in [90], that uses both gradient and function value information. In the sequence a switching mechanism is incorporated into the algorithm so that the appropriate stepsize to be chosen according to the status of the current iterative point. Finally, the global convergence of the proposed algorithm is studied and the results of some numerical experiments are presented.

In Chapter 4, some efficient training algorithms, based on conjugate gradient optimization methods, are presented [43]. In addition to the existing conjugate gradient training algorithms, we introduce Perry's conjugate gradient method as a training algorithm [60]. Furthermore, a new class of conjugate gradient methods is proposed,

called self-scaled conjugate gradient methods, which are derived from the principles of Hestenes-Stiefel, Fletcher-Reeves, Polak-Ribière and Perry’s method. This class is based on the spectral scaling parameter introduced in [5]. Furthermore, we incorporate to the conjugate gradient training algorithms an efficient line search technique based on the Wolfe conditions and on safeguarded cubic interpolation [77]. In addition, the initial learning rate parameter, fed to the line search technique, was automatically adapted at each iteration by a closed formula proposed in [77] and [80]. Finally, an efficient restarting procedure was employed in order to further improve the effectiveness of the conjugate gradient training algorithms and prove their global convergence. Experimental results show that, in general, the new class of methods can perform better with a much lower computational cost and better success performance.

In the last chapter of this dissertation, the Perry’s self-scaled conjugate gradient training algorithm that was presented in the previous chapter was isolated and modified. More specifically, the main characteristics of the training algorithm were maintained but in this case a different line search strategy based on the nonmonotone Wolfe conditions was utilized. Furthermore, a new initial learning rate parameter was introduced ([41]) for use in conjunction with the self-scaled conjugate gradient training algorithm that seems to be more effective from the initial learning rate parameter, proposed by Shanno in [77], when used with the nonmonotone line search technique. In the sequence the experimental results for different training problems are presented. Finally, a feedforward neural network with the proposed algorithm for the problem of brain astrocytomas grading was trained and compared the results with those achieved by a probabilistic neural network [42].

The dissertation is concluded with the Appendix A’, where the training problems used for the evaluation of the proposed training algorithms are presented.

Ευχαριστίες

Η παρούσα διδακτορική διατριβή θα ήταν αδύνατο να εκπονηθεί χωρίς την καθοριστική βοήθεια και συμπαράσταση πολλών ανθρώπων. Θα ήθελα καταρχήν να ευχαριστήσω θερμά την επιβλέπουσα της διδακτορικής διατριβής, επίκουρο καθηγήτρια κα. Θεοδούλα Γράψα. Η παρουσία της ήταν καθοριστική σε όλη τη διάρκεια φοίτησής μου στο Τμήμα Μαθηματικών του Πανεπιστημίου Πατρών. Η καθοδήγησή της ήταν ουσιαστική και με βοήθησε να ξεπεράσω τις διάφορες δυσκολίες που αντιμετώπισα κατά την έρευνα και συγγραφή της διδακτορικής διατριβής. Θα ήθελα, επίσης, να ευχαριστήσω και τα υπόλοιπα μέλη της τριμελούς συμβουλευτικής επιτροπής, τους καθηγητές κ.κ. Κοσμά Ιορδανίδη και Χαράλαμπο Μπότσαρη για την βοήθεια που μου προσέφεραν για την ολοκλήρωση της διατριβής αυτής, αλλά και για τις πολύτιμες γνώσεις που μου προσέφεραν κατά την διάρκεια της φοίτησής μου.

Στο σημείο αυτό, θα ήθελα να ευχαριστήσω την καθηγήτρια και πρύτανη του Tampere University of Technology κα. Ulla Ruotsalainen, τον διδάκτορα κ. Antti Happonen, αλλά και τα υπόλοιπα μέλη της ερευνητικής ομάδας M^2oBSI , με τους οποίους είχα την τιμή να συνεργαστώ σε θέματα επεξεργασίας σημάτων στα πλαίσια της υποτροφίας Marie Curie της Ευρωπαϊκής Ένωσης. Επίσης, θα ήθελα να ευχαριστήσω για την ανεκτίμητη βοήθεια που μου προσέφεραν τους υπόλοιπους συνεργάτες μου, κ.κ. Ιωάννη Νίκα, Γεράσιμο Αντζουλάτο, Χριστίνα Νικολακάκου, Αριστοτέλη Κλαμαργιά, Δημήτριο Γκλότσο και Στυλιανό Δήμα.

Επιπλέον, θα ήταν παράλειψή μου να μην ευχαριστήσω την διευθύντρια της Βιβλιοθήκης και Κέντρου Πληροφόρησης του Πανεπιστημίου Πατρών κα. Αικατερίνη Συνέλλη αλλά και το υπόλοιπο προσωπικό για την βοήθεια και την στήριξη τους κατά την διάρκεια της φοίτησής μου. Θα ήθελα, επίσης, να απευθύνω ευχαριστίες και στο προσωπικό του Εργαστηρίου Ηλεκτρονικών Υπολογιστών και Πληροφορικής του Τμήματος Μαθηματικών για τη διάθεση των απαραίτητων υπολογιστικών πόρων.

Τέλος, θα ήθελα να ευχαριστήσω θερμά τους γονείς μου, την αδερφή μου και τη σύζυγό μου οι οποίοι με στήριζαν με όλη τους τη δύναμη και όλα τα μέσα κατά τη διάρκεια των σπουδών μου.

Περιεχόμενα

I	ΕΙΣΑΓΩΓΗ	19
Κεφάλαιο 1	Τεχνητά Νευρωνικά Δίκτυα	21
1.1	Οφέλη των Νευρωνικών Δικτύων	22
1.2	Βιολογικοί Νευρώνες	24
1.3	Τεχνητοί Νευρώνες	26
1.3.1	Τύποι Συναρτήσεων Ενεργοποίησης	30
1.4	Αρχιτεκτονικές Δικτύων	33
1.4.1	Μονοεπίπεδα Εμπρόσθια Τροφοδοτούμενα Δίκτυα	33
1.4.2	Πολυεπίπεδα Εμπρόσθια Τροφοδοτούμενα Δίκτυα	34
1.5	Η Ιστορία των Τεχνητών Νευρωνικών Δικτύων	36
1.6	Εφαρμογές των Τεχνητών Νευρωνικών Δικτύων	38
Κεφάλαιο 2	Εκπαίδευση των Τεχνητών Νευρωνικών Δικτύων	41
2.1	Μέθοδοι Εκπαίδευσης	42
2.2	Επιφάνεια Σφάλματος και ο Χώρος των Βαρών	47
2.2.1	Αρχικοποίηση των Βαρών	49

II ΝΕΟΙ ΑΛΓΟΡΙΘΜΟΙ ΕΚΠΑΙΔΕΥΣΗΣ ΤΝΔ 53

Κεφάλαιο 3	Αλγόριθμος Εκπαίδευσης Προσέγγισης Δύο Σημείων	55
3.1	Η μέθοδος των Barzilai και Borwein	55
3.2	Νέα Έξισωση Χορδής και Παραγωγή Νέων Βημάτων	59
3.3	Νέος Αλγόριθμος Εκπαίδευσης και η Ολική Σύγκλιση του	62
3.4	Πειραματικά Αποτελέσματα	65
3.4.1	Αποκλειστικό-EITE	67
3.4.2	Ισοτιμία των 3-bit	68
3.4.3	Αναγνώριση Κεφαλαίων Γραμμάτων	69
3.4.4	Προσέγγιση Συνεχούς Τριγωνομετρικής Συνάρτησης	70
3.5	Συμπεράσματα	70
Κεφάλαιο 4	Νέα Οικογένεια Αλγορίθμων Εκπαίδευσης Συζυγών Κλίσεων	73
4.1	Αλγόριθμοι Εκπαίδευσης Συζυγών Κλίσεων	74
4.2	Κλιμακωτοί Αλγόριθμοι Εκπαίδευσης Συζυγών Κλίσεων	76
4.2.1	Παράμετρος Κλιμάκωσης	77
4.3	Γενικός Αλγόριθμος Εκπαίδευσης	78
4.3.1	Διαδικασίες Επανεκκίνησης	79
4.3.2	Διαδικασίες Γραμμικής Αναζήτησης	80
4.3.3	Αρχικός Ρυθμός Εκπαίδευσης	82
4.3.4	Περιγραφή του Αλγορίθμου Εκπαίδευσης	82
4.3.5	Ολική Σύγκλιση	84
4.4	Πειραματικά Αποτελέσματα	86
4.4.1	Αποκλειστικό-EITE	88
4.4.2	Ισοτιμία των 3-bit	90

4.4.3	$N - M - N$ Κωδικοποιητής/Αποκωδικοποιητής	92
4.4.4	Αναγνώριση Κεφαλαίων Γραμμάτων	102
4.4.5	Αναγνώριση Αριθμών	104
4.4.6	Ταξινόμηση των Φυτών της Οικογένειας Iris	106
4.4.7	Οπτική Αναγνώριση Χειρόγραφων Αριθμών	108
4.4.8	Αναγνώριση Φωνηέντων	110
4.5	Συμπεράσματα	112
Κεφάλαιο 5 Μη Μονότονοι Αλγόριθμοι Εκπαίδευσης Συζυγών Κλίσεων		115
5.1	Ο Αλγόριθμος του Perry	115
5.2	Μη Μονότονος Κλιμακωτός Αλγόριθμος Εκπαίδευσης του Perry	116
5.3	Πειραματικά Αποτελέσματα	121
5.3.1	Αποκλειστικό-EITE	122
5.3.2	Ισοτιμία των 3-bit	123
5.3.3	Αναγνώριση Κεφαλαίων Γραμμάτων	124
5.3.4	Προσέγγιση Συνεχούς Τριγωνομετρικής Συνάρτησης	125
5.3.5	Ταξινόμηση Καρκινικών Κυττάρων Εγκεφάλου	126
5.4	Συμπεράσματα	129

III ΠΑΡΑΡΤΗΜΑΤΑ—ΒΙΒΛΙΟΓΡΑΦΙΑ

131

Κεφάλαιο Α' Προβλήματα Εκπαίδευσης Νευρωνικών Δικτύων		133
A'.1	Αποκλειστικό-EITE (XOR)	133
A'.2	Ισοτιμία των 3-bit	134
A'.3	Αναγνώριση Κεφαλαίων Γραμμάτων	135
A'.4	Προσέγγιση Συνεχούς Τριγωνομετρικής Συνάρτησης	136

A'.5	$N - M - N$ Κωδικοποιητής/Αποκωδικοποιητής	136
A'.6	Αναγνώριση Αριθμών	137
A'.7	Ταξινόμηση των Φυτών της Οικογένειας Iris	138
A'.8	Οπτική Αναγνώριση Χειρόγραφων Αριθμών	139
A'.8.1	Αναγνώριση Φωνηέντων	140
Κεφάλαιο Β'	Δημοσιεύσεις - Αναφορές	141

Κατάλογος Σχημάτων

1.1	Σχέδιο ενός βιολογικού νευρώνα.	25
1.2	Μετάδοση σήματος προσυναπτικού άξονα μέσω συναπτικής σχισμής στον μετασυναπτικό νευρώνα.	26
1.3	Μη γραμμικό μοντέλο ενός νευρώνα.	27
1.4	Ομοπαράλληλικός μετασχηματισμός παραγόμενος από την παρουσία μεροληψίας. Σημειώνεται ότι $v_k = b_k$ στο $u_k = 0$	28
1.5	Ένα άλλο μη γραμμικό μοντέλο νευρώνα.	30
1.6	Συνάρτηση Κατωφλιού.	31
1.7	Κατά Τμήματα Γραμμική Συνάρτηση.	32
1.8	Σιγμοειδής Συνάρτηση.	33
1.9	Εμπρόσθια τροφοδοτούμενο δίκτυο με ένα μόνο επίπεδο από νευρώνες.	34
1.10	Πλήρως διασυνδεδεμένο εμπρόσθια τροφοδοτούμενο δίκτυο με ένα κρυφό επίπεδο και ένα επίπεδο εξόδου.	35
2.1	Η επιφάνεια της συνάρτησης σφάλματος πάνω στον χώρο των βαρών σε ένα νευρωνικό δίκτυο με ένα μόνο νευρώνα.	47
2.2	Η επιφάνεια της συνάρτησης σφάλματος πάνω στον χώρο των βαρών σε ένα νευρωνικό δίκτυο με ένα μόνο νευρώνα υπό την επίδραση: α. ενός προτύπου, β. δύο προτύπων, γ. τριών προτύπων και δ. τεσσάρων προτύπων.	48
2.3	Ισοϋφείς της επιφάνειας της συνάρτησης σφάλματος και τροχιά του αλγόριθμου εκπαίδευσης. Αριστερά ο αλγόριθμος συγκλίνει σε ένα τοπικό ελάχιστο, ενώ δεξιά ο αλγόριθμος συγκλίνει στο ολικό ελάχιστο.	49
4.1	Μέσος χρόνος για το πρόβλημα του αποκλειστικού-EITE	89

4.2	Καμπύλες εκπαίδευσης των αλγορίθμων εκπαίδευσης για το πρόβλημα του αποκλειστικού-EITE	89
4.3	Μέσος χρόνος για το πρόβλημα της ισοτιμίας των 3-bit	91
4.4	Καμπύλες εκπαίδευσης των αλγορίθμων εκπαίδευσης για το πρόβλημα της ισοτιμίας των 3-bit	91
4.5	Μέσος χρόνος για το πρόβλημα του 4 – 2 – 4 Κωδικοποιητή / Αποκωδικοποιητή	94
4.6	Καμπύλες εκπαίδευσης των αλγορίθμων εκπαίδευσης για το πρόβλημα του 4 – 2 – 4 Κωδικοποιητή / Αποκωδικοποιητή	95
4.7	Μέσος χρόνος για το πρόβλημα του 8 – 3 – 8 Κωδικοποιητή / Αποκωδικοποιητή	96
4.8	Καμπύλες εκπαίδευσης των αλγορίθμων εκπαίδευσης για το πρόβλημα του 8 – 3 – 8 Κωδικοποιητή / Αποκωδικοποιητή	96
4.9	Μέσος χρόνος για το πρόβλημα του 16 – 4 – 16 Κωδικοποιητή / Αποκωδικοποιητή	97
4.10	Μέσος χρόνος για το πρόβλημα του 32 – 5 – 32 Κωδικοποιητή / Αποκωδικοποιητή	98
4.11	Μέσος χρόνος για το πρόβλημα του 64 – 6 – 64 Κωδικοποιητή / Αποκωδικοποιητή	99
4.12	Μέσος χρόνος για το πρόβλημα του 128 – 7 – 128 Κωδικοποιητή / Αποκωδικοποιητή	100
4.13	Μέσος χρόνος για το πρόβλημα του 256 – 8 – 256 Κωδικοποιητή / Αποκωδικοποιητή	101
4.14	Μέσος χρόνος για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων	102
4.15	Καμπύλες εκπαίδευσης των αλγορίθμων εκπαίδευσης για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων	103
4.16	Μέσος χρόνος για το πρόβλημα της αναγνώρισης αριθμών	105
4.17	Καμπύλες εκπαίδευσης των αλγορίθμων εκπαίδευσης για το πρόβλημα της αναγνώρισης αριθμών	105
4.18	Μέσος χρόνος για το πρόβλημα της ταξινόμησης των φυτών της οικογένειας Iris	107

4.19 Μέσος χρόνος για το πρόβλημα της οπτικής αναγνώρισης χειρόγραφων αριθμών	109
4.20 Μέσος χρόνος για το πρόβλημα της αναγνώρισης φωνηέντων	111
5.1 Παραδείγματα χαμηλού (επάνω) και υψηλού βαθμού (κάτω) εγκεφαλικών αστροκυτωμάτων μαζί με την τμηματοποιημένη εικόνα.	127
5.2 Καλύτερες επιδόσεις για το εμπρόσθια τροφοδοτούμενο νευρωνικό δίκτυο για διαφορετικό αριθμό νευρώνων στο κρυφό επίπεδο.	130
A'.1 Το πρόβλημα του Αποκλειστικού-EITE στο επίπεδο. Από το σχήμα φαίνεται καθαρά ότι τα πρότυπα εκπαίδευσης δεν είναι γραμμικά διαχωρίσιμα.	134
A'.2 Η κωδικοποίηση των γραμμάτων A, B, C και D και οι αντίστοιχες δυαδικές κωδικοποιήσεις.	135
A'.3 Η γραφική παράσταση της $f(x) = \sin(x) \cos(2x)$ στο $[0, 2\pi]$. Τα 20 πρότυπα εκπαίδευσης απεικονίζονται πάνω στη γραφική παράσταση με κόκκινες κουκίδες.	136
A'.4 Η κωδικοποίηση των αριθμών 0 και 1 και οι αντίστοιχες δυαδικές κωδικοποιήσεις.	138
A'.5 Iris Setosa.	138
A'.6 Iris Versicolor.	139
A'.7 Iris Virginica.	139

Κατάλογος Πινάκων

3.1	Συγκριτικά αποτελέσματα για το πρόβλημα του αποκλειστικού-EITE . .	67
3.2	Συγκριτικά αποτελέσματα για το πρόβλημα της ισοτιμίας των 3-bit . . .	68
3.3	Συγκριτικά αποτελέσματα για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων	69
3.4	Συγκριτικά αποτελέσματα για το πρόβλημα της προσέγγισης συνεχούς τριγωνομετρικής συνάρτησης.	70
4.1	Συγκριτικά αποτελέσματα για το πρόβλημα του αποκλειστικού-EITE . .	90
4.2	Συγκριτικά αποτελέσματα για το πρόβλημα της ισοτιμίας των 3-bit . . .	92
4.3	Προδιαγραφές των προβλημάτων $N-M-N$ Κωδικοποιητή/Αποκωδικοποιητή. Οι αναφερόμενες παράμετροι είναι: <i>Δοκιμές</i> , ο αριθμός των εξομοιώσεων, <i>Όριο Έπ.</i> , ο μέγιστος επιτρεπτός αριθμός συναρτησιακών υπολογισμών, <i>Βάρη</i> , ο αριθμός των συναπτικών βαρών, <i>Μεροληψίες</i> , ο αριθμός των μεροληψιών, και <i>e, g</i> , η συνθήκη τερματισμού.	93
4.4	Συγκριτικά αποτελέσματα για το πρόβλημα του $4-2-4$ Κωδικοποιητή/Αποκωδικοποιητή	94
4.5	Συγκριτικά αποτελέσματα για το πρόβλημα του $8-3-8$ Κωδικοποιητή/Αποκωδικοποιητή	95
4.6	Συγκριτικά αποτελέσματα για το πρόβλημα του $16-4-16$ Κωδικοποιητή/Αποκωδικοποιητή	97
4.7	Συγκριτικά αποτελέσματα για το πρόβλημα του $32-5-32$ Κωδικοποιητή/Αποκωδικοποιητή	98
4.8	Συγκριτικά αποτελέσματα για το πρόβλημα του $64-6-64$ Κωδικοποιητή/Αποκωδικοποιητή	99

4.9	Συγκριτικά αποτελέσματα για το πρόβλημα του 128 – 7 – 128 Κωδικοποιητή/Αποκωδικοποιητή	100
4.10	Συγκριτικά αποτελέσματα για το πρόβλημα του 256 – 8 – 256 Κωδικοποιητή/Αποκωδικοποιητή	101
4.11	Συγκριτικά αποτελέσματα για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων	103
4.12	Συγκριτικά αποτελέσματα για το πρόβλημα της αναγνώρισης αριθμών .	106
4.13	Συγκριτικά αποτελέσματα για το πρόβλημα της ταξινόμησης των φυτών της οικογένειας Iris	107
4.14	Συγκριτικά αποτελέσματα για το πρόβλημα της οπτικής αναγνώρισης χειρόγραφων αριθμών	109
4.15	Συγκριτικά αποτελέσματα για το πρόβλημα της αναγνώρισης φωνηέντων	110
5.1	Συγκριτικά αποτελέσματα για το πρόβλημα του αποκλειστικού-EITE . .	123
5.2	Συγκριτικά αποτελέσματα για το πρόβλημα της ισοτιμίας των 3-bit . . .	123
5.3	Συγκριτικά αποτελέσματα για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων	124
5.4	Συγκριτικά αποτελέσματα για το πρόβλημα της προσέγγισης συνεχούς τριγωνομετρικής συνάρτησης.	125
5.5	Συγκριτικά αποτελέσματα για το πρόβλημα της Ταξινόμησης Καρκινικών Κυττάρων Εγκεφάλου μεταξύ του εμπρόσθια τροφοδοτούμενου νευρωνικού δικτύου με 7 νευρώνες (MLP) στο κρυφό επίπεδο και του πιθανοτικού νευρωνικού δικτύου (PNN).	129
A'.1	Τα πρότυπα εκπαίδευσης της Ισοτιμίας των 3-bit.	135
A'.2	Τα πρότυπα εκπαίδευσης 4 – 2 – 4 Κωδικοποιητή/Αποκωδικοποιητή. .	137
A'.3	Τα πρότυπα εκπαίδευσης 8 – 3 – 8 Κωδικοποιητή/Αποκωδικοποιητή. .	137

Μέρος Ι

ΕΙΣΑΓΩΓΗ

Τεχνητά Νευρωνικά Δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ) έχουν εμπνευστεί από την διαπίστωση ότι ο ανθρώπινος εγκέφαλος κάνει υπολογισμούς με ένα εντελώς διαφορετικό τρόπο από τους συμβατικούς ψηφιακούς υπολογιστές. Ο ανθρώπινος εγκέφαλος είναι φτιαγμένος από ένα τεράστιο δίκτυο από υπολογιστικά στοιχεία, που καλούνται *νευρώνες*, ζευγαρωμένα με αισθητήριους δέκτες και επιδραστές. Ο μέσος ανθρώπινος εγκέφαλος υπολογίζεται ότι περιέχει περίπου 100 δισεκατομμύρια κύτταρα ποικίλου είδους. Ένας νευρώνας είναι ένα ειδικό κύτταρο που μεταφέρει ένα ηλεκτρικό σήμα. Υπάρχουν περίπου 10 δισεκατομμύρια νευρώνες στον ανθρώπινο εγκέφαλο. Τα εναπομείναντα 90 δισεκατομμύρια κύτταρα ονομάζονται *νευρογλοιακά* ή *κολλώδη κύτταρα*, και αυτά υπηρετούν ως κύτταρα υποστήριξης για τους νευρώνες. Οι νευρώνες αλληλεπιδρούν μέσω επαφών που ονομάζονται *συνάψεις*. Κατά μέσο όρο κάθε νευρώνας λαμβάνει σήματα μέσω χιλιάδων συνάψεων.

Ο εγκέφαλος οργανώνει αυτόν τον τεράστιο αριθμό από νευρώνες σε ένα μαζικό παράλληλο περίπλοκο δίκτυο, στο οποίο αυτοί οι νευρώνες αλληλεπιδρούν μεταξύ τους δυναμικά για να παράγουν ένα πανίσχυρο επεξεργαστή πληροφορίας. Τυπικά, οι νευρώνες είναι πέντε έως έξι τάξεις μεγέθους αργότεροι από τις λογικές πύλες πυριτίου. Ο μοντέρνος υπολογιστής εύκολα υπερτερεί του ανθρώπου στον προγραμματισμό επαναληπτικών υπολογισμών. Ωστόσο, η κατανόηση ομιλίας σε πραγματικό χρόνο και η οπτική αντίληψη, τις οποίες ο άνθρωπος υλοποιεί χωρίς κόπο, είναι ακόμη πέρα από την ικανότητα των ψηφιακών υπολογιστών. Για παράδειγμα, ο ανθρώπινος εγκέφαλος επιτυγχάνει αναγνώριση γνωστών προσώπων σε άγνωστα περιβάλλοντα σε περίπου 100 με 200 χιλιοστά του δευτερολέπτου, ενώ η ίδια λειτουργία σε ένα συμβατικό υπολογιστή θα χρειαζόταν αρκετές ημέρες για να εκτελεστεί.

Ο λόγος που ο ανθρώπινος εγκέφαλος υπερσχύει οφείλεται στην εκπληκτική δομή του και στην ικανότητά του να σχηματίζει τους δικούς του κανόνες μέσω μιας διαδικασίας γνωστής ως εμπειρία. Η εμπειρία χτίζεται με την άροδο του χρόνου. Η πιο δραματική

ανάπτυξή της λαμβάνει χώρα κατά τα δυο πρώτα έτη από την γέννηση, αλλά συνεχίζεται κι αργότερα.

Το αναπτυσσόμενο νευρικό σύστημα χαρακτηρίζεται από μια ιδιότητα που ονομάζεται πλαστικότητα, η οποία του επιτρέπει να προσαρμόζεται στο περιβάλλον. Η πλαστικότητα είναι ουσιώδης στη λειτουργία των νευρώνων ως μονάδες επεξεργασίας πληροφοριών στον ανθρώπινο εγκέφαλο. Με τον ίδιο τρόπο η πλαστικότητα εμφανίζεται να είναι ουσιώδης και στα νευρωνικά δίκτυα με τεχνητούς νευρώνες. Ένα νευρωνικό δίκτυο στην απλούστερη μορφή του είναι μια μηχανή που είναι σχεδιασμένη να προσομοιώνει τον τρόπο με τον οποίο ο εγκέφαλος εκτελεί μια συγκεκριμένη εργασία. Τα νευρωνικά δίκτυα για να μπορέσουν να εκτελέσουν χρήσιμους υπολογισμούς χρησιμοποιούν μια διαδικασία που ονομάζεται μάθηση. Για να επιτύχουν καλή απόδοση, χρησιμοποιούν ένα τεράστιο αριθμό διασυνδεδεμένων απλών υπολογιστικών κυττάρων που ονομάζονται νευρώνες. Σύμφωνα με τα παραπάνω, ένα νευρωνικό δίκτυο ορίζεται ως εξής:

Ορισμός 1.1. Ένα νευρωνικό δίκτυο είναι ένας μαζικός παράλληλος κατανεμημένος επεξεργαστής αποτελούμενος από απλές επεξεργαστικές μονάδες, τους νευρώνες, που έχουν τη φυσική ροπή να αποθηκεύουν εμπειρική γνώση και να τη διαθέτουν προς χρήση. Μοιάζουν με τον εγκέφαλο σε δύο σημεία:

1. Το δίκτυο αποκτά την γνώση από το περιβάλλον μέσω μιας διαδικασίας μάθησης.
2. Οι σύνδεσμοι μεταξύ των νευρώνων που ονομάζονται συναπτικά βάρη ή απλά βάρη, χρησιμοποιούνται για την αποθήκευση της γνώσης.

Η διαδικασία που χρησιμοποιείται για την εκτέλεση της διαδικασίας μάθησης καλείται αλγόριθμος εκπαίδευσης. Ο αλγόριθμος εκπαίδευσης ουσιαστικά τροποποιεί τα βάρη του νευρωνικού δικτύου με μεθοδικό τρόπο για την επίτευξη του επιθυμητού σχεδιαστικού στόχου.

1.1 Οφέλη των Νευρωνικών Δικτύων

Η χρήση των νευρωνικών δικτύων προσφέρει τις ακόλουθες χρήσιμες ιδιότητες και δυνατότητες:

1. **Μη γραμμικότητα.** Ένας τεχνητός νευρώνας μπορεί να είναι γραμμικός ή μη γραμμικός. Ένα νευρωνικό δίκτυο αποτελούμενο από μη γραμμικούς νευρώνες είναι μη γραμμικό. Επιπλέον, η μη γραμμικότητα είναι ιδιαίτερη υπό την έννοια ότι είναι κατανεμημένη σε ολόκληρο το δίκτυο. Η μη γραμμικότητα είναι μια εξαιρετικά σημαντική ιδιότητα, ειδικότερα όταν ο φυσικός μηχανισμός που είναι υπεύθυνος για την δημιουργία του σήματος εισόδου είναι εκ φύσεως μη γραμμικός.

2. **Συσχέτιση εισόδου–εξόδου.** Η επιβλεπόμενη μάθηση, η οποία αποτελεί ένα δημοφιλές πρότυπο μάθησης, τροποποιεί τα συναπτικά βάρη του νευρωνικού δικτύου χρησιμοποιώντας ένα σύνολο από πρότυπα εκπαίδευσης. Κάθε πρότυπο αποτελείται από ένα μοναδικό σήμα εισόδου και την αντίστοιχη επιθυμητή έξοδο. Τα συναπτικά βάρη του δικτύου τροποποιούνται έτσι ώστε η διαφορά μεταξύ της επιθυμητής εξόδου και της πραγματικής εξόδου που παράγεται από το σήμα εισόδου να ελαχιστοποιείται σύμφωνα με ένα στατιστικό κριτήριο. Η εκπαίδευση του δικτύου συνεχίζεται για πολλά πρότυπα εκπαίδευσης μέχρι το δίκτυο να φτάσει σε μια σταθερή κατάσταση όπου η τροποποίηση των συναπτικών βαρών να μην είναι πλέον σημαντική. Έτσι το δίκτυο μαθαίνει από τα πρότυπα εκπαίδευσης δημιουργώντας μια συσχέτιση εισόδου–εξόδου για το δοθέν πρόβλημα.
3. **Προσαρμοστικότητα.** Τα νευρωνικά δίκτυα έχουν την ικανότητα να προσαρμόζουν τα συναπτικά τους βάρη όταν το περιβάλλον τους αλλάζει. Πιο συγκεκριμένα, ένα νευρωνικό δίκτυο που εκπαιδεύεται σε ένα συγκεκριμένο περιβάλλον μπορεί εύκολα να επανεκπαιδευτεί ώστε να αντιμετωπίσει μικρές αλλαγές στο περιβάλλον. Επιπλέον, όταν το δίκτυο λειτουργεί σε μη σταθερό περιβάλλον, μπορεί να σχεδιαστεί έτσι ώστε να προσαρμόζει τα συναπτικά του βάρη σε πραγματικό χρόνο.
4. **Απόκριση βασισμένη σε ενδείξεις.** Ένα νευρωνικό δίκτυο μπορεί να σχεδιαστεί έτσι ώστε να παρέχει πληροφορίες όχι μόνο για την ταξινόμηση των εισόδων σε κλάσεις, αλλά και για τον βαθμό εμπιστοσύνης αυτής της ταξινόμησης. Η τελευταία πληροφορία μπορεί να χρησιμοποιηθεί για την ταξινόμηση νέων αγνώστων προτύπων κατά την διάρκεια της εκπαίδευσης.
5. **Συναφείς πληροφορίες.** Η γνώση αντιπροσωπεύεται από την δομή και την κατάσταση ενεργοποίησης του νευρωνικού δικτύου. Κάθε νευρώνας του δικτύου είναι ενδεχομένως επηρεασμένος από την ολική δραστηριότητα όλων των υπόλοιπων νευρώνων του δικτύου. Συνεπώς, συναφείς πληροφορίες αντιμετωπίζονται με φυσικό τρόπο από το νευρωνικό δίκτυο.
6. **Ανοχή σε σφάλματα.** Ένα νευρωνικό δίκτυο που έχει υλοποιηθεί σε ηλεκτρονική μορφή υλικού (hardware) έχει την δυνατότητα να είναι εγγενώς ανεκτικό σε σφάλματα, με την έννοια ότι η απόδοση του μειώνετε ομαλά κάτω από αντίξοες συνθήκες λειτουργίας. Για παράδειγμα, αν ένας νευρώνας του δικτύου υποστεί ζημιά, η απόδοση του δικτύου θα μειωθεί ποιοτικά. Ωστόσο, λόγω της κατανεμημένης φύσεως της πληροφορίας που είναι αποθηκευμένη στο δίκτυο, η ζημιά θα πρέπει να είναι εκτεταμένη για να παρατηρηθούν σημαντικά σφάλματα στην απόκριση του δικτύου.
7. **Υλοποίηση VLSI.** Η μαζικά παράλληλη φύση του νευρωνικού δικτύου το καθιστά ικανό για γρήγορους υπολογισμούς. Αυτό το χαρακτηριστικό καθιστά τα νευρωνικά δίκτυα ιδανικά για υλοποίηση σε υλικό με χρήση της τεχνολογίας ολοκλήρωσης πολύ μεγάλης κλίμακας (VLSI). Μια σημαντική αρετή της υλοποίησης

VLSI είναι ότι παρέχει την δυνατότητα να συλλαμβάνει πραγματικά πολύπλοκη συμπεριφορά με ένα ιεραρχικό τρόπο.

8. **Ομοιομορφία ανάλυσης και σχεδιασμού.** Τα νευρωνικά δίκτυα μοιράζονται ίδια σημειογραφία σε όλες τις εφαρμογές τους. Αυτή η ιδιότητα εκδηλώνεται με διαφορετικούς τρόπους:

- Οι νευρώνες αποτελούν ένα κοινό συστατικό σε όλα τα νευρωνικά δίκτυα.
- Τα κοινά χαρακτηριστικά δίνουν την δυνατότητα για κοινούς αλγόριθμους εκπαίδευσης σε διαφορετικές εφαρμογές των νευρωνικών δικτύων.

9. **Βιολογική αναλογία.** Ο σχεδιασμός των νευρωνικών δικτύων είναι εμπνευσμένος από τον ανθρώπινο εγκέφαλο. Οι νευροβιολόγοι συχνά χρησιμοποιούν τα νευρωνικά δίκτυα ως ένα εργαλείο για την εξήγηση νευροβιολογικών φαινομένων. Από την άλλη μεριά, οι μηχανικοί χρησιμοποιούν την νευροβιολογία για να αφομοιώσουν νέες ιδέες που θα τους βοηθήσουν στην επίλυση πολύπλοκων προβλημάτων.

1.2 Βιολογικοί Νευρώνες

Υπάρχουν πολλοί διαφορετικοί τύποι βιολογικών νευρώνων. Οι βασικές ιδιότητές τους όμως μοιράζονται από πολλούς νευρώνες. Ένας νευρώνας (Εικόνα 1.1) αποτελείται από τα παρακάτω κύρια μέρη:

1. Πυρήνα
2. Σώμα
3. Δενδρίτες
4. Αξονικό λοφίσκο
5. Άξονα
6. Τερματικοί βολβοί

Κάθε νευρώνας είναι ένα κύτταρο το οποίο χρησιμοποιεί βιοχημικές αντιδράσεις για τη λήψη, επεξεργασία και μετάδοση πληροφοριών.

Το δενδριτικό δέντρο ενός νευρώνα είναι συνδεδεμένο με χιλιάδες γειτονικούς νευρώνες. Όταν ένας από αυτούς τους νευρώνες εκπέμπει σήμα, ένα αρνητικό ή θετικό φορτίο λαμβάνεται από έναν δενδρίτη. Τα δυναμικά όλων των λαμβανόμενων φορτίων

προστίθενται μέσω μιας διαδικασίας χωρικής και χρονικής πρόσθεσης. Η χωρική πρόσθεση λαμβάνει χώρα όταν πολλά αδύναμα σήματα μετατρέπονται σε ένα μεγάλο, ενώ η χρονική πρόσθεση μετατρέπει μια σειρά από ταχεία αδύναμα σήματα από μια πηγή σε ένα μεγάλο σήμα. Η συνολική είσοδος στη συνέχεια εισέρχεται στο σώμα. Το σώμα και ο πυρήνας δεν παίζουν κανένα ρόλο στην επεξεργασία το εισερχομένων και εξερχομένων δεδομένων. Η λειτουργία τους περιορίζεται στην συντήρηση του κυττάρου ώστε ο νευρώνας να παραμένει λειτουργικός. Το μέρος του σώματος το οποίο είναι σημαντικό για την διάδοση του σήματος είναι ο αξονικός λοφίσκος. Αν η συνολική είσοδος είναι μεγαλύτερη από την τιμή του κατωφλιού του αξονικού λοφίσκου, τότε ο νευρώνας εκπέμπει και το σήμα του μεταδίδεται στον άξονα. Το δυναμικό της εξόδου είναι σταθερό είτε αν το σήμα εισόδου είναι μόλις επάνω από την τιμή του κατωφλιού, ή εκατό φορές πιο πάνω. Το δυναμικό του σήματος εξόδου δεν επηρεάζεται από τις πολλές διακλαδώσεις του άξονα και φτάνει στους τερματικούς βολβούς με την ίδια ένταση που είχε στον αξονικό λοφίσκο.

Σχήμα 1.1: Σχέδιο ενός βιολογικού νευρώνα.

Ένας νευρώνας είναι εγκλεισμένος από την μεμβράνη του. Η διέγερση ενός νευρώνα επηρεάζει άλλους νευρώνες μέσω των συνδέσεων των τερματικών βολβών με τις μεμβράνες των άλλων νευρώνων. Τέτοιες συνδέσεις ονομάζονται συνάψεις (Εικόνα 1.2). Η μεμβράνη του τερματικού βολβού ονομάζεται *προσυναπτική μεμβράνη*. Η μεμβράνη πάνω στην οποία επιδρά ο τερματικός βολβός ονομάζεται *μετασυναπτική μεμβράνη*. Ο μικρός χώρος μεταξύ των προσυναπτικών και των μετασυναπτικών μεμβρανών ονομάζεται *συναπτική σχισμή*. Η επικοινωνία μεταξύ των νευρώνων διευκολύνεται από την απελευθέρωση μικρών πακέτων χημικών μέσα στο κενό αυτό. Ένας νευρώνας μπορεί να επικοινωνεί με πάνω από 100000 άλλους νευρώνες.

Όταν ένα δυναμικό δράσης φτάσει στον τερματικό βολβό της προσυναπτικής μεμβράνης, προξενεί μια αλλαγή δυναμικού απέναντι στην μετασυναπτική μεμβράνη. Η άμεση

αίτια της αλλαγής στο ηλεκτρικό δυναμικό της μετασυναπτικής μεμβράνης είναι αρχικά χημική και όχι ηλεκτρική. Η ηλεκτρική ώση που μεταδίδεται διαμέσω κάθε συναπτικού κενού είναι χημικώς παρακινούμενη και ελεγχόμενη.

Σχήμα 1.2: Μετάδοση σήματος προσυναπτικού άξονα μέσω συναπτικής σχισμής στον μετασυναπτικό νευρώνα.

Ένα δυναμικό δράσης που διαδίδεται στον τερματικό βολβό προξενεί την απελευθέρωση χημικών ουσιών που ονομάζονται νευροδιαβιβαστές από μικρά πακέτα (συναπτικά κυστίδια) στην προσυναπτική μεμβράνη. Οι νευροδιαβιβαστές διαχέονται διαμέσου της συναπτικής σχισμής στη μετασυναπτική μεμβράνη και δημιουργούν μια αλλαγή στο ηλεκτρικό δυναμικό απέναντι στη μετασυναπτική μεμβράνη.

1.3 Τεχνητοί Νευρώνες

Ένας τεχνητός νευρώνας είναι μια μονάδα επεξεργασίας πληροφορίας η οποία είναι θεμελιώδης για την λειτουργία ενός νευρωνικού δικτύου. Το διάγραμμα της Εικόνας 1.3 δείχνει το μοντέλο ενός τεχνητού νευρώνα, ο οποίος σχηματίζει την βάση για τον σχεδιασμό των τεχνητών νευρωνικών δικτύων. Εδώ αναγνωρίζονται τρία βασικά στοιχεία των νευρωνικών μοντέλων:

1. Ένα σύνολο από συνάψεις, καθεμία από τις οποίες χαρακτηρίζεται από ένα βάρος. Συγκεκριμένα, ένα σήμα x_j στην είσοδο της συνάψης j συνδεδεμένη με τον νευρώνα k πολλαπλασιάζεται με το συναπτικό βάρος w_{kj} . Είναι σημαντικό να γίνει μια σημείωση για τον τρόπο με τον οποίο γράφονται οι υποδείκτες των συναπτικών βαρών w_{kj} . Ο πρώτος υποδείκτης αναφέρεται στον νευρώνα υπό εξέταση και ο δεύτερος υποδείκτης αναφέρεται στο τέλος της εισόδου της συνάψης στην οποία αναφέρεται το βάρος. Αντίθετα με μια συνάψη στον εγκέφαλο, το συναπτικό βάρος

ενός τεχνητού νευρώνα μπορεί να βρίσκεται σε ένα πεδίο που περιέχει αρνητικές, καθώς επίσης και θετικές τιμές.

2. Ένας αθροιστής για την άθροιση των σημάτων εισόδου, βεβαρμένων από τις αντίστοιχες συνάψεις του νευρώνα. Οι λειτουργίες που περιγράφονται εδώ συνιστούν ένα γραμμικό συνδυαστή.
3. Μια συνάρτηση ενεργοποίησης για τον περιορισμό του πλάτους της εξόδου ενός νευρώνα. Η συνάρτηση ενεργοποίησης επίσης αναφέρεται ως συνάρτηση πολτοποίησης επειδή πολτοποιεί (περιορίζει) το επιτρεπόμενο πλάτος του πεδίου του σήματος εξόδου σε κάποια πεπερασμένη τιμή. Τυπικώς, το κανονικοποιημένο πλάτος του πεδίου της εξόδου ενός νευρώνα γράφεται ως το κλειστό μοναδιαίο διάστημα $[0, 1]$ ή εναλλακτικά $[-1, 1]$.

Το νευρωνικό μοντέλο της Εικόνας 1.3 επίσης περιλαμβάνει μια εξωτερικά εφαρμοζόμενη μεροληψία, που δηλώνεται ως b_k . Η μεροληψία b_k έχει επίδραση στην αύξηση ή στην μείωση της καθαρής εισόδου της συνάρτησης ενεργοποίησης, εξαρτώμενη από πότε είναι θετική ή αρνητική, αντίστοιχα.

Σχήμα 1.3: Μη γραμμικό μοντέλο ενός νευρώνα.

Σε μαθηματικούς όρους, μπορεί να περιγραφεί ένας νευρώνας k γράφοντας το ακόλουθο ζεύγος εξισώσεων:

$$u_k = \sum_{j=1}^m w_{kj} x_j \quad (1.1)$$

και

$$y_k = f(u_k + b_k) \quad (1.2)$$

όπου $x_1, x_2, \dots, x_m$ είναι τα σήματα εισόδου, $w_{k1}, w_{k2}, \dots, w_{km}$ είναι τα συναπτικά βάρη του νευρώνα k , u_k είναι η έξοδος του γραμμικού συσχετιστή λόγω των σημάτων εισόδου, b_k είναι η μεροληψία, $f(\cdot)$ είναι η συνάρτηση ενεργοποίησης και y_k είναι το σήμα εξόδου του νευρώνα. Η χρήση της μεροληψίας b_k έχει ως σκοπό την εφαρμογή ενός ομοπαράλληλικού μετασχηματισμού στην έξοδο u_k του γραμμικού συνδυαστή στο μοντέλο της Εικόνας 1.3, όπως φαίνεται από την

$$v_k = u_k + b_k \quad (1.3)$$

Σχήμα 1.4: Ομοπαράλληλικός μετασχηματισμός παραγόμενος από την παρουσία μεροληψίας. Σημειώνεται ότι $v_k = b_k$ στο $u_k = 0$.

Συγκεκριμένα, εξαρτώμενα από πότε η μεροληψία b_k είναι θετική ή αρνητική, η σχέση μεταξύ του επηρεασμένου τοπικού πεδίου v_k του νευρώνα k και της εξόδου u_k του

γραμμικού συσχετιστή, τροποποιείται με ένα τρόπο που φαίνεται στην Εικόνα 1.4. Σημειώνεται ότι ως ένα αποτέλεσμα αυτού του ομοπαράλληλικού μετασχηματισμού, το γράφημα του v_k κατά του u_k δεν περνάει πλέον από την αρχή των συντεταγμένων.

Η μεροληψία b_k είναι μια εξωτερική παράμετρος του νευρώνα k . Ισοδύναμα, μπορεί να διατυπωθεί ο συνδυασμός των εξισώσεων (1.1) και (1.3) όπως ακολουθεί:

$$v_k = \sum_{j=0}^m w_{kj} x_j \quad (1.4)$$

και

$$y_k = f(v_k) \quad (1.5)$$

Στην εξίσωση (1.4) προστέθηκε μια νέα σύναψη. Η είσοδος της είναι

$$x_0 = +1 \quad (1.6)$$

και το βάρος της

$$w_{k0} = b_k \quad (1.7)$$

Γι' αυτό το λόγο, μπορεί να επαναδιατυπωθεί το μοντέλο του νευρώνα k όπως στην Εικόνα 1.5. Σε αυτή την εικόνα η επίδραση της μεροληψίας εξηγείται κάνοντας δύο πράγματα:

- (1) προσθέτοντας ένα νέο σήμα εισόδου σταθερό στην τιμή $+1$, και
- (2) προσθέτοντας ένα νέο συναπτικό βάρος ίσο με την μεροληψία b_k .

Παρόλο που τα μοντέλα των Εικόνων 1.3 και 1.5 είναι διαφορετικά στην εμφάνιση, είναι μαθηματικά ισοδύναμα.

Σχήμα 1.5: Ένα άλλο μη γραμμικό μοντέλο νευρώνα.

1.3.1 Τύποι Συναρτήσεων Ενεργοποίησης

Η συνάρτηση ενεργοποίησης, που δηλώνεται από την $f(v)$, ορίζει την έξοδο ενός νευρώνα σε σχέση με το επηρεασμένο τοπικό πεδίο. Στο σημείο αυτό παρουσιάζονται τρεις βασικοί τύποι συναρτήσεων ενεργοποίησης.

Συνάρτηση Κατωφλιού

Γι' αυτό τον τύπο συνάρτησης ενεργοποίησης, που περιγράφεται στην Εικόνα 1.6, έχουμε:

$$f(v) = \begin{cases} 1 & \text{αν } v \geq 0 \\ 0 & \text{αν } v < 0 \end{cases} \quad (1.8)$$

Στην βιβλιογραφία της μηχανικής, αυτή η μορφή μιας συνάρτησης κατωφλιού κοινά αναφέρεται ως μια *Heaviside συνάρτηση*. Αντιστοίχως, η έξοδος ενός νευρώνα k που κατέχει μια τέτοια συνάρτηση κατωφλιού, εκφράζεται ως

$$y_k = \begin{cases} 1 & \text{αν } v_k \geq 0 \\ 0 & \text{αν } v_k < 0 \end{cases} \quad (1.9)$$

όπου v_k είναι το επηρεασμένο τοπικό πεδίο του νευρώνα, δηλαδή

$$v_k = \sum_{j=1}^m w_{kj} x_j + b_k \quad (1.10)$$

Ένας τέτοιος νευρώνας αναφέρεται στην βιβλιογραφία ως *το μοντέλο McCulloch–Pitts* σε αναγνώριση της πρωτοπόρας δουλειάς που έγινε από τους McCulloch και Pitts (1943). Σε αυτό το μοντέλο, η έξοδος ενός νευρώνα παίρνει την τιμή 1 αν το επηρεασμένο τοπικό πεδίο του νευρώνα είναι μη αρνητικό, και 0 διαφορετικά. Αυτός ο ισχυρισμός περιγράφει την ιδιότητα *όλα ή τίποτα* του μοντέλου McCulloch–Pitts.

Σχήμα 1.6: Συνάρτηση Κατωφλιού.

Κατά Τμήματα Γραμμική Συνάρτηση

Για την κατά τμήματα γραμμική συνάρτηση που περιγράφεται στην Εικόνα 1.7 έχουμε

$$f(v) = \begin{cases} 1, & v \geq +\frac{1}{2} \\ v, & +\frac{1}{2} > v > -\frac{1}{2} \\ 0, & v \leq -\frac{1}{2} \end{cases} \quad (1.11)$$

όπου ο παράγοντας ενίσχυσης μέσα στην γραμμική περιοχή της δράσης υποτίθεται ότι είναι μοναδιαίος. Αυτός ο τύπος συνάρτησης ενεργοποίησης μπορεί να θεωρηθεί ως μια προσέγγιση ενός μη γραμμικού ενισχυτή. Οι ακόλουθες δύο περιπτώσεις μπορούν να θεωρηθούν ως ειδικές μορφές της κατά τμήματα γραμμικής συνάρτησης:

- Ένας *γραμμικός συσχετιστής* παρουσιάζεται αν η γραμμική περιοχή δράσης διατηρείται χωρίς να τρέχει σε κορεσμό.
- Η κατά τμήματα γραμμική συνάρτηση ανάγεται σε μια *συνάρτηση κατωφλιού* αν ο παράγοντας ενίσχυσης της γραμμικής περιοχής γίνεται απείρως μεγάλος.

Σχήμα 1.7: Κατά Τμήματα Γραμμική Συνάρτηση.

Σιγμοειδής Συνάρτηση

Η σιγμοειδής συνάρτηση, της οποίας το γράφημα έχει σχήμα s, είναι μακράν ο πιο κοινός τύπος συνάρτησης ενεργοποίησης που χρησιμοποιείται στην κατασκευή τεχνητών νευρωνικών δικτύων. Ορίζεται ως μια αυστηρά αύξουσα συνάρτηση η οποία επιδεικνύει μια «χαριτωμένη» ισορροπία μεταξύ της μη γραμμικής και γραμμικής συμπεριφοράς. Ένα παράδειγμα της σιγμοειδής συνάρτησης είναι η *λογιστική συνάρτηση*, που ορίζεται από

$$f(v) = \frac{1}{1 + \exp(-av)} \quad (1.12)$$

όπου a είναι η *παράμετρος κλίσης* της σιγμοειδούς συνάρτησης. Διαφοροποιώντας την παράμετρο a , επιτυγχάνουμε σιγμοειδείς συναρτήσεις με διαφορετικές κλίσεις, όπως φαίνεται στην Εικόνα 1.8. Στην πραγματικότητα, η κλίση στην αρχή των συντεταγμένων είναι ίση με $a/4$. Στο όριο, καθώς η παράμετρος κλίσης πλησιάζει στο άπειρο, η σιγμοειδής συνάρτηση γίνεται απλώς μια συνάρτηση κατωφλιού. Ενώ μια συνάρτηση κατωφλιού λαμβάνει την τιμή 0 ή 1, μια σιγμοειδής συνάρτηση λαμβάνει ένα συνεχές πεδίο τιμών από 0 έως το 1. Σημειώνεται, επίσης, ότι η σιγμοειδής συνάρτηση είναι διαφορίσιμη, ενώ η συνάρτηση κατωφλιού δεν είναι.

Οι συναρτήσεις ενεργοποίησης που ορίστηκαν στις εξισώσεις (1.8), (1.11) και (1.12) διακυμαίνονται από το 0 στο +1. Είναι επιθυμητό μερικές φορές να έχουμε το πεδίο της συνάρτησης ενεργοποίησης από το -1 στο +1. Στην περίπτωση αυτή, η συνάρτηση ενεργοποίησης λαμβάνει ένα αντισυμμετρικό τύπο σε σχέση με την αρχή των συντεταγμένων. Δηλαδή, η συνάρτηση ενεργοποίησης είναι μια περιττή συνάρτηση του επηρεασμένου τοπικού πεδίου. Συγκεκριμένα, η συνάρτηση κατωφλιού της εξίσωσης (1.8)

Σχήμα 1.8: Σιγμοειδής Συνάρτηση.

ορίζεται τώρα ως

$$f(v) = \begin{cases} 1, & \text{αν } v > 0 \\ 0, & \text{αν } v = 0 \\ -1, & \text{αν } v < 0 \end{cases} \quad (1.13)$$

η οποία κοινά αναφέρεται ως *συνάρτηση προσήμου (signum function)*. Για τον αντίστοιχο τύπο μιας σιγμοειδούς συνάρτησης μπορεί να χρησιμοποιηθεί η *συνάρτηση υπερβολικής εφαπτομένης (hyperbolic tangent function)*, που ορίζεται από

$$f(v) = \tanh(v) \quad (1.14)$$

Επιτρέποντας μια συνάρτηση σιγμοειδούς τύπου να λάβει αρνητικές τιμές όπως περιγράφηκε από την εξίσωση (1.14) έχει αναλυτικά πλεονεκτήματα.

1.4 Αρχιτεκτονικές Δικτύων

Ο τρόπος με τον οποίο οι νευρώνες ενός νευρωνικού δικτύου είναι δομημένοι είναι στενά συνδεδεμένος με τον αλγόριθμο μάθησης που χρησιμοποιείται για την εκπαίδευση του δικτύου. Υπάρχουν πολλές αρχιτεκτονικές νευρωνικών δικτύων, αλλά σε αυτή τη διατριβή θα ασχοληθούμε με τα εμπρόσθια τροφοδοτούμενα νευρωνικά δίκτυα.

1.4.1 Μονοεπίπεδα Εμπρόσθια Τροφοδοτούμενα Δίκτυα

Σε ένα νευρωνικό δίκτυο οι νευρώνες οργανώνονται με την μορφή επιπέδων. Στην απλούστερη μορφή ενός δικτύου, έχουμε ένα επίπεδο εισόδου που προβάλλεται πάνω σε

ένα επίπεδο εξόδου από νευρώνες, αλλά όχι το αντίστροφο. Με άλλα λόγια, αυτό το δίκτυο είναι εμπρόσθια τροφοδοτούμενο. Ένα τέτοιο δίκτυο παρουσιάζεται στην Εικόνα 1.9 για την περίπτωση των m νευρώνων στο επίπεδο εισόδου και των n νευρώνων στο επίπεδο εξόδου. Ένα τέτοιο δίκτυο καλείται *μονοεπίπεδο δίκτυο*, με τον χαρακτηρισμό «μονοεπίπεδο» να αναφέρεται στο επίπεδο εξόδου των νευρώνων. Δεν υπολογίζουμε το επίπεδο εισόδου γιατί δεν εκτελείται κανένας υπολογισμός εκεί.

Σχήμα 1.9: Εμπρόσθια τροφοδοτούμενο δίκτυο με ένα μονό επίπεδο από νευρώνες.

1.4.2 Πολυεπίπεδα Εμπρόσθια Τροφοδοτούμενα Δίκτυα

Τα πολυεπίπεδα εμπρόσθια τροφοδοτούμενα νευρωνικά δίκτυα διακρίνονται από την παρουσία ενός ή περισσοτέρων *κρυφών επιπέδων*, των οποίων οι νευρώνες καλούνται *αντιστοιχώς κρυφοί νευρώνες*. Η λειτουργία των κρυφών νευρώνων είναι να μεσολαβεί μεταξύ της εξωτερικής εισόδου και της εξόδου του δικτύου με κάποιο χρήσιμο τρόπο. Προσθέτοντας ένα ή περισσότερα κρυφά επίπεδα, το δίκτυο καθίσταται ικανό να εξαγάγει στατιστικά υψηλότερης τάξης. Με μια μάλλον ασαφή έννοια το δίκτυο αποκτά μια *ολική*

όψη παρά την τοπική του συνδεσιμότητα λόγω του επιπλέον συνόλου από συναπτικές συνδέσεις και την επιπλέον διάσταση των νευρωνικών αλληλεπιδράσεων (Churchland και Sejnowski, 1992). Η ικανότητα των κρυφών νευρώνων να εξάγουν στατιστικά υψηλότερης τάξης είναι ιδιαίτερος πολύτιμη όταν το μέγεθος του επιπέδου εισόδου είναι μεγάλο.

Το επίπεδο εισόδου του δικτύου παρέχει τα αντίστοιχα στοιχεία του προτύπου ενεργοποίησης, τα οποία αποτελούν τα σήματα εισόδου που εφαρμόζονται στους νευρώνες στο δεύτερο επίπεδο (δηλαδή στο πρώτο κρυφό επίπεδο). Τα σήματα εξόδου του δεύτερου επιπέδου χρησιμοποιούνται ως εισοδοί για το τρίτο επίπεδο, και αυτό επαναλαμβάνεται για το υπόλοιπο δίκτυο. Τυπικώς οι νευρώνες σε κάθε επίπεδο του δικτύου έχουν ως εισόδους τους τα σήματα εξόδου του προηγούμενου επιπέδου μόνο. Το σύνολο των σημάτων εξόδου των νευρώνων στο επίπεδο εξόδου του δικτύου αποτελεί την συνολική απόκριση του δικτύου για το παρεχόμενο πρότυπο ενεργοποίησης στο επίπεδο εισόδου.

Σχήμα 1.10: Πλήρως διασυνδεδεμένο εμπρόσθια τροφοδοτούμενο δίκτυο με ένα κρυφό επίπεδο και ένα επίπεδο εξόδου.

Ο αρχιτεκτονικός γράφος της Εικόνας 1.10 παρουσιάζει τη διάταξη του πολυεπίπεδου

εμπρόσθια τροφοδοτούμενου νευρωνικού δικτύου για την περίπτωση ενός μόνο κρυφού επιπέδου. Για συντομία το δίκτυο στην Εικόνα 1.10 αναφέρεται ως ένα $m-l-n$ δίκτυο γιατί έχει m εισόδους, l κρυφούς νευρώνες και n νευρώνες εξόδου. Ως ένα άλλο παράδειγμα, ένα εμπρόσθια τροφοδοτούμενο δίκτυο με m εισόδους, h_1 νευρώνες στο πρώτο κρυφό επίπεδο, h_2 νευρώνες στο δεύτερο κρυφό επίπεδο, και q νευρώνες στο επίπεδο εξόδου αναφέρεται ως ένα $m-h_1-h_2-q$ δίκτυο.

Το νευρωνικό δίκτυο στην Εικόνα 1.10 λέγεται ότι είναι *πλήρως διασυνδεδεμένο* με την έννοια ότι κάθε κόμβος σε κάθε επίπεδο του δικτύου συνδέεται με κάθε άλλο κόμβο του διπλανού εμπρόσθιου επιπέδου. Αν, ωστόσο, κάποιες από τις συναπτικές συνδέσεις απουσιάζουν από το δίκτυο, λέμε ότι το δίκτυο είναι *μερικώς διασυνδεδεμένο*.

1.5 Η Ιστορία των Τεχνητών Νευρωνικών Δικτύων

Η νέα εποχή των νευρωνικών δικτύων ξεκίνησε με την πρωτοποριακή δουλειά των McCulloch και Pitts (1943) [50]. Στην εργασία αυτή περιγράφουν μια λογική ανάλυση των νευρωνικών δικτύων που συνένωσε τις μελέτες της νευροφυσιολογίας και τη μαθηματική λογική. Το τυπικό τους μοντέλο ενός νευρώνα υποτέθηκε ότι ακολουθεί έναν «όλα-ή-τίποτα» κανόνα. Με ένα ικανό αριθμό από τέτοιες μονάδες, και με σωστά εφαρμοσμένες συναπτικές συνδέσεις, οι McCulloch και Pitts έδειξαν ότι ένα τέτοιο δίκτυο θα μπορούσε, κατ' αρχήν, να υπολογίσει οποιαδήποτε υπολογίσιμη συνάρτηση. Αυτό ήταν ένα πολύ σημαντικό αποτέλεσμα και με αυτό, γενικώς, συμφωνήθηκε ότι γεννήθηκαν οι αρχές των νευρωνικών δικτύων και της τεχνητής νοημοσύνης.

Η επόμενη μεγάλη ανάπτυξη στα νευρωνικά δίκτυα ήρθε το 1949 με τη δημοσίευση του βιβλίου του Hebb *The Organization of Behavior* [32], στο οποίο παρουσιάστηκε μια αναμφίβολη αναφορά ενός φυσιολογικού κανόνα μάθησης για συναπτική τροποποίηση για πρώτη φορά. Ειδικότερα, ο Hebb πρότεινε ότι η συνδετικότητα του εγκεφάλου αλλάζει συνεχώς ως ένας οργανισμός που μαθαίνει διαφορετικές λειτουργικές εργασίες, και ότι οι *νευρωνικές συναρμολογήσεις* δημιουργούνται από τέτοιες αλλαγές. Ο Hebb ακολουθώντας μια προηγούμενη πρόταση από τον Ramón y Cajal [68] εισήγαγε το τώρα διάσημο *αξίωμα της μάθησης*, το οποίο δηλώνει ότι η αποτελεσματικότητα μιας μεταβλητής σύναψης μεταξύ δύο νευρώνων αυξάνεται από την συνεχόμενη ενεργοποίηση του ενός νευρώνα από τον άλλο δια μέσου εκείνης της σύναψης. Το βιβλίο του Hebb είχε πάρα πολύ μεγάλη επιρροή στους ψυχολόγους, αλλά δυστυχώς είχε λίγη έως καθόλου επίδραση στην κοινότητα της μηχανικής.

Περίπου 15 χρόνια μετά την δημοσίευση της κλασσικής εργασίας των McCulloch και Pitts, μια νέα προσέγγιση στο πρόβλημα της αναγνώρισης προτύπου εισήχθη από τον Rosenblatt (1958) στην εργασία του πάνω στο *perceptron*, μια νέα μέθοδο στην επιβλεπόμενη μάθηση [71]. Το αναγνωρισμένο κατόρθωμα της δουλειά του Rosenblatt

ήταν το επονομαζόμενο *θεώρημα σύγκλισης του perceptron*, η πρώτη απόδειξη για το οποίο περιγράφηκε από τον Rosenblatt (1960) [72]. Το 1960, οι Widrow και Hoff εισήγαγαν τον αλγόριθμο των ελαχίστων μέσων τετραγώνων (*least mean square (LMS) algorithm*) και τον χρησιμοποίησαν για να διατυπώσουν το γραμμικό προσαρμοστικό στοιχείο *Adaline* (*adaptive linear element*) [88]. Η διαφορά μεταξύ του perceptron και του Adaline εντοπίζεται στην διαδικασία εκπαίδευσης. Ένα από τα πρώιμα εκπαιδευσιμα νευρωνικά δίκτυα με επίπεδα με πολλαπλά προσαρμοστικά στοιχεία ήταν η δομή *Madaline* (*multiple-adaline*) που προτάθηκε από τον Widrow και τους μαθητές του (Widrow, 1962 [87]). Κατά την διάρκεια της κλασσικής περιόδου του perceptron στην δεκαετία του 1960, φαινόταν ότι τα νευρωνικά δίκτυα μπορούσαν να κάνουν τα πάντα. Αλλά τότε ήρθε το βιβλίο από τους Minsky και Papert (1969) [51], που χρησιμοποίησαν μαθηματικά για να δείξουν ότι υπάρχουν θεμελιώδεις όρια για το τι μπορούν να υπολογίσουν τα μονοεπίπεδα perceptron. Σε ένα συνοπτικό κεφάλαιο στο πολυεπίπεδο perceptron, δήλωσαν ότι δεν υπήρχε λόγος να υποθέσουν ότι κάποιοι από τους περιορισμούς των μονοεπίπεδων perceptron μπορούσαν να ξεπεραστούν στην πολυεπίπεδη έκδοση.

Η εργασία αυτή συνεισέφερε με τον ένα τρόπο ή με τον άλλο στον μετριασμό του ενδιαφέροντος για τα νευρωνικά δίκτυα στην δεκαετία του 1970. Πολλοί από τους ερευνητές, εκτός από εκείνους της ψυχολογίας και των νευροεπιστημών, εγκατέλειψαν την περιοχή κατά την διάρκεια αυτής της δεκαετίας. Όντως, μόνο μια χούφτα από τους πρώιμους πρωτοπόρους διατήρησαν την αφοσίωσή τους στα νευρωνικά δίκτυα.

Στην δεκαετία του 1980 έγιναν σημαντικές συνεισφορές στην θεωρία και τον σχεδιασμό των νευρωνικών δικτύων σε πολλά μέτωπα, και εξαιτίας του λόγου αυτού, υπήρξε αναγέννηση του ενδιαφέροντος για τα νευρωνικά δίκτυα. Ίσως περισσότερο από κάθε άλλη δημοσίευση, η ανάπτυξη του αλγορίθμου οπισθοδρομικής διάδοσης του σφάλματος (*backpropagation*) των Rumelhart, Hinton και Williams (1986) [73] ήταν η πιο σημαντική δημοσίευση, υπεύθυνη για την αναγέννηση του ενδιαφέροντος στα νευρωνικά δίκτυα. Σε αυτό τον ίδιο χρόνο, δημοσιεύτηκε το πασίγνωστο δίτομο βιβλίο, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition* που συντάχθηκε από τους Rumelhart και McClelland [74]. Αυτό το πρόσφατο βιβλίο είχε μια μεγάλη επιρροή στη χρήση της μάθησης οπισθοδρομικής διάδοσης του σφάλματος, η οποία εμφανίστηκε στο προσκήνιο ως ο πιο δημοφιλής αλγόριθμος μάθησης για την εκπαίδευση των πολυεπίπεδων perceptron. Στην πραγματικότητα η μάθηση οπισθοδρομικής διάδοσης του σφάλματος ανακαλύφθηκε ανεξάρτητα σε δύο άλλες περιοχές περίπου τον ίδιο καιρό (Parker, 1985 [58], LeCun, 1985 [44]). Μετά την ανακάλυψη του αλγορίθμου οπισθοδρομικής διάδοσης του σφάλματος στα μέσα της δεκαετίας του 1980, αποδείχθηκε ότι είχε περιγραφεί νωρίτερα από τον Werbos στην διδακτορική του διατριβή στο Πανεπιστήμιο Harvard τον Αύγουστο το 1974 [86]. Η βασική ιδέα της οπισθοδρομικής διάδοσης του σφάλματος μπορεί να εντοπιστεί ακόμη πιο πίσω στο βιβλίο *Applied Optimal Control* από τους Bryson και Ho (1969) [13]. Σε τελευταία ανάλυση, ωστόσο, η περισσότερη αναγνώριση για τον αλγόριθμο οπισθοδρομικής διάδοσης του σφάλματος πρέπει να δοθεί στους Rumelhart, Hinton και Williams (1986) [73] για την πρότασή τους να χρησιμοποιηθεί στην μάθηση μηχανής και για την επίδειξη του τρόπου

που μπορεί να γίνει.

Τα νευρωνικά δίκτυα έχουν πλέον διανύσει ένα μακρύ δρόμο από τις πρώιμες μέρες των McCulloch και Pitts. Πραγματικά, καθιερώθηκαν ως ένα περιλαμβάνων διαφορετικές αρχές θέμα με βαθιές ρίζες στις νευροεπιστήμες, την ψυχολογία, τα μαθηματικά, στις φυσικές επιστήμες και της μηχανολογίας. Πολλοί ερευνητές πλέον ασχολούνται με τα νευρωνικά δίκτυα τόσο σε θεωρητικό όσο και πρακτικό επίπεδο. Δεν χρειάζεται να αναφερθεί ότι είναι εδώ για να μείνουν και να συνεχίσουν να αναπτύσσονται σε θεωρία, σχεδιασμό και εφαρμογές.

1.6 Εφαρμογές των Τεχνητών Νευρωνικών Δικτύων

Σήμερα, προβλήματα μεγάλης οικονομικής σημασίας που δεν μπορούσαν να προσεγγιστούν προηγουμένως με κανένα πρακτικό τρόπο, μπορούν τώρα να επιλυθούν με τα τεχνητά νευρωνικά δίκτυα. Προβλήματα στην επιστήμη των μαθηματικών, της φυσικής, αλλά και της ιατρικής, της μηχανικής και της οικονομίας αντιμετωπίζονται με την χρήση των νευρωνικών δικτύων. Η εφαρμογή τους σε μια μεγάλη ποικιλία προβλημάτων σε πολλά πεδία τα έκανε πολύ ελκυστικά. Επίσης, οι γρηγορότεροι ηλεκτρονικοί υπολογιστές και οι γρηγορότεροι αλγόριθμοι έκαναν δυνατή τη χρήση των νευρωνικών δικτύων για την επίλυση πολύπλοκων βιομηχανικών προβλημάτων που μέχρι τώρα απαιτούσαν πολύ μεγάλο υπολογιστικό κόπο.

Η μελέτη DARPA για τα νευρωνικά δίκτυα του 1988, καταγράφει ποικίλες εφαρμογές των νευρωνικών δικτύων, αρχίζοντας από το προσαρμοστικό κανάλι εξισωτή περίπου το 1984. Αυτή η συσκευή, η οποία είχε μια καταπληκτική εμπορική επιτυχία, είναι ένα δίκτυο μονού νευρώνα που χρησιμοποιείται στα τηλεφωνικά συστήματα μακρινής απόστασης για την σταθεροποίηση των φωνητικών σημάτων. Η αναφορά DARPA συνεχίζει να καταγράφει άλλες εμπορικές εφαρμογές, συμπεριλαμβανομένου ενός μικρού αναγνωριστή λέξεων, ενός ελεγκτή διαδικασιών, ενός ταξινομητή σόναρ και ενός συστήματος ανάλυσης κινδύνου.

Τα νευρωνικά δίκτυα έχουν εφαρμοστεί σε πολλά πεδία από τότε που γράφηκε η αναφορά DARPA. Στη συνέχεια ακολουθεί μια λίστα από εφαρμογές που αναφέρονται στην βιβλιογραφία [31].

Αεροδιαστημική. Υψηλής απόδοσης αυτόματοι πιλότοι αεροσκαφών, εξομοιωτές μονοπατιών πτήσης, συστήματα ελέγχου αεροσκαφών, εμπλουτισμός αυτόματων πιλότων, εξομοιώσεις συστατικών αεροσκαφών, ανιχνευτές σφαλμάτων συστατικών αεροσκαφών.

Αυτοκινητοβιομηχανία. Αυτόματα συστήματα κατεύθυνσης αυτοκινήτων, αναλυτές εγγύησης κίνησης.

Τραπεζική Λειτουργία. Αναγνώστες επιταγών και άλλων εγγράφων, αποτιμή-
τες εφαρμογών πίστωσης.

Άμυνα. Πηδαλιούχηση όπλου, εγκλωβισμός στόχου, διάκριση αντικειμένων, αναγνώριση προσώπου, νέα είδη αισθητήρων, σόναρ, ραντάρ και επεξεργασία σημάτων εικόνας συμπεριλαμβανομένου συμπίεσης δεδομένων, εξαγωγή χαρακτηριστικών και καταστολή θορύβου, αναγνώριση σήματος/εικόνας.

Ηλεκτρονική. Πρόβλεψη ακολουθίας κωδικού, διάταξη τσιπ ολοκληρωμένου κυκλώματος, έλεγχος διαδικασίας, ανάλυση αποτυχίας τσιπ, όραση μηχανής, σύνθεση φωνής, μη γραμμική μοντελοποίηση.

Διασκέδαση. Κινούμενα σχέδια, ειδικά εφέ, πρόβλεψη αγοράς.

Οικονομικά. Εκτίμηση ακίνητης περιουσίας, σύμβουλος δανείου, παρακολούθηση υποθήκης, αποτίμηση αξίας εταιρικού γραμματίου, σημείωμα αναφοράς πηγής είδησης χρησιμοποιώντας ανάλυση, χαρτοφυλάκιο εμπορικού προγράμματος, ανάλυση εταιρικών οικονομικών, πρόβλεψη τιμής συναλλάγματος.

Ασφάλεια. Πολιτική εφαρμογής αποτίμησης, βελτιστοποίηση προϊόντος.

Κατασκευές. Έλεγχος διαδικασίας κατασκευής, σχεδίαση και ανάλυση προϊόντος, διάγνωση διαδικασίας και μηχανής, αναγνώριση σωματιδίων σε πραγματικό χρόνο, συστήματα οπτικής επιθεώρησης ποιότητας, δοκιμή μύρας, ανάλυση ποιότητας συγκόλλησης, πρόβλεψη ποιότητας χαρτιού, ανάλυση ποιότητας τσιπ υπολογιστών, ανάλυση λειτουργιών αλέσματος, ανάλυση σχεδιασμού χημικών προϊόντων, ανάλυση συντήρησης μηχανής, δυναμική μοντελοποίηση συστημάτων χημικής διαδικασίας, σχεδιασμός και διαχείριση.

Ιατρική. Ανάλυση καρκινικών ιστών στήθους, ανάλυση EEG και ECG, ανάλυση τεχνητού μέλους, βελτιστοποίηση χρόνων μεταμόσχευσης, μείωση εξόδων νοσοκομείου, βελτίωση ποιότητας νοσοκομείων.

Πετρέλαιο και Αέριο. Εξερεύνηση.

Ρομποτική. Έλεγχος τροχιάς, περονοφόρο ανυψωτικό ρομπότ, ελεγκτές παραπλάνησης, συστήματα όρασης.

Ομιλία. Αναγνώριση ομιλίας, συμπίεση ομιλίας, ταξινόμηση φωνηέντων, σύνθεση ομιλίας από κείμενο.

Τηλεπικοινωνίες. Συμπίεση εικόνας και δεδομένων, αυτόματες υπηρεσίες πληροφόρησης, μετάφραση πραγματικού χρόνου ομιλούμενης γλώσσας, συστήματα επεξεργασίας πληρωμής πελατών.

Μεταφορικά Μέσα. Συστήματα διάγνωσης φρένων φορτηγού, χρονοδιάγραμμα οχημάτων.

Ο αριθμός των εφαρμογών των νευρωνικών δικτύων, τα χρήματα τα οποία έχουν επενδυθεί στο λογισμικό και στην υλική ανάπτυξη των νευρωνικών δικτύων, και το βάθος και το πλάτος του ενδιαφέροντος για αυτές τις συσκευές, αυξάνονται ταχύτατα.

Εκπαίδευση των Τεχνητών Νευρωνικών Δικτύων

Οι μέθοδοι εκπαίδευσης τεχνητών νευρωνικών δικτύων χωρίζονται σε δύο κύριες κατηγορίες, τις μεθόδους εκπαίδευσης με επίβλεψη (supervised learning) και τις μεθόδους εκπαίδευσης χωρίς επίβλεψη (unsupervised learning). Σε αυτή την εργασία αντιμετωπίζονται μόνο μέθοδοι εκπαίδευσης με επίβλεψη.

Οι μέθοδοι εκπαίδευσης με επίβλεψη προϋποθέτουν την παρουσία ενός «δασκάλου». Μπορεί να θεωρηθεί ότι ο δάσκαλος έχει γνώση του περιβάλλοντος που αναπαριστάται από ένα σύνολο παραδειγμάτων εισόδου-εξόδου. Το περιβάλλον είναι, ωστόσο, άγνωστο στο υπό εξέταση νευρωνικό δίκτυο. Ας υποτεθεί τώρα ότι ο δάσκαλος και το νευρωνικό δίκτυο είναι και τα δυο εκτεθειμένα σε ένα διάνυσμα εκπαίδευσης (παράδειγμα) εξαγόμενο από το περιβάλλον. Από το προτέρημα της τμηματικά αυξανόμενης γνώσης, ο δάσκαλος είναι ικανός να παρέχει στο νευρωνικό δίκτυο τις επιθυμητές αποκρίσεις για το διάνυσμα εκπαίδευσης. Πραγματικά, οι επιθυμητές αποκρίσεις αποτελούν την βέλτιστη ενέργεια που πρέπει να εκτελεστεί από το νευρωνικό δίκτυο. Οι παράμετροι του δικτύου ρυθμίζονται κάτω από την συνδυασμένη επιρροή του διανύσματος εκπαίδευσης και του σήματος σφάλματος. Το *σήμα σφάλματος* ορίζεται ως η διαφορά μεταξύ της επιθυμητής απόκρισης και της τρέχουσας απόκρισης του δικτύου. Αυτή η ρύθμιση εκτελείται επαναληπτικά βήμα-βήμα με τη βοήθεια της οποίας τελικά το νευρωνικό δίκτυο *εξομοιώνει* το δάσκαλο. Η εξομοίωση συνεπάγεται ότι είναι η βέλτιστη με μια στατιστική έννοια. Με αυτό τον τρόπο διαθέσιμη γνώση του περιβάλλοντος στον δάσκαλο, μεταφέρεται στο νευρωνικό δίκτυο μέσω της εκπαίδευσης όσο το δυνατό πληρέστερα. Όταν η συνθήκη αυτή ικανοποιείται, μπορεί τότε να απαλλαγεί ο δάσκαλος και να αφαιρεθεί το νευρωνικό δίκτυο να συναλλάσσεται με το περιβάλλον εντελώς μόνο του.

Η μορφή της εκπαίδευσης με επίβλεψη που μόλις περιγράφηκε είναι η εκπαίδευση σφάλματος - διόρθωσης. Είναι ένα σύστημα ανάδρασης κλειστού βρόγχου όπου το ά-

γνωστο περιβάλλον δεν είναι μέσα στον βρόγχο. Ως ένα μέτρο απόδοσης για το σύστημα μπορούμε να θεωρήσουμε το μέσο τετραγωνικό σφάλμα ή το άθροισμα των τετραγώνων των σφαλμάτων πάνω στο δείγμα εκπαίδευσης, ορισμένο ως μια συνάρτηση των ελεύθερων παραμέτρων του συστήματος. Αυτή η συνάρτηση μπορεί να απεικονιστεί ως μια πολυδιάστατη επιφάνεια σφάλματος - απόδοσης ή απλά επιφάνεια σφάλματος, με τις ελεύθερες παραμέτρους ως συντεταγμένες. Η πραγματική επιφάνεια σφάλματος είναι στρογγυλεμένη κατά μέσο όρο πάνω σε όλα τα πιθανά παραδείγματα εισόδου-εξόδου. Κάθε δοθείσα λειτουργία του συστήματος κάτω από την επίβλεψη του δασκάλου αναπαρίσταται ως ένα σημείο στην επιφάνεια σφάλματος. Για να βελτιώσει το σύστημα την απόδοση του στο χρόνο και επομένως να μάθει από τον δάσκαλο, το σημείο λειτουργίας πρέπει να μετακινηθεί προς τα κάτω διαδοχικά προς ένα ελάχιστο σημείο της επιφάνειας σφάλματος. Το ελάχιστο σημείο μπορεί να είναι ένα τοπικό ελάχιστο ή ένα ολικό ελάχιστο. Ένα σύστημα εκπαίδευσης με επίβλεψη είναι ικανό να κάνει αυτή την ενέργεια με την χρήσιμη πληροφορία που έχει για την κλίση την επιφάνειας σφάλματος σχετικά με την τρέχουσα συμπεριφορά του συστήματος. Η κλίση μιας επιφάνειας σφάλματος σε κάθε σημείο είναι ένα διάνυσμα που δείχνει στην κατεύθυνση της πιο απότομης καθόδου (*steepest descent*). Στην πραγματικότητα, στην περίπτωση της εκπαίδευσης με επίβλεψη, το σύστημα μπορεί να χρησιμοποιήσει μια στιγμιαία εκτίμηση του διανύσματος της κλίσης, παίρνοντας ως δεδομένους τους δείκτες των παραδειγμάτων σε εκείνο το χρόνο. Η χρήση μιας τέτοιας εκτίμησης έχει ως αποτέλεσμα την κίνηση του σημείου λειτουργίας πάνω στην επιφάνεια σφάλματος που είναι τυπικά της μορφής του «τυχαίου περιπάτου». Ωστόσο, δοθέντος ενός αλγορίθμου σχεδιασμένου για την ελαχιστοποίηση της συνάρτησης κόστους, ενός επαρκούς συνόλου από παραδείγματα εισόδου-εξόδου, και αρκετού επιτρεπόμενου χρόνου για να γίνει η εκπαίδευση, ένα σύστημα εκπαίδευσης με επίβλεψη είναι συνήθως ικανό να εκτελέσει λειτουργίες όπως ταξινόμηση προτύπων και προσέγγιση συναρτήσεων.

2.1 Μέθοδοι Εκπαίδευσης

Σύμφωνα με τον αρχιτεκτονικό γράφο του πολυεπίπεδου νευρωνικού δικτύου της Εικόνας 1.10, δίνεται το μαθηματικό μοντέλο της μετάδοσης του σήματος εισόδου στο δίκτυο. Έστω ένα πολυεπίπεδο νευρωνικό δίκτυο του οποίου το l επίπεδο περιέχει N_l νευρώνες, όπου $l = 1, \dots, M$. Τότε το νευρωνικό δίκτυο περιγράφεται από τις ακόλουθες σχέσεις:

$$u_j^l = \sum_{i=1}^{N_{l-1}} w_{ij}^{l-1,l} y_i^{l-1} + b_j, \quad y_j^l = f(u_j^l)$$

όπου u_j^l είναι η είσοδος του j νευρώνα στο l επίπεδο ($j = 1, \dots, N_j$), $w_{ij}^{l-1,l}$, είναι τα βάρη από τον i νευρώνα στο $(l-1)$ επίπεδο στον j νευρώνα στο l επίπεδο, b_j είναι η μεροληψία του j νευρώνα, y_j^l είναι η έξοδος του j νευρώνα που ανήκει στο l επίπεδο και $f(u_j^l)$ είναι η συνάρτηση ενεργοποίησης του j νευρώνα.

Για την απλοποίηση των εξισώσεων, θεωρούμε ότι w είναι το διάνυσμα n διαστάσεως που περιέχει όλα τα συναπτικά βάρη και τις μεροληψίες. Η διαδικασία εκπαίδευσης του νευρωνικού δικτύου μπορεί να πραγματοποιηθεί τώρα ελαχιστοποιώντας την συνάρτηση σφάλματος E , δηλαδή αναζητώντας ένα βέλτιστο διάνυσμα $w^* \in \mathbb{R}^n$ έτσι ώστε

$$w^* = \min_{w^* \in \mathbb{R}^n} E(w) \quad (2.1)$$

όπου

$$E(w) = \frac{1}{2} \sum_{p=1}^P \sum_{j=1}^{N_M} (y_{j,p}^M - t_{j,p})^2 = \sum_{p=1}^P E_p \quad (2.2)$$

όπου $(y_{j,p}^M - t_{j,p})^2$ είναι το τετράγωνο της διαφοράς μεταξύ της τιμής της τρέχουσας απόκρισης στον j νευρώνα του επιπέδου εξόδου για το πρότυπο p και της τιμής της επιθυμητής απόκρισης του δικτύου. Το p είναι ένας δείκτης για τα ζεύγη εισόδου–εξόδου.

Έστω ότι το επαναληπτικό σχήμα ανανέωσης των βαρών έχει την ακόλουθη μορφή

$$w_{k+1} = w_k + \alpha_k d_k, \quad k = 0, 1, 2, \dots \quad (2.3)$$

όπου το $w_0 \in \mathbb{R}^n$ είναι το τυχαίο αρχικό σημείο, α_k είναι ο ρυθμός εκπαίδευσης με $\alpha_k > 0$ και d_k είναι η κατεύθυνση αναζήτησης που ικανοποιεί την συνθήκη καθόδου, δηλαδή $g_k^T d_k \leq 0$. Η κλίση $g_k = \nabla E(w_k)$ μπορεί εύκολα να υπολογιστεί εφαρμόζοντας τον κανόνα της αλυσίδας στα επίπεδα του νευρωνικού δικτύου. Πολλές μέθοδοι τοπικής βελτιστοποίησης έχουν εφαρμοστεί στην εκπαίδευση των τεχνητών νευρωνικών δικτύων [31].

Ο πιο γνωστός αλγόριθμος εκπαίδευσης είναι η οπισθοδρομική διάδοση του σφάλματος (back-propagation) [74], η οποία ελαχιστοποιεί την συνάρτηση σφάλματος χρησιμοποιώντας την κατεύθυνση της πιο απότομης καθόδου (steepest descent), δηλαδή $d_k = -g_k$, και ένα σταθερό (ευρετικά καθορισμένο) ρυθμό εκπαίδευσης α . Στην πράξη, η παράμετρος του ρυθμού εκπαίδευσης α ορίζεται σε μια μικρή τιμή μέσα στο διάστημα $(0, 1)$ έτσι ώστε να εξασφαλιστεί η σύγκλιση του αλγορίθμου της οπισθοδρομικής διάδοσης του σφάλματος και να αποφευχθούν οι ταλαντώσεις όταν η συνάρτηση σφάλματος είναι απότομη και στενή. Ωστόσο, ακόμη και με ένα μικρό ρυθμό εκπαίδευσης, ο αλγόριθμος της οπισθοδρομικής διάδοσης του σφάλματος μπορεί να παρουσιάσει ταλαντώσεις όταν συναντά απότομες κοιλάδες. Επιπλέον, η χρήση μικρού ρυθμού εκπαίδευσης επιβραδύνει σημαντικά την διαδικασία εκπαίδευσης αφού μπορεί να μην είναι κατάλληλος για όλες τις περιοχές της επιφάνειας σφάλματος. Ακόμη, ο αλγόριθμος συνήθως συγχλίνει σε τοπικό ελάχιστο παρά σε ένα ολικό, αφού η επιφάνεια που ορίζεται από την συνάρτηση σφάλματος στον χώρο των βαρών μπορεί να είναι ανώμαλη με πολλά τοπικά ελάχιστα. Άλλωστε, σε πραγματικά προβλήματα, η διάσταση του διανύσματος των βαρών w μπορεί να είναι τόσο μεγάλη ώστε ο χρόνος που απαιτεί ο αλγόριθμος της οπισθοδρομικής διάδοσης του σφάλματος για την εκπαίδευση του δικτύου να είναι υπερβολικά μεγάλος, γεγονός που καθιστά τον αλγόριθμο μη πρακτικό.

Προκειμένου να περιοριστεί το εγγενές πρόβλημα του σταθερού ρυθμού εκπαίδευσης του αλγορίθμου της οπισθοδρομικής διάδοσης του σφάλματος, ερευνήθηκε η περίπτωση της αλλαγής του ρυθμού εκπαίδευσης δυναμικά κατά την διάρκεια της εκπαίδευσης. Πολλές στρατηγικές για την προσαρμογή του ρυθμού εκπαίδευσης ερευνήθηκαν από τους ερευνητές της περιοχής. Ο Vogl [84] πρότεινε ένα προσαρμοστικό αλγόριθμο εκπαίδευσης οπισθοδρομικής διάδοσης του σφάλματος όπου η διαδικασία εκπαίδευσης ξεκινά με ένα μικρό ρυθμό εκπαίδευσης ο οποίος αυξάνεται εκθετικά εάν διαδοχικές επαναλήψεις μειώνουν την τιμή της συνάρτησης σφάλματος, ή μειώνεται γρήγορα εάν σημειωθεί μια σημαντική αύξηση της τιμής της συνάρτησης σφάλματος. Οι Chan και Fallside [15] χρησιμοποίησαν ένα προσαρμοστικό σχήμα για τον ρυθμό εκπαίδευσης όπου η διαδικασία εκπαίδευσης ξεκινά με ένα μικρό ρυθμό εκπαίδευσης ο οποίος αυξάνεται εάν διαδοχικές επαναλήψεις διατηρούν την κλίση της κατεύθυνσης αρκούντως σταθερή, μειώνεται γρήγορα εάν η κλίση της κατεύθυνσης ποικίλλει σε μεγάλο βαθμό σε κάθε επανάληψη. Άλλη μια προσέγγιση του θέματος που χρησιμοποιήθηκε στις εργασίες [4], [3], [35], [36], [79], [62] και [70] ήταν η σύνδεση ενός ατομικού ρυθμού εκπαίδευσης σε κάθε βάρος. Ο ατομικός ρυθμός εκπαίδευσης αυξάνεται στην περίπτωση όπου διαδοχικές αλλαγές στα βάρη συμβαίνουν στην ίδια κατεύθυνση, η μειώνεται διαφορετικά. Επίσης, κλειστοί τύποι για τον υπολογισμό της παραμέτρου του ρυθμού εκπαίδευσης προτάθηκαν στις εργασίες [34], [48] και [63]. Κλειστοί τύποι για τον υπολογισμό του ατομικού ρυθμού εκπαίδευσης για κάθε βάρος προτάθηκαν στις εργασίες [23], [49] και [85]. Ωστόσο, οι περισσότερες από τις παραπάνω στρατηγικές χρησιμοποιούν ευρετικές παραμέτρους ώστε να επιβάλλουν μείωση στην συνάρτηση σφάλματος σε κάθε επανάληψη και να εξασφαλίσουν την ασφαλή σύγκλιση του αλγορίθμου εκπαίδευσης. Αυτές οι ευρετικές παράμετροι είναι εξαρτώμενες από τον χρήστη και συνήθως πρέπει να ρυθμιστούν για κάθε πρόβλημα.

Άλλες προσεγγίσεις για την αποτελεσματική εκπαίδευση των νευρωνικών δικτύων έχουν προταθεί από την θεωρία της αριθμητικής βελτιστοποίησης. Πιο συγκεκριμένα, έχουν χρησιμοποιηθεί μέθοδοι δεύτερης τάξης. Στις εργασίες [8], [59], και [6] έχουν προταθεί αλγόριθμοι εκπαίδευσης βασισμένες στην μέθοδο Newton και σε διάφορες τροποποιήσεις της. Αυτοί οι αλγόριθμοι αποδίδουν καλά όταν το αρχικό σημείο βρίσκεται μέσα σε κυρτή περιοχή και συγκλίνουν γρήγορα αν η περιοχή είναι τετραγωνική, ή περίπου τετραγωνική. Στην πράξη όμως υπάρχουν πολλά μειονεκτήματα στην χρήση μεθόδων Newton που τις κάνουν ακατάλληλες για την εκπαίδευση νευρωνικών δικτύων. Για να συγκλίνουν οι μέθοδοι Newton απαιτείται μια καλή αρχική εκτίμηση της λύσης, η οποία δεν είναι διαθέσιμη. Επιπλέον, κάθε επανάληψη απαιτεί τον υπολογισμό του Εσσιανού πίνακα και του αντιστρόφου του, το οποίο είναι υπολογιστικά πολύ ακριβό ($O(n^3)$ πολυπλοκότητα ανά επανάληψη) και οι απαιτήσεις σε χώρο αποθήκευσης είναι πολύ μεγάλες ($O(n^2)$ εκχωρήσεις μνήμης), ειδικά σε μεγάλης κλίμακας προβλήματα. Τέλος, για μια μη κυρτή συνάρτηση, όπως η συνάρτηση σφάλματος (2.2), οι μέθοδοι Newton μπορεί να συγκλίνουν σε ένα τοπικό μέγιστο, σε ένα σαγματικό σημείο ή σε ένα τοπικό ελάχιστο [57].

Οι μέθοδοι ψευδό-Newton (Quasi-Newton) έχουν επίσης χρησιμοποιηθεί για την

εκπαίδευση των τεχνητών νευρωνικών δικτύων. Συνδυάζουν την μέθοδο Newton με ένα αλγόριθμο ολικής σύγκλισης όπως η γραμμική αναζήτηση [57]. Η πιο γνωστή και ευρέως χρησιμοποιούμενη μέθοδος αυτής της κατηγορίας είναι η Broyden–Fletcher–Goldfarb–Shanno (BFGS) [8], [6]. Αυτή η μέθοδος έχει πολλά πλεονεκτήματα. Έχει καλές ιδιότητες σύγκλισης τόσο στην θεωρία όσο και στην πράξη. Υπολογίζει μόνο μια προσέγγιση του αντιστρόφου του Εσσιανού πίνακα, που σημαίνει ότι δεν απαιτείται ο υπολογισμός των παραγώγων δευτέρας τάξεως. Επίσης, ο προσεγγιστικός αντίστροφος Εσσιανός πίνακας της BFGS είναι συμμετρικός και θετικά ορισμένος, γεγονός που καθιστά τον αλγόριθμο αριθμητικά σταθερό. Ωστόσο, η BFGS και οι υπόλοιποι ψευδό–Newton μέθοδοι έχουν επίσης μερικά σοβαρά μειονεκτήματα στην εκπαίδευση των νευρωνικών δικτύων. Ένα από αυτά είναι η ανάγκη για τον υπολογισμό της προσέγγισης του αντιστρόφου του Εσσιανού πίνακα σε κάθε επανάληψη, που καθιστά τις μεθόδους αυτές υπολογιστικά ακριβές ($O(n^2)$ πολυπλοκότητα ανά επανάληψη). Επίσης η αποθήκευση του πίνακα σε μεγάλης κλίμακας προβλήματα αυξάνει τις απαιτήσεις σε μνήμη του αλγορίθμου σημαντικά ($O(n^2)$ εκχωρήσεις μνήμης). Γι’ αυτούς τους λόγους οι μέθοδοι ψευδό–Newton μπορούν να εφαρμοστούν σε μικρά και μεσαίας τάξεως προβλήματα [10].

Στην εργασία [6], ο Battiti πρότεινε μια νέα ψευδό–Newton μέθοδο χωρίς μνήμη, που την ονόμασε One Step Secant (OSS). Αυτή η μέθοδος είναι υλοποιημένη χωρίς την χρήση πίνακα. Γι’ αυτό το λόγο, η αποθήκευση και το υπολογιστικό κόστος, ειδικά σε προβλήματα μεγάλης κλίμακας, έχει μειωθεί σημαντικά ($O(n)$ εκχωρήσεις μνήμης και πολυπλοκότητα ανά επανάληψη). Στην εργασία [6] φαίνεται ότι η μέθοδος OSS μπορεί να είναι εξαιρετικά ανταγωνιστική με την κλασική BFGS μέθοδο στην εκπαίδευση μεγάλων τεχνητών νευρωνικών δικτύων. Ωστόσο, εξαιτίας της φύσεως της χωρίς μνήμης προσέγγισης, η ποσότητα της πληροφορίας δευτέρας τάξεως στην OSS έχει μειωθεί σε σχέση με την κλασική BFGS μέθοδο. Βασισμένος στην ιδέα της OSS μεθόδου, ο Di Fiore στην εργασία [21] παρουσίασε μια γενίκευση της μεθόδου OSS. Πιο συγκεκριμένα, πρότεινε δυο αλγορίθμους που τους ονόμασε Generalized Battiti 1 και 2 (GB1 και GB2). Οι μέθοδοι GB περιέχουν περισσότερη πληροφορία δευτέρας τάξεως και συχνά παρουσιάζονται ανώτερες της κλασικής μεθόδου OSS. Άλλη μια τροποποίηση της BFGS μεθόδου είναι οι περιορισμένης μνήμης BFGS μέθοδοι (L–BFGS) [2], [46], [57]. Οι L–BFGS μέθοδοι ενημερώνουν συνεχώς μια προσέγγιση του Εσσιανού πίνακα χρησιμοποιώντας την πιο πρόσφατη πληροφορία δευτέρας τάξεως που είναι διαθέσιμη με τη μορφή διανυσμάτων. Ο ρυθμός σύγκλισης των μεθόδων L–BFGS μπορεί να βελτιωθεί αν χρησιμοποιηθεί περισσότερη πληροφορία (μεγαλύτερος αριθμός διανυσμάτων). Ωστόσο, περισσότερα διανύσματα αυξάνουν τις απαιτήσεις σε μνήμη των μεθόδων ($O(mn)$ εκχωρήσεις μνήμης, όπου m είναι ο αριθμός των διανυσμάτων). Ο Bortoletti στην εργασία [11] εισήγαγε μια νέα κατηγορία μεθόδων ψευδό–Newton, τις $\mathcal{LQN}$. Αυτές οι μέθοδοι χρησιμοποιούν κατάλληλες προσεγγίσεις ολόκληρου του Εσσιανού πίνακα που καθορίζονται από μια άλγεβρα $\mathcal{L}$ πινάκων διαγωνοποιημένων από γρήγορους μοναδιαίους μετασχηματισμούς [20]. Στην εργασία αυτή παρουσιάστηκε ότι οι μέθοδοι αυτές χρειάζονται $O(n)$ εκχωρήσεις μνήμης και παρουσιάζουν $O(n \log n)$ πολυπλοκότητα ανά επανάληψη. Αυτό επιτρέπει την επίλυση προβλημάτων εκπαίδευσης μεγάλης κλίμακας,

σε αντίθεση με την μέθοδο BFGS.

Μια εναλλακτική πρόταση στους αλγορίθμους εκπαίδευσης που προτάθηκαν προηγουμένως είναι οι γνωστές μέθοδοι συζυγών κλίσεων (Conjugate Gradient (CG)). Είναι κοινά αποδεκτό στην κοινότητα της αριθμητικής ανάλυσης ότι οι μέθοδοι συζυγών κλίσεων είναι οι πιο κατάλληλες μέθοδοι για την επίλυση προβλημάτων μεγάλης κλίμακας γιατί απαιτούν $O(n)$ εκχωρήσεις μνήμης και έχουν $O(n)$ πολυπλοκότητα ανά επανάληψη [10]. Πολλές μέθοδοι συζυγών κλίσεων έχουν προταθεί πρόσφατα ως αλγόριθμοι εκπαίδευσης τεχνητών νευρωνικών δικτύων [37], [38], [52]. Οι πιο δημοφιλείς και ευρέως χρησιμοποιούμενοι αλγόριθμοι εκπαίδευσης από αυτή την κατηγορία μεθόδων είναι οι Hestenes-Stiefel (HS) [33], Fletcher-Reeves (FR) [26] και Polak-Ribière (PR) [65]. Έχει αποδειχθεί ότι αυτοί οι αλγόριθμοι παρέχουν σταθερή εκπαίδευση, ανθεκτικότητα στις ταλαντώσεις και βελτιωμένους ρυθμούς σύγκλισης. Αυτό προέρχεται από το γεγονός ότι η κατεύθυνση αναζήτησης συζυγών κλίσεων λαμβάνει υπόψη την προηγούμενη κατεύθυνση και ότι ο ρυθμός εκπαίδευσης καθορίζεται από μια αποτελεσματική μέθοδο γραμμικής αναζήτησης [57]. Η χρησιμοποίηση των τεχνικών γραμμικής αναζήτησης προϋποθέτει μερικούς επιπρόσθετους υπολογισμούς της συνάρτησης σφάλματος και τις κλίσεις της σε κάθε επανάληψη, αλλά ο υπολογιστικός αυτός κόπος είναι σχετικά μικρός ακόμη και σε προβλήματα μεγάλης κλίμακας. Στην εργασία [52] παρουσιάζεται μια άλλη εκδοχή της μεθόδου συζυγών κλίσεων όπου η γραμμική αναζήτηση αντικαθίσταται από μια κλιμακοποίηση της παραμέτρου του ρυθμού εκπαίδευσης που εξαρτάται από την επιτυχημένη μείωση του σφάλματος και την πιστότητα μιας μονοδιάστατης τετραγωνικής προσέγγισης. Αυτή η μέθοδος καλείται Scaled Conjugate Gradient (SCG) και ενσωματώνει ιδέες από τις μεθόδους εμπιστοσύνης περιοχής trust region και κάποιες διαδικασίες ασφαλείας που απουσιάζουν από τις κλασικές μεθόδους συζυγών κλίσεων. Ωστόσο, ακόμη και χωρίς την χρήση τεχνικών γραμμικής αναζήτησης, ο αλγόριθμος εκπαίδευσης SCG δεν είναι ανώτερος των κλασικών αλγορίθμων εκπαίδευσης συζυγών κλίσεων όταν χρησιμοποιείται ένας κατάλληλος αρχικός ρυθμός εκπαίδευσης και μια αποτελεσματική τεχνική γραμμικής αναζήτησης.

Όπως μπορεί να γίνει εύκολα αντιληπτό, από την εποχή της δημοσίευσης της εργασίας των Rumelhart, Hinton και Williams [73] το 1986, όπου παρουσιάστηκε ο αλγόριθμος εκπαίδευσης της οπισθοδρομικής διάδοσης του σφάλματος, οι ερευνητές της περιοχής έχουν αναπτύξει πάρα πολλές μεθόδους για την εκπαίδευση των νευρωνικών δικτύων. Μέχρι σήμερα δεν έχει υπάρξει κάποια μέθοδος η οποία να μπορεί να επιλύσει όλου του είδους τα προβλήματα εκπαίδευσης με βέλτιστο τρόπο. Κάθε μέθοδος μπορεί να επιλύσει κάποια υποκατηγορία προβλημάτων καλύτερα από τις υπόλοιπες. Στην παρούσα διατριβή θα παρουσιαστούν μερικοί νέοι αλγόριθμοι εκπαίδευσης τεχνητών νευρωνικών δικτύων οι οποίοι φαίνεται να μπορούν να επιλύσουν ορισμένα προβλήματα με τον καλύτερο δυνατό τρόπο.

2.2 Επιφάνεια Σφάλματος και ο Χώρος των Βαρών

Η εκπαίδευση των τεχνητών νευρωνικών δικτύων, δηλαδή η ελαχιστοποίηση της συνάρτησης σφάλματος (2.2), είναι ένα πολύ δύσκολο πρόβλημα. Ας θεωρήσουμε την απλή περίπτωση, όπου το νευρωνικό δίκτυο αποτελείται από ένα μόνο νευρώνα με σιγμοειδή συνάρτηση ενεργοποίησης. Το δίκτυο δέχεται μια μόνο είσοδο, που σημαίνει ότι το διάνυσμα των βαρών $w \in \mathbb{R}^2$. Ο σκοπός του δικτύου είναι να μάθει 8 ζεύγη εισόδου-εξόδου. Στην Εικόνα 2.1 απεικονίζεται η επιφάνεια της συνάρτησης σφάλματος πάνω στον χώρο των βαρών.

Σχήμα 2.1: Η επιφάνεια της συνάρτησης σφάλματος πάνω στον χώρο των βαρών σε ένα νευρωνικό δίκτυο με ένα μόνο νευρώνα.

Το επιθυμητό ελάχιστο βρίσκεται στο κέντρο του σχήματος ενώ υπάρχουν δύο κοιλάδες (δεξιά και αριστερά του σχήματος) που οδηγούν σε τοπικά ελάχιστα. Κατά την εκπαίδευση του νευρωνικού δικτύου, υπάρχει πιθανότητα ο αλγόριθμος εκπαίδευσης να συγκλίνει στα ανεπιθύμητα τοπικά ελάχιστα. Στο παράδειγμα αυτό ο χώρος των βαρών και η επιφάνεια της συνάρτησης σφάλματος είναι σχετικά απλές. Σε προβλήματα πολύ

μεγάλης κλίμακας ο χώρος των βαρών γίνεται πιο πολύπλοκος. Η συνάρτηση σφάλματος πλέον παρουσιάζει στενές περιοχές με πολλά τοπικά ελάχιστα όπου είναι πιο εύκολο να παγιδευτούν οι μέθοδοι εκπαίδευσης.

Επίσης σημαντικό ρόλο παίζει και ο αριθμός των προτύπων εκπαίδευσης. Όσο περισσότερα είναι τα πρότυπα εκπαίδευσης τόσο πιο πολύπλοκος γίνεται ο χώρος των βαρών και η συνάρτηση σφάλματος. Δημιουργούνται μεγάλες επίπεδες περιοχές με σχεδόν μηδενική κλίση και απότομες χαράδρες με πολλά τοπικά ελάχιστα που κάνουν την εκπαίδευση των τεχνητών νευρωνικών δικτύων πραγματική πρόκληση. Στην Εικόνα 2.3 φαίνεται η επίδραση που έχει ο αριθμός των προτύπων στην επιφάνεια σφάλματος στην περίπτωση του απλού νευρωνικού δικτύου, αποτελούμενο από ένα νευρώνα.

Σχήμα 2.2: Η επιφάνεια της συνάρτησης σφάλματος πάνω στον χώρο των βαρών σε ένα νευρωνικό δίκτυο με ένα μόνο νευρώνα υπό την επίδραση: α. ενός προτύπου, β. δύο προτύπων, γ. τριών προτύπων και δ. τεσσάρων προτύπων.

2.2.1 Αρχικοποίηση των Βαρών

Οι αρχικές τιμές των συναπτικών βαρών παίζουν πολύ σημαντικό ρόλο στην επιτυχία της εκπαίδευσης του τεχνητού νευρωνικού δικτύου. Ας θεωρήσουμε για μια ακόμη φορά την απλή περίπτωση, όπου το νευρωνικό δίκτυο αποτελείται από ένα μόνο νευρώνα και η επιφάνεια της συνάρτησης σφάλματος δίνεται από την Εικόνα 2.1. Ξεκινώντας από δύο διαφορετικά σημεία πάνω στον χώρο των βαρών, το νευρωνικό δίκτυο εκπαιδεύεται δύο φορές χρησιμοποιώντας τον αλγόριθμο εκπαίδευσης της οπισθοδρομικής διάδοσης του σφάλματος σταθεροποιώντας και τις δύο φορές όλες τις υπόλοιπες παραμέτρους. Όπως φαίνεται από την Εικόνα 2.3, ο αλγόριθμος εκπαίδευσης συμπεριφέρεται διαφορετικά, ανάλογα με το αρχικό σημείο.

Σχήμα 2.3: Ισοϋψείς της επιφάνειας της συνάρτησης σφάλματος και τροχιά του αλγόριθμου εκπαίδευσης. Αριστερά ο αλγόριθμος συγκλίνει σε ένα τοπικό ελάχιστο, ενώ δεξιά ο αλγόριθμος συγκλίνει στο ολικό ελάχιστο.

Στην πρώτη περίπτωση (αριστερά), ο αλγόριθμος εκπαίδευσης αποκλίνει από το ολικό ελάχιστο και παγιδεύεται σε μια κοιλάδα που οδηγεί σε ένα τοπικό ελάχιστο. Στην δεύτερη περίπτωση (δεξιά) ο αλγόριθμος εκπαίδευσης συγκλίνει γρήγορα στο ολικό ελάχιστο.

Η επιλογή αρχικών τιμών για τα συναπτικά βάρη είναι ένα πολύ δύσκολο και ανοιχτό θέμα. Όταν τα συναπτικά βάρη αρχικοποιούνται με μεγάλες τιμές, είναι πολύ πιθανό οι νευρώνες του δικτύου να οδηγηθούν σε κορεσμό. Σε αυτή την περίπτωση, η κλίση της συνάρτησης παίρνει πολύ μικρές τιμές γεγονός που επιβραδύνει τον αλγόριθμο

εκπαίδευσης σημαντικά. Ωστόσο, όταν τα συναπτικά βάρη αρχικοποιούνται με μικρές τιμές ο αλγόριθμος εκπαίδευσης μπορεί να λειτουργεί σε μια πολύ επίπεδη περιοχή κοντά στην αρχή των συντεταγμένων της επιφάνειας σφάλματος, ειδικά στην περίπτωση που χρησιμοποιούνται αντισυμμετρικές συναρτήσεις ενεργοποίησης, όπως η υπερβολική εφαπτομένη. Δυστυχώς όμως, η αρχή των συντεταγμένων είναι ένα *σαγματικό σημείο* (*saddle point*), το οποίο αναφέρεται σε ένα στάσιμο σημείο όπου η καμπυλότητα της επιφάνειας σφάλματος εγκάρσια του σάγματος είναι αρνητική και κατά μήκος του σάγματος είναι θετική. Γι' αυτούς τους λόγους πρέπει να αποφεύγεται η χρήση μεγάλων και μικρών τιμών για την αρχικοποίηση των συναπτικών βαρών. Η σωστή επιλογή για την αρχικοποίηση βρίσκεται κάπου μεταξύ αυτών των δυο ακραίων περιπτώσεων.

Ας υποθέσουμε ότι ένα πολυεπίπεδο νευρωνικό δίκτυο χρησιμοποιεί ως συνάρτηση ενεργοποίησης για τους νευρώνες του την υπερβολική εφαπτομένη. Έστω ότι η μεροληψία που εφαρμόζεται σε κάθε νευρώνα του δικτύου είναι ίση με μηδέν. Μπορούμε τότε να περιγράψουμε το επηρεασμένο τοπικό πεδίο του νευρώνα j ως

$$v_j = \sum_{i=1}^m w_{ji} y_i$$

Έστω ότι οι είσοδοι που εφαρμόζονται σε κάθε νευρώνα στο δίκτυο έχουν μηδενικό μέσο και μοναδιαία διακύμανση, όπως φαίνεται από την

$$\mu_y = E[y_i] = 0 \quad \text{για όλα τα } i$$

και την

$$\sigma_y^2 = E[(y_i - \mu_i)^2] = E[y_i^2] = 1 \quad \text{για όλα τα } i$$

Έστω ότι υποθέτουμε επιπλέον ότι οι είσοδοι είναι μη συσχετιζόμενες, όπως φαίνεται από την

$$E[y_i y_k] = \begin{cases} 1 & \text{για } k = i \\ 0 & \text{για } k \neq i \end{cases}$$

και ότι τα συναπτικά βάρη επιλέγονται από ένα ομοιόμορφα κατανομημένο σύνολο από αριθμούς με μηδενικό μέσο

$$\mu_w = E[w_{ji}] = 0 \quad \text{για όλα τα ζεύγη } (j, i)$$

και διακύμανση

$$\sigma_w^2 = E[(w_{ji} - \mu_w)^2] = E[w_{ji}^2] \quad \text{για όλα τα ζεύγη } (j, i)$$

Επομένως μπορούμε να εκφράσουμε τον μέσο και την διακύμανση του επηρεασμένου τοπικού πεδίου v_j ως

$$\mu_v = E[v_j] = E\left[\sum_{i=1}^m w_{ji} y_i\right] = \sum_{i=1}^m E[w_{ji}] E[y_i] = 0$$

και

$$\begin{aligned}
\sigma_v^2 &= E[(v_j - m_v)^2] = E[v_j^2] \\
&= E\left[\sum_{i=1}^m \sum_{k=1}^m w_{ji} w_{jk} y_i y_k\right] \\
&= \sum_{i=1}^m \sum_{k=1}^m E[w_{ji} w_{jk}] E[y_i y_k] \\
&= \sum_{i=1}^m E[w_{ji}^2] \\
&= m\sigma_w^2
\end{aligned} \tag{2.4}$$

όπου m είναι ο αριθμός των συναπτικών συνδέσεων ενός νευρώνα.

Σύμφωνα με τον παραπάνω αποτέλεσμα, μπορούμε να περιγράψουμε μια καλή στρατηγική για την αρχικοποίηση των συναπτικών βαρών έτσι ώστε η τυπική απόκλιση του επηρεασμένου τοπικού πεδίου ενός νευρώνα να βρίσκεται στην περιοχή μετάβασης μεταξύ του γραμμικού και των κορεσμένων τμημάτων της σιγμοειδούς συνάρτησης ενεργοποίησης του. Για παράδειγμα, για την περίπτωση μιας συνάρτησης υπερβολικής εφαιτομένης της μορφής

$$f(v) = \alpha \tanh(bv)$$

όπου $\alpha = 1.7159$ και $b = 2/3$, αυτός ο στόχος ικανοποιείται θέτοντας $\sigma_v = 1$ στην εξίσωση (2.4), στην οποία περίπτωση παίρνουμε (LeCun, 1993 [45])

$$\sigma_w = m^{-1/2}$$

Έτσι είναι επιθυμητό για την ομοιόμορφη κατανομή, από την οποία επιλέχθηκαν τα συναπτικά βάρη, να έχει μηδενικό μέσο και διακύμανση ίση με τον αντίστροφο του αριθμού των συναπτικών συνδέσεων ενός νευρώνα.

Μια άλλη ιδιαίτερα αποτελεσματική τεχνική αρχικοποίησης των βαρών προτάθηκε από τους Nguyen και Widrow [55]. Αυτή η τεχνική αποτρέπει τον πρόωρο κορεσμό στους κρυφούς νευρώνες υπολογίζοντας το διάστημα στο οποίο λαμβάνονται τα βάρη που συνδέουν τις εισόδους με τους κρυφούς νευρώνες σύμφωνα με τον αριθμό των εισόδων $\mathcal{N}$ και των αριθμό των κρυφών νευρώνων $\mathcal{M}$. Για την υλοποίηση της τεχνικής, πρώτα υπολογίζεται η παράμετρος ρ :

$$\rho = 0.7(\mathcal{M}^{1/\mathcal{N}})$$

Στη συνέχεια επιλέγονται τα βάρη $w^r = (w_{11}^r, \dots, w_{nm}^r, \dots, w_{\mathcal{N}\mathcal{M}}^r)$ τυχαία μέσα από το διάστημα $[-1, 1]$. Τέλος τα βάρη μεταξύ της εισόδου και του κρυφού επιπέδου αρχικοποιούνται σύμφωνα με τη σχέση

$$w_{nm}^0 = \frac{\rho w_{nm}^r}{\|w^r\|}$$

Αυτή η διαδικασία αρχικοποίησης των συναπτικών βαρών έχει ως αποτέλεσμα την κατανομή των βαρών των κρυφών νευρώνων με τέτοιο τρόπο ώστε να είναι πιο πιθανό κάθε πρότυπο να προκαλεί αποδοτική εκπαίδευση των κρυφών νευρώνων, για να επιταχύνεται η σύγκλιση και να αποφεύγετε ο πρόωρος κορεσμός. Σε αυτή την διατριβή, χρησιμοποιείται η τεχνική αρχικοποίησης των συναπτικών βαρών των Nguyen και Widrow εκτός κι αν σημειωθεί διαφορετικά.

Μέρος II

ΝΕΟΙ ΑΛΓΟΡΙΘΜΟΙ ΕΚΠΑΙΔΕΥΣΗΣ ΤΝΔ

Αλγόριθμος Εκπαίδευσης Προσέγγισης Δύο Σημείων

Σε αυτό το κεφάλαιο παρουσιάζεται ένας νέος αλγόριθμος εκπαίδευσης τεχνητών νευρωνικών δικτύων βασισμένος στον αλγόριθμο της οπισθοδρομικής διάδοσης του σφάλματος και την αυτόματη προσαρμογή του ρυθμού εκπαίδευσης χρησιμοποιώντας πληροφορία δύο σημείων. Η κατεύθυνση αναζήτησης είναι πάντα η κατεύθυνση της πιο απότομης καθόδου, αλλά για τον προσδιορισμό του ρυθμού εκπαίδευσης χρησιμοποιούνται προσεγγίσεις δύο σημείων της εξίσωσης χορδής των μεθόδων ψευδό-Newton [5]. Επιπλέον, παράγουμε ένα νέο ρυθμό εκπαίδευσης προσεγγίζοντας μια νέα εξίσωση χορδής, που προτάθηκε από τον Zhang [90], η οποία χρησιμοποιεί πληροφορία παραγώγων και συναρτησιακών τιμών. Στην συνέχεια, ένας κατάλληλος μηχανισμός επιλογής του ρυθμού εκπαίδευσης ενσωματώνεται στον αλγόριθμο εκπαίδευσης ώστε να επιλέγεται κάθε φορά ο κατάλληλος ρυθμός εκπαίδευσης. Τέλος, γίνεται μελέτη της σύγκλισης του αλγορίθμου εκπαίδευσης και παρουσιάζονται τα πειραματικά αποτελέσματα για διάφορα προβλήματα εκπαίδευσης.

3.1 Η μέθοδος των Barzilai και Borwein

Ας θεωρήσουμε το πρόβλημα της βελτιστοποίησης χωρίς περιορισμούς

$$\min f(x), x \in \mathbb{R}^n \quad (3.1)$$

όπου η f είναι συνεχής με διαθέσιμες τις παραγώγους πρώτης τάξεως. Ο αλγόριθμος της πιο απότομης καθόδου για την επίλυση του προβλήματος (3.1) είναι ένα επαναληπτικό σχήμα της μορφής

$$x_{k+1} = x_k - \alpha_k g_k \quad (3.2)$$

όπου $g_k = \nabla f(x_k)$ και α_k είναι ένα βήμα.

Είναι γνωστό ότι η κατεύθυνση της πιο απότομης καθόδου έχει την ακόλουθη βέλτιστη ιδιότητα

$$-g_k = \min_{d \in \mathbb{R}^n} \lim_{\alpha \rightarrow 0+} [f(x_k) - f(x_k + \alpha d / \|d\|_2^2)] / \alpha \quad (3.3)$$

Στην κλασική μέθοδο της πιο απότομης καθόδου [14], το βήμα προσδιορίζεται χρησιμοποιώντας ακριβής γραμμική αναζήτηση, δηλαδή

$$\alpha_k = \arg \min_{\alpha} f(x_k - \alpha g_k) \quad (3.4)$$

Ωστόσο, παρόλες τις βέλτιστες ιδιότητες (3.3) και (3.4), η μέθοδος της πιο απότομης καθόδου συγκλίνει πολύ αργά και επηρεάζεται σημαντικά από την κακή κατάσταση του προβλήματος [1].

Οι Barzilai και Borwein [5], πρότειναν μια τροποποίηση του αλγορίθμου της πιο απότομης καθόδου όπου το βήμα υπολογίζεται χρησιμοποιώντας προσεγγίσεις δύο σημείων της εξίσωσης χορδής των μεθόδων ψευδό-Newton. Η μέθοδος αυτή ονομάζεται BB.

Η εξίσωση (3.2) μπορεί να γραφεί ως

$$x_{k+1} = x_k - B_k g_k \quad (3.5)$$

όπου $B_k = \alpha_k I$. Για να έχει ο πίνακας B_k την ιδιότητα ψευδό-Newton, το α_k υπολογίζεται έτσι ώστε να ελαχιστοποιείται η ποσότητα

$$\|s_{k-1} - B_k y_{k-1}\| \quad (3.6)$$

Ως συνέπεια της απαίτησης αυτής προκύπτει

$$\alpha_k = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}} \quad (3.7)$$

όπου $s_{k-1} = x_k - x_{k-1}$ και $y_{k-1} = g_k - g_{k-1}$.

Συμμετρικά, μπορούμε να λύσουμε το πρόβλημα $\min \|B_k^{-1} s_{k-1} - y_{k-1}\|$ ως προς το α_k , το οποίο δίνει ως λύση το βήμα

$$\alpha_k = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}} \quad (3.8)$$

Επομένως, η μέθοδος των Barzilai και Borwein δίνεται από τον παρακάτω αλγόριθμο.

Αλγόριθμος 1 Η μέθοδος των Barzilai και Borwein.

BB($x_0, \epsilon, \mu, x_*, f(x_*)$)

- 1: **Για** $k := 0$ **έως** μ **Κάνε**
- 2: Υπολόγισε το $f_k = f(x_k)$ και το $g_k = \nabla f(x_k)$.
- 3: **Αν** $\|g_k\| \leq \epsilon$ **Τότε**
- 4: **Επέστρεψε** $x_* = x_k, f(x_*) = f(x_k)$.
- 5: **Αλλιώς**
- 6: Θέσε $d_k = -g_k$.
- 7: **Αν** $k := 0$ **Τότε**
- 8: Βρες το α_0 χρησιμοποιώντας γραμμική αναζήτηση.
- 9: **Αλλιώς**
- 10: Υπολόγισε το α_k χρησιμοποιώντας την (3.7) ή την (3.8).
- 11: **Τέλος**
- 12: Υπολόγισε $x_{k+1} = x_k + \alpha_k d_k$.
- 13: **Τέλος**
- 14: **Τέλος**
- 15: **Επέστρεψε** Το ελάχιστο δεν βρέθηκε.

Από τον παραπάνω αλγόριθμο φαίνεται ότι δεν χρειάζονται καθόλου υπολογισμοί πινάκων. Επίσης δεν χρειάζονται γραμμικές αναζητήσεις, με εξαίρεση το πρώτο βήμα ($k = 0$). Η μέθοδος των Barzilai και Borwein είναι μια μέθοδος κλίσης που απαιτεί μικρότερο υπολογιστικό κόπο και επιταχύνει σημαντικά την σύγκλιση της μεθόδου της πιο απότομης καθόδου. Οι Barzilai και Borwein έδειξαν ότι ο παραπάνω αλγόριθμος είναι R -υπεργραμμικής σύγκλισης για την περίπτωση των τετραγωνικών συναρτήσεων.

Μια πιο λεπτομερής εξέταση της μεθόδου των Barzilai και Borwein φανερώνει ότι τα βήματα α_k είναι προσεγγίσεις του ηλίκου Rayleigh μιας τιμής $1/\lambda_i$. Έχοντας στο μυαλό την εξίσωση χορδής, $B_k s_{k-1} = y_{k-1}$, μπορεί να επαληθευτεί ότι το ηλίκιο Rayleigh

$$R_q = s_{k-1}^T \left[\int_0^1 \nabla^2 f(x_{k-1} + \tau s_{k-1}) d\tau \right] s_{k-1} / s_{k-1}^T s_{k-1}$$

αντιστοιχεί στον μέσο όρο του Εσσιανού πίνακα

$$\int_0^1 \nabla^2 f(x_{k-1} + ts_k) dt$$

στο ευθύγραμμο τμήμα μεταξύ του x_{k-1} και x_k . Επομένως το R_q είναι μια καλή προσέγγιση της ιδιοτιμής λ_i για την οποία το s_{k-1} είναι το αντίστοιχο ιδιοδιάνυσμα. Επομένως είναι προφανές ότι το βήμα α_k σχετίζεται με τις ιδιοτιμές του Εσσιανού πίνακα στην τιμή του ελαχίστου και όχι με την συναρτησιακή τιμή.

Στην γενική περίπτωση των μη τετραγωνικών συναρτήσεων, ο Raydan [69] πρότεινε την χρησιμοποίηση μιας μη μονότονης στρατηγικής γραμμικής αναζήτησης που ταιριάζει με την μέθοδο των Barzilai και Borwein, αποδεικνύοντας την ολική σύγκλιση της

μεθόδου. Η μη μονότονη συνθήκη του Armijo, που χρησιμοποιείται στην γραμμική αναζήτηση, δίνεται από την σχέση

$$f(x_k - \alpha_k g_k) \leq \max_{0 \leq j \leq \min\{k, M\}} f(x_{k-j}) - \sigma \alpha_k g_k^T g_k \quad (3.9)$$

όπου M είναι ένας ακέραιος αριθμός που ελέγχει το πλήθος των πιο πρόσφατων τιμών της αντικειμενικής συνάρτησης και το $\sigma \in (0, 1)$ είναι ένας μικρός θετικός αριθμός. Αυτή η συνθήκη, που προτάθηκε από τους Grippo, Lampariello και Lucidi [29], επιτρέπει την αποδοχή οποιουδήποτε σημείου που βελτιώνει ικανοποιητικά την μεγαλύτερη από τις πιο πρόσφατες συναρτησιακές τιμές.

Επιπλέον το βήμα α_k που υπολογίζεται από του τύπους (3.7) και (3.8) μπορεί να είναι απαράδεκτα μεγάλο ή μικρό. Σε αυτή την περίπτωση, πρέπει να υποθέσουμε ότι το βήμα α_k ικανοποιεί τη συνθήκη

$$0 < \alpha^{(l)} \leq \alpha_k \leq \alpha^{(u)} \quad \text{για όλα τα } k$$

όπου $\alpha^{(l)}$ και $\alpha^{(u)}$ είναι προκαθορισμένοι αριθμοί.

Αν χρησιμοποιήσουμε την επανάληψη

$$x_{k+1} = x_k - \frac{1}{\alpha_k} g_k = x_k - \lambda_k g_k \quad (3.10)$$

με

$$\alpha_k = \frac{s_{k-1}^T y_{k-1}}{s_{k-1}^T s_{k-1}}, \quad \lambda_k = \frac{1}{\alpha_k} \quad (3.11)$$

τότε, έχοντας υπόψιν ότι $s_k = -\frac{1}{\alpha_k} g_k = -\lambda_k g_k$, έχουμε

$$\alpha_{k+1} = \frac{s_k^T y_k}{s_k^T s_k} = \frac{-\lambda_k g_k^T y_k}{\lambda_k^2 g_k^T g_k} = -\frac{g_k^T y_k}{\lambda_k g_k^T g_k}$$

Η μέθοδος των Barzilai και Borwein με την χρήση της μη μονότονης στρατηγικής αναζήτησης των Grippo, Lampariello και Lucidi δίνεται από τον παρακάτω αλγόριθμο. Σημειώνεται ότι $M \geq 0$, $\sigma \in (0, 1)$, $\delta > 0$ και $0 < \sigma_1 < \sigma_2 < 1$.

Αλγόριθμος 2 Η μέθοδος των Barzilai και Borwein με την χρήση της μη μονότονης στρατηγικής αναζήτησης.

NMBB($x_0, \epsilon, \mu, M, \sigma, \delta, \sigma_1, \sigma_2, \alpha^{(l)}, \alpha^{(u)}, x_*, f(x_*)$)

- 1: **Για** $k := 0$ **έως** μ **Κάνε**
- 2: Υπολόγισε το $f_k = f(x_k)$ και το $g_k = \nabla f(x_k)$.
- 3: **Αν** $\|g_k\| \leq \epsilon$ **Τότε**
- 4: **Επέστρεψε** $x_* = x_k, f(x_*) = f(x_k)$.
- 5: **Αλλιώς**
- 6: **Αν** $\alpha_k \leq \alpha^{(l)}$ **ή** $\alpha_k \geq \alpha^{(u)}$ **Τότε**
- 7: Θέσε $\alpha_k = \delta$.
- 8: **Τέλος**
- 9: Θέσε $\lambda = 1/\alpha_k$.
- 10: **Όσο** $f(x_k - \lambda g_k) > \max_{0 \leq j \leq \min\{k, M\}} f(x_{k-j}) - \sigma \lambda g_k^T g_k$ **Κάνε**
- 11: Επέλεξε $\sigma \in [\sigma_1, \sigma_2]$ και θέσε $\lambda = \sigma \lambda$.
- 12: **Τέλος**
- 13: Θέσε $\lambda_k = \lambda$ και $x_{k+1} = x_k - \lambda_k g_k$.
- 14: Θέσε $\alpha_{k+1} = -(g_k^T y_k) / (\lambda_k g_k^T g_k)$
- 15: **Τέλος**
- 16: **Τέλος**
- 17: **Επέστρεψε** Το ελάχιστο δεν βρέθηκε.

3.2 Νέα Έξισωση Χορδής και Παραγωγή Νέων Βημάτων

Οι μέθοδοι ψευδό-Newton προσεγγίζουν τον Εσσιανό πίνακα $\nabla^2 f(x_k)$ με ένα $n \times n$ συμμετρικό πίνακα B_k ο οποίος υπολογίζεται ενημερώνοντας τον πίνακα B_{k-1} και αξιοποιώντας την διαθέσιμη πληροφορία. Η ενημέρωση βασίζεται στην εξίσωση χορδής που ικανοποιείται από τον πίνακα B_k . Η κλασική εξίσωση χορδής προκύπτει ακολουθώντας την παρακάτω διαδικασία.

Έστω $g(x) = \nabla f(x)$ και $s_{k-1} = x_k - x_{k-1}$. Τότε έχουμε

$$g_k - g_{k-1} = \left[\int_0^1 \nabla^2 f(x_{k-1} + \tau s_{k-1}) d\tau \right] s_{k-1} \quad (3.12)$$

Αφού ο B_k πρέπει να προσεγγίζει τον Εσσιανό πίνακα, η εξίσωση χορδής γίνεται

$$B_k s_{k-1} = y_{k-1} \quad (3.13)$$

όπου $y_{k-1} = g_k - g_{k-1}$.

Προφανώς, μόνο δύο διαδοχικές κλίσεις χρησιμοποιούνται στην εξίσωση χορδής. Ο Zhang [90] παρήγαγε μια νέα εξίσωση χορδής η οποία χρησιμοποιεί πληροφορία κλίσεων και συναρτησιακών τιμών. Πιο συγκεκριμένα, προσπάθησε να προσεγγίσει το $\nabla^2 f(x_k) s_{k-1}$ χρησιμοποιώντας τις τιμές g_{k-1}, g_k, f_{k-1} και f_k .

Ορίζοντας

$$x(\tau) = x_{k-1} + \tau s_{k-1} / \|s_{k-1}\| \quad (3.14)$$

έχουμε

$$\begin{aligned} (d/d\tau)g(x(\tau))|_{\tau=\|s_{k-1}\|} &= \nabla^2 f(x(\tau)) s_{k-1} / \|s_{k-1}\| |_{\tau=\|s_{k-1}\|} \\ &= \nabla^2 f(x_k) s_{k-1} / \|s_{k-1}\| \end{aligned}$$

ή

$$\nabla^2 f(x_k) s_{k-1} = \|s_{k-1}\| (d/d\tau)g(x(\tau))|_{\tau=\|s_{k-1}\|} \quad (3.15)$$

Η εξίσωση (3.15) υποδηλώνει ότι μπορούμε να πάρουμε μια καλή προσέγγιση του $\nabla^2 f(x_k) s_{k-1}$ αν πετύχουμε μια καλή προσέγγιση του $g(x(\tau))$. Υποθέτουμε ότι η $g(x(\tau))$ προσεγγίζεται από ένα πολυώνυμο $q(\tau)$. Αφού η διαθέσιμη πληροφορία περιέχει $2n+2$ στοιχεία, ο βαθμός του πολυωνύμου $q(\tau)$ πρέπει να είναι μεγαλύτερος της μονάδας. Επομένως, υποθέτουμε ότι το $q(\tau)$ είναι μια τετραγωνική συνάρτηση που δίνεται από την παρακάτω σχέση

$$q(\tau) = \alpha\tau^2 + b\tau + c \quad (3.16)$$

όπου $\alpha, b, c \in \mathbb{R}^n$. Λαμβάνοντας υπόψιν την διαθέσιμη πληροφορία, η προσεγγιστική συνάρτηση $q(\tau)$ απαιτείται να ικανοποιεί τις ακόλουθες συνθήκες

$$q(0) = g(x(0)) = g(x_{k-1}) = g_{k-1} \quad (3.17)$$

$$q(\|s_{k-1}\|) = g(x(\|s_{k-1}\|)) = g(x_k) = g_k \quad (3.18)$$

$$\int_0^{\|s_{k-1}\|} q(\tau)^T x'(\tau) d\tau = f_k - f_{k-1} \quad (3.19)$$

όπου η τελευταία εξίσωση προκύπτει από την ταυτότητα

$$\begin{aligned} \int_0^{\|s_{k-1}\|} g(x(\tau))^T x'(\tau) d\tau &= \int_0^{\|s_{k-1}\|} g(x(\tau))^T dx(\tau) \\ &= f(x(\tau))|_0^{\|s_{k-1}\|} \\ &= f(x_k) - f(x_{k-1}) \end{aligned}$$

Οι συνθήκες (3.17)-(3.19) μπορούν να διατυπωθούν ως

$$c = g_{k-1} \quad (3.20)$$

$$\|s_{k-1}\|^2 \alpha + \|s_{k-1}\| b = g_k - g_{k-1} = y_{k-1} \quad (3.21)$$

$$\|s_{k-1}\|^2 \alpha^T s_{k-1} = \gamma \quad (3.22)$$

όπου

$$\gamma = 3(g_k - g_{k-1})^T s_{k-1} - 6(f_k - f_{k-1}) \quad (3.23)$$

Η εξίσωση (3.22) αποφέρει την

$$\|s_{k-1}\|^2 \alpha = \gamma s_{k-1} / \|s_{k-1}\|^2 + v \quad (3.24)$$

όπου το v είναι ένα οποιοδήποτε διάνυσμα στον $\mathbb{R}^n$ κάθετο στο s_{k-1} . Έτσι από τις σχέσεις (3.15), (3.16), (3.21) και (3.24) έχουμε

$$\begin{aligned} \nabla^2 f(x_k) s_{k-1} &= \|s_{k-1}\| (d/d\tau) g(x(\tau))|_{\tau=\|s_{k-1}\|} \\ &\approx \|s_{k-1}\| (d/d\tau) q(\tau)|_{\tau=\|s_{k-1}\|} \\ &= 2\|s_{k-1}\|^2 \alpha + \|s_{k-1}\| b \\ &= y_{k-1} + \gamma s_{k-1} / \|s_{k-1}\|^2 + v \end{aligned} \quad (3.25)$$

όπου το γ και το v ορίζονται από τις (3.23) και (3.24), αντίστοιχα. Επομένως η νέα εξίσωση χορδής έχει τη μορφή

$$B_k s_{k-1} = y_{k-1} + \gamma s_{k-1} / \|s_{k-1}\|^2 + v \quad (3.26)$$

Για να καθορίσουμε το διάνυσμα v , παρατηρούμε ότι η κλασική εξίσωση χορδής (3.13) λειτουργεί αποτελεσματικά στην πράξη. Επομένως, είναι λογικό να κάνουμε το δεξί μέρος της εξίσωσης (3.26) να πλησιάσει το δεξί μέρος της εξίσωσης (3.13). Ελαχιστοποιώντας την διαφορά μεταξύ του $y_{k-1} + \gamma s_{k-1} / \|s_{k-1}\|^2 + v$ και του y_{k-1} έχουμε

$$v = 0 \quad (3.27)$$

Επομένως η νέα εξίσωση χορδής δίνεται από τις σχέσεις

$$B_k s_{k-1} = \tilde{y}_{k-1} \quad (3.28)$$

$$\tilde{y}_{k-1} = y_{k-1} + \gamma s_{k-1} / \|s_{k-1}\|^2 \quad (3.29)$$

$$\gamma = 3(g_k - g_{k-1})^T s_{k-1} - 6(f_k - f_{k-1}) \quad (3.30)$$

Έχοντας προσδιορίσει τώρα την νέα εξίσωση χορδής, είναι δυνατό να υπολογιστούν νέα βήματα με τον ίδιο τρόπο που οι Barzilai και Borwein υπολόγισαν για την κλασική εξίσωση χορδής. Πιο συγκεκριμένα, υποθέτουμε ότι ο πίνακας B_k της εξίσωσης (3.5) μπορεί να γραφεί ως $B_k = \tilde{\alpha}_k I$. Για να ικανοποιεί ο πίνακας B_k την νέα εξίσωση χορδής, το $\tilde{\alpha}_k$ υπολογίζεται έτσι ώστε να ελαχιστοποιείται η ποσότητα

$$\|s_{k-1} - B_k \tilde{y}_{k-1}\| \quad (3.31)$$

Η παραπάνω απαίτηση οδηγεί στην επιλογή

$$\tilde{\alpha}_k = \frac{s_{k-1}^T \tilde{y}_{k-1}}{\tilde{y}_{k-1}^T \tilde{y}_{k-1}} \quad (3.32)$$

όπου $s_{k-1} = x_k - x_{k-1}$ και $y_{k-1} = g_k - g_{k-1}$.

Συμμετρικά, μπορεί να επιλυθεί το πρόβλημα $\min \|B_k^{-1}s_{k-1} - \tilde{y}_{k-1}\|$ ως προς το α_k , το οποίο δίνει ως λύση το βήμα

$$\tilde{\alpha}_k = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T \tilde{y}_{k-1}} \quad (3.33)$$

Τα νέα βήματα (3.32) και (3.33) έχουν μεγαλύτερη ακρίβεια από τα βήματα (3.7) και (3.8) αφού προέρχονται από την νέα εξίσωση χορδής η οποία περιέχει περισσότερη πληροφορία σε σχέση με την κλασική εξίσωση χορδής. Αν η συνάρτηση f είναι τετραγωνική στο ευθύγραμμο τμήμα μεταξύ του x_{k-1} και x_k , τότε οι (3.32) και (3.33) είναι ισοδύναμες με τις (3.7) και (3.8), αντιστοίχως. Δηλαδή, σε αυτή την περίπτωση έχουμε $\tilde{\alpha}_k = \alpha_k$. Επιπλέον, αν η συνάρτηση f είναι αυστηρώς κυρτή, τότε έχουμε $0 \leq \tilde{\alpha}_k \leq 2\alpha_k$.

3.3 Νέος Αλγόριθμος Εκπαίδευσης και η Ολική Σύγκλιση του

Σε αυτή την παράγραφο, δίνεται ένας νέος αλγόριθμος εκπαίδευσης τεχνητών νευρωνικών δικτύων ο οποίος βασίζεται στο βήμα των Barzilai και Borwein και το νέο βήμα που βασίζεται στην νέα εξίσωση χορδής. Επιπλέον, χρησιμοποιείται μια τεχνική γραμμικής αναζήτησης που βασίζεται στην μη μονότονη συνθήκη του Armijo (3.9). Πιο συγκεκριμένα, γίνεται προσπάθεια να ελαχιστοποιηθεί η συνάρτηση σφάλματος $E(w)$ χρησιμοποιώντας το ακόλουθο επαναληπτικό σχήμα

$$w_{k+1} = w_k - \eta_k g_k \quad (3.34)$$

όπου η_k είναι ο ρυθμός εκπαίδευσης και $g_k = \nabla E(w_k)$.

Για τον ρυθμό εκπαίδευσης η_k χρησιμοποιούνται τα βήματα (3.8) και (3.33). Η επιλογή του κατάλληλου ρυθμού εκπαίδευσης κατά την διαδικασία εκπαίδευσης του νευρωνικού δικτύου επιτυγχάνεται χρησιμοποιώντας έναν μηχανισμό αλλαγής του ρυθμού εκπαίδευσης. Εφόσον ο ρυθμός εκπαίδευσης (3.33) είναι υψηλότερης ακρίβειας από τον ρυθμό εκπαίδευσης (3.8), χρησιμοποιείται μια αμέραιη παράμετρος j στον αλγόριθμο για να αποφασιστεί αν αξίζει τον κόπο η χρησιμοποίηση του ρυθμού εκπαίδευσης (3.33). Η τιμή του j εξαρτάται από την ακολουθία u_k , όπου $u_k = \left| \frac{\alpha_k}{\tilde{\alpha}_k} - 1 \right|$ είναι μια ποσότητα που δείχνει πόσο κοντά είναι η $E(w)$ σε μια τετραγωνική συνάρτηση στο ευθύγραμμο τμήμα που ορίζεται από τα w_{k-1} και w_k . Πιο συγκεκριμένα, υποθέτουμε ότι $c_3 > c_2 > c_1$ είναι τρεις θετικές σταθερές. Αν $u_k \leq c_1$, ή $\max(u_k, u_{k-1}) \leq c_2$, ή $\max(u_k, u_{k-1}, u_{k-2}) \leq c_3$, τότε μπορεί να θεωρηθεί ότι η $E(w)$ προσεγγίζει μια τετραγωνική συνάρτηση στο ευθύγραμμο τμήμα που ορίζεται από τα w_{k-1} και w_k , το οποίο υποδηλώνει ότι ο ρυθμός

εκπαίδευσης υψηλότερης ακριβείας είναι καταλληλότερος. Σε αυτή την περίπτωση ο α-κέραιος αριθμός j τίθεται ίσος με 1 και χρησιμοποιείται ο ρυθμός εκπαίδευσης (3.33). Διαφορετικά, τίθεται $j = 0$ και χρησιμοποιείται ο ρυθμός εκπαίδευσης (3.8). Ο μηχανισμός αλλαγής του βήματος περιγράφεται από τον παρακάτω αλγόριθμο.

Αλγόριθμος 3 Μηχανισμός αλλαγής βήματος.

SSM($k, \delta, w_k, w_{k-1}, g_k, g_{k-1}, f_k, f_{k-1}, c_1, c_2, c_3, \text{lambda}$)

- 1: **Αν** $s_{k-1}^T y_{k-1} \leq 0$ **Τότε**
- 2: Θέσε $\lambda = \delta, u_k = 1$ και **επέστρεψε** το λ .
- 3: **Αλλιώς**
- 4: Υπολόγισε τα α_k και $\tilde{\alpha}_k$ χρησιμοποιώντας τις σχέσεις (3.8) και (3.33), αντίστοιχα.
- 5: Υπολόγισε το $u_k = \left| \frac{\alpha_k}{\tilde{\alpha}_k} - 1 \right|$.
- 6: Θέσε $j = 0$.
- 7: **Αν** $u_k \leq c_1$, ή $\max_{i=0,1} u_{k-i} \leq c_2$, ή $\max_{i=0,1,2} u_{k-i} \leq c_3$ **Τότε**
- 8: Θέσε $j = 1$.
- 9: **Τέλος**
- 10: **Αν** $j = 1$ **Τότε**
- 11: Θέσε $\lambda = \tilde{\alpha}_k$.
- 12: **Αλλιώς**
- 13: Θέσε $\lambda = \max(1/\delta, \min(\alpha_k, \delta))$
- 14: **Τέλος**
- 15: **Επέστρεψε** το λ .
- 16: **Τέλος**

Ο παρονομαστής α_k του κλάσματος στην σχέση του u_k μπορεί να γίνει μηδέν. Ωστόσο, από τις σχέσεις (3.8) και (3.33) έχουμε

$$\frac{\alpha_k}{\tilde{\alpha}_k} = \frac{s_{k-1}^T \tilde{y}_{k-1}}{s_{k-1}^T y_{k-1}}$$

το οποίο είναι καλά ορισμένο αφού έχουμε ότι $s_{k-1}^T y_{k-1} > 0$.

Ο παραπάνω μηχανισμός επιλογής του κατάλληλου ρυθμού μάθησης συνδυάζεται με μια στρατηγική γραμμικής αναζήτησης που χρησιμοποιεί την μη μονότονη συνθήκη του Armijo (3.9). Η μέθοδος της γραμμικής αναζήτησης που χρησιμοποιείται στον αλγόριθμο εκπαίδευσης είναι απλή και στηρίζεται στον υποδιπλασιασμό του ρυθμού εκπαίδευσης όταν αυτός δεν ικανοποιεί την μη μονότονη συνθήκη του Armijo (3.9). Η μέθοδος εκπαίδευσης προσέγγισης δύο σημείων δίνεται από τον παρακάτω αλγόριθμο.

Αλγόριθμος 4 Αλγόριθμος Εκπαίδευσης Προσέγγισης Δύο Σημείων.

$ABB(x_0, \epsilon, \mu, M, \sigma, \delta, c_1, c_2, c_3, w_*, E(w_*))$
 1: **Για** $k := 0$ **έως** μ **Κάνε**
 2: Υπολόγισε το $E_k = E(w_k)$ και το $g_k = \nabla E(w_k)$.
 3: **Αν** $E_k \leq \epsilon$ **Τότε**
 4: **Επέστρεψε** $w_* = w_k$, $E(w_*) = E(w_k)$.
 5: **Αλλιώς**
 6: **Αν** $k = 0$ **Τότε**
 7: Θέσε $\lambda = 1/\|g_k\|$.
 8: **Αλλιώς**
 9: Θέσε $\lambda = SSM(k, \delta, w_k, w_{k-1}, g_k, g_{k-1}, f_k, f_{k-1}, c_1, c_2, c_3, \lambda)$
 10: **Τέλος**
 11: **Όσο** $f(x_k - \lambda g_k) > \max_{0 \leq j \leq \min\{k, M\}} f(x_{k-j}) - \sigma \lambda g_k^T g_k$ **Κάνε**
 12: Θέσε $\lambda = \lambda/2$.
 13: **Τέλος**
 14: Θέσε $\eta_k = \lambda$ και $w_{k+1} = w_k - \eta_k g_k$.
 15: **Τέλος**
 16: **Τέλος**
 17: **Επέστρεψε** Το δίκτυο δεν εκπαιδεύτηκε.

Ας υποθεθεί ότι το $\nabla E(w)$ είναι Lipschitz συνεχές. Τότε έχουμε ότι

$$s_{k-1}^T y_{k-1} \leq L s_{k-1}^T s_{k-1} \quad (3.35)$$

όπου L είναι μια θετική σταθερά. Αν το $s_{k-1}^T y_{k-1} > 0$, τότε από τις σχέσεις (3.8) και (3.35) έχουμε ότι

$$\alpha_k \geq L^{-1} \quad (3.36)$$

Από την παραπάνω σχέση και τον Αλγόριθμο 3 έχουμε ότι, αν $j = 1$, τότε πρέπει να έχουμε ότι

$$\lambda = \tilde{\alpha}_k \geq (1 + c_3)^{-1} \alpha_k \geq [(1 + c_3)L]^{-1} \quad (3.37)$$

Διαφορετικά, αν το $j = 0$ ή το λ λαμβάνεται από το δεύτερο βήμα του Αλγόριθμου 3, επίσης έχουμε

$$\lambda \geq \delta^{-1} \quad (3.38)$$

Χρησιμοποιώντας τις σχέσεις (3.37) και (3.38) μπορεί να αποδειχθεί η ολική σύγκλιση του Αλγορίθμου 4 όπως στο Θεώρημα 2.1 στην εργασία [69].

Θεώρημα 3.1. Έστω ότι το σύνολο $\mathcal{L}_0 = \{w \in \mathbb{R}^n : E(w) \leq E(w_0)\}$ είναι φραγμένο και ότι το $\nabla E(w)$ είναι Lipschitz συνεχές σε μια γειτονιά $\mathcal{N}$ του $\mathcal{L}_0$. Έστω ότι $\{w_k\}_{k=0}^\infty$ είναι η ακολουθία που παράγεται από τον Αλγόριθμο 4. Τότε, είτε το $\nabla E(w_k) = 0$ για κάποιο πεπερασμένο k , ή ισχύουν οι ακόλουθες ιδιότητες:

- (i) $\lim_{k \rightarrow \infty} \|\nabla E(w_k)\| = 0$,
- (ii) κανένα οριακό σημείο της $\{w_k\}$ είναι ένα τοπικό μέγιστο της E ,
- (iii) αν ο αριθμός των στάσιμων σημείων της E στο $\mathcal{L}_0$ είναι πεπερασμένος, τότε η ακολουθία $\{w_k\}$ συγκλίνει.

Απόδειξη. Για να αποδειχθεί το (i), μπορεί να χρησιμοποιηθεί το πρώτο μέρος της απόδειξης του θεωρήματος σύγκλισης στην παράγραφο 3 της εργασίας [29].

Ας υποθέσουμε ότι $m(k) = \min(k, M)$. Είναι φανερό ότι $m(0) = 0$ και ότι

$$0 \leq m(k) \leq \min(m(k-1) + 1, M), \quad \text{για } k \geq 1.$$

Επιπλέον υπάρχει μια θετική σταθερά λ τέτοια ώστε $0 < \eta_k \leq \lambda$ για όλα τα k . Όντως για τον Αλγόριθμο 4 το $\lambda = \max(1/\delta, \min(\alpha_k, \delta))$. Επίσης, υπάρχουν θετικοί αριθμοί σ_1 και σ_2 τέτοιοι ώστε η κατεύθυνση αναζήτησης d_k να ικανοποιεί την $g_k^T d_k \leq -\sigma_1 \|g_k\|^2$ και την $\|d_k\| \leq \sigma_2 \|g_k\|$. Όντως, στον Αλγόριθμο 4, η κατεύθυνση αναζήτησης d_k είναι ίση με το $-g_k$ για όλα τα k και επομένως $\sigma_1 = \sigma_2 = 1$. Επομένως, παίρνουμε την εξίσωση (14) από την εργασία [29], η οποία στην περίπτωση μας παίρνει την μορφή

$$\lim_{k \rightarrow \infty} \eta_k \|g_k\| = 0$$

Εφόσον το $\eta_k \geq 1/\delta$ ή $\eta_k \geq [(1 + c_3)L]^{-1}$ (ανάλογα με την τιμή του j) για όλα τα k , ο ισχυρισμός (i) ισχύει. Οι ισχυρισμοί (ii) και (iii) απορρέουν άμεσα από το θεώρημα σύγκλισης στην εργασία [29]. $\square$

3.4 Πειραματικά Αποτελέσματα

Στην παράγραφο αυτή αξιολογείται η επίδοση του προτεινόμενου αλγόριθμου εκπαίδευσης τεχνητών νευρωνικών δικτύων (ABB) [81] και συγκρίνεται με την επίδοση των ακόλουθων γνωστών αλγορίθμων εκπαίδευσης:

1. Τον αλγόριθμο της οπισθοδρομικής διάδοσης του σφάλματος με σταθερό ρυθμό εκπαίδευσης (BP) [74].
2. Τον αλγόριθμο της οπισθοδρομικής διάδοσης του σφάλματος με σταθερό ρυθμό εκπαίδευσης και ορμή (BPM) [74].
3. Τον αλγόριθμο της οπισθοδρομικής διάδοσης του σφάλματος με προσαρμοστικό ρυθμό εκπαίδευσης και ορμή (ABP) [84].
4. Τον μη μονότονο αλγόριθμο εκπαίδευσης των Barzilai και Borwein (BBP) [64].

Για τους αλγόριθμους εκπαίδευσης BBP και ABB οι παράμετροι M , σ και δ έχουν τις τιμές 10 , 10^{-4} και 10^3 , αντίστοιχα. Επιπλέον, οι παράμετροι c_1 , c_2 και c_3 του αλγόριθμου ABB έχουν τις τιμές 0.5×10^{-4} , 0.1 και 0.5 , αντίστοιχα. Οι τιμές των παραπάνω παραμέτρων είναι οι ίδιες για όλα τα πειράματα που πραγματοποιήθηκαν. Και οι δύο αλγόριθμοι χρησιμοποιούν την ίδια μη μονότονη στρατηγική γραμμικής αναζήτησης. Επιπλέον, για την αρχικοποίηση των βαρών και των μεροληψιών εφαρμόστηκε η τεχνική που προτάθηκε από τους Nguyen και Widrow [55] για όλους τους αλγορίθμους.

Τα προβλήματα που χρησιμοποιήθηκαν για την αξιολόγηση της ταχύτητας σύγκλισης του προτεινόμενου αλγόριθμου εκπαίδευσης ABB σε σχέση με τους προαναφερθέντες αλγόριθμους εκπαίδευσης είναι τα ακόλουθα:

1. Το πρόβλημα του Αποκλειστικού-EITE.
2. Το πρόβλημα της ισοτιμίας των 3-bit.
3. Το πρόβλημα της αναγνώρισης των κεφαλαίων γραμμάτων της Αγγλικής αλφαβήτου.
4. Το πρόβλημα της προσέγγισης της συνεχούς τριγωνομετρικής συνάρτησης.

Για κάθε πρόβλημα, παρατίθεται ένας πίνακας ο οποίος συνοψίζει την απόδοση των αλγορίθμων για προσομοιώσεις που έφτασαν σε λύση εντός ενός προκαθορισμένου ορίου υπολογισμών της συνάρτησης σφάλματος που ορίζεται σε κάθε πρόβλημα. Οι αναφερόμενες στους πίνακες παράμετροι είναι: *Min* ο ελάχιστος αριθμός επαναλήψεων, *Max* ο μέγιστος αριθμός επαναλήψεων, μ η μέση τιμή των επαναλήψεων, σ η τυπική απόκλιση, και *Επιτυχία* ο αριθμός των πετυχημένων προσομοιώσεων από ένα σύνολο 1000 δοκιμών. Σε περίπτωση που ένας αλγόριθμος αποτύχει να συγκλίνει μέσα στο προκαθορισμένο όριο των υπολογισμών της συνάρτησης σφάλματος, θεωρείται ότι απέτυχε να εκπαιδεύσει το τεχνητό νευρωνικό δίκτυο και οι επαναλήψεις της συγκεκριμένης προσομοίωσης δεν περιλαμβάνονται στην στατιστική ανάλυση του αλγορίθμου.

Σε αυτό το σημείο, πρέπει να επισημανθεί ότι γενικά μια επανάληψη ενός αλγορίθμου εκπαίδευσης σημαίνει ότι όλα τα πρότυπα εκπαίδευσης έχουν παρουσιαστεί στο τεχνητό νευρωνικό δίκτυο. Στους κλασικούς αλγορίθμους, όπως η BP, κάθε επανάληψη αντιστοιχεί σε ένα συναρτησιακό υπολογισμό της συνάρτησης σφάλματος και έναν υπολογισμό της κλίσης (παράγωγοι πρώτης τάξεως). Στους αλγορίθμους BBP και ABB λαμβάνουν χώρα περισσότεροι συναρτησιακοί υπολογισμοί ανά επανάληψη, γεγονός που οφείλεται στην παρουσία της μη μονότονης τεχνικής γραμμικής αναζήτησης. Εξαιτίας της φύσεως της γραμμικής αναζήτησης που χρησιμοποιείται στους αλγορίθμους αυτούς, ενώ μπορεί να απαιτούνται περισσότεροι συναρτησιακοί υπολογισμοί ανά επανάληψη, αυτό δεν ισχύει για τον υπολογισμό της κλίσεως η οποία υπολογίζεται μια φορά ανά επανάληψη. Λαμβάνοντας υπόψιν ότι οι υπολογισμοί κλίσεως είναι ακριβότεροι

από τους συναρτησιακούς υπολογισμούς του σφάλματος, είναι φανερό ότι οι αλγόριθμοι αυτοί είναι γρηγορότεροι αφού απαιτούν λιγότερες πράξεις κινητής υποδιαστολής. Για τους παραπάνω λόγους, στους πίνακες των αποτελεσμάτων, για τις μεθόδους BBP και ABB αναφέρονται δύο σειρές αποτελεσμάτων, η πρώτη για τους υπολογισμούς των τιμών της συνάρτησης σφάλματος και η δεύτερη για τους υπολογισμούς των κλίσεων.

3.4.1 Αποκλειστικό-EITE

Για το πρόβλημα του αποκλειστικού-EITE (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα $2 - 2 - 1$ τεχνητό νευρωνικό δίκτυο (6 βάρη και 3 μεροληψίες). Το κρυφό επίπεδο του νευρωνικού δικτύου είναι βασισμένο σε νευρώνες με συνάρτηση ενεργοποίησης την υπερβολική εφαπτομένη, ενώ το επίπεδο εξόδου είναι βασισμένο σε νευρώνες με συνάρτηση ενεργοποίησης την γραμμική συνάρτηση. Η συνθήκη τερματισμού είναι $E \leq 0.01$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 1000. Για τις μεθόδους σταθερού ρυθμού εκπαίδευσης (BP και BPM) ο ρυθμός εκπαίδευσης ορίστηκε στην τιμή 0.1, αντί της προεπιλεγμένης τιμής 0.01, έτσι ώστε να επιταχυνθεί η σύγκλιση τους, αφού με την προεπιλεγμένη τιμή συγκλίνουν πολύ αργά σε αυτό το πρόβλημα. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 3.1.

Αλγόριθμος	Min	Max	μ	σ	Επιτυχία
BP	31	975	100.9	124.8	71.1%
BPM	24	962	128.3	161.8	75.7%
ABP	19	865	47.3	69.1	70.8%
BBP	7	998	65.3	131.4	82.7%
	4	595	40.6	79.4	
ABB	7	931	56.1	125.1	82.7%
	6	556	39.2	77.7	

Πίνακας 3.1: Συγκριτικά αποτελέσματα για το πρόβλημα του αποκλειστικού-EITE

Ο αλγόριθμος BPM απέδωσε καλύτερα όσον αφορά το ποσοστό επιτυχίας σε σχέση με τον αλγόριθμο BP, αλλά χρειάστηκε περισσότερες επαναλήψεις. Ο αλγόριθμος ABP, αν και είχε σημαντικά μικρότερο μέσο αριθμό επαναλήψεων, ήταν ο χειρότερος αλγόριθμος σε ποσοστό επιτυχίας. Ο αλγόριθμος εκπαίδευσης BBP ήταν σημαντικά καλύτερος από τους κλασικούς αλγόριθμους εκπαίδευσης τόσο σε ποσοστό επιτυχίας όσο και σε κόστος. Πιο συγκεκριμένα, αν και χρειάστηκε περισσότερους υπολογισμούς των τιμών της συνάρτησης σφάλματος σε σχέση με την ABP, χρειάστηκε αρκετά λιγότερους υπολογισμούς κλίσης. Όσον αφορά τον προτεινόμενο αλγόριθμο εκπαίδευσης ABB, παρουσίασε το ίδιο ποσοστό επιτυχίας με τον αλγόριθμο BBP, αλλά βελτιώθηκε όσον αφορά τους συναρτησιακούς υπολογισμούς και τους υπολογισμούς κλίσεως, γεγονός που τον καθιστά καλύτερο.

3.4.2 Ισοτιμία των 3-bit

Για το πρόβλημα της ισοτιμίας των 3-bit (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα $3 - 2 - 1$ τεχνητό νευρωνικό δίκτυο (8 βάρη και 3 μεροληψίες). Το κρυφό επίπεδο του νευρωνικού δικτύου είναι βασισμένο σε νευρώνες με συνάρτηση ενεργοποίησης την υπερβολική εφαπτομένη, ενώ το επίπεδο εξόδου είναι βασισμένο σε νευρώνες με συνάρτηση ενεργοποίησης την γραμμική συνάρτηση. Η συνθήκη τερματισμού είναι $E \leq 0.01$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 1000. Για τις μεθόδους σταθερού ρυθμού εκπαίδευσης (BP και BPM) ο ρυθμός εκπαίδευσης ορίστηκε στην τιμή 0.1, αντί της προεπιλεγμένης τιμής 0.01, έτσι ώστε να επιταχυνθεί η σύγκλιση τους, αφού με την προεπιλεγμένη τιμή συγκλίνουν πολύ αργά σε αυτό το πρόβλημα. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 3.2.

Αλγόριθμος	Min	Max	μ	σ	Επιτυχία
BP	-	-	-	-	0.0%
BPM	178	995	486.1	203.3	52.6%
ABP	426	959	602.9	123.4	48.5%
BBP	50	999	233.5	192.3	72.9%
	33	615	147.1	115.3	
ABB	50	983	222.4	185.5	76.0%
	33	619	148.4	114.7	

Πίνακας 3.2: Συγκριτικά αποτελέσματα για το πρόβλημα της ισοτιμίας των 3-bit

Ο αλγόριθμος εκπαίδευσης BP απέτυχε να συγκλίνει μέσα στο όριο των υπολογισμών των τιμών της συνάρτησης σφάλματος σε όλες τις προσομοιώσεις. Αξίζει να αναφερθεί ότι δοκιμάσαμε διαφορετικές τιμές για την τιμή της παραμέτρου του ρυθμού εκπαίδευσης, αλλά το αποτέλεσμα ήταν το ίδιο. Ο αλγόριθμος BPM απέδωσε αρκετά καλά, ενώ ήταν καλύτερος τόσο σε ποσοστό επιτυχίας όσο και σε υπολογισμούς τιμών της συνάρτησης σφάλματος (και υπολογισμούς κλίσεως), αναφορικά με τον αλγόριθμο εκπαίδευσης ABP. Ο αλγόριθμος BBP απέδωσε πολύ καλύτερα όσον αφορά το ποσοστό επιτυχίας αφού η διαφορά του σε ποσοστό επιτυχίας από τον αλγόριθμο BPM ήταν 20%. Επιπλέον, είχε πολύ μικρότερο κόστος σε συναρτησιακούς υπολογισμούς και υπολογισμούς κλίσεως. Ο προτεινόμενος αλγόριθμος εκπαίδευσης ABB παρουσίασε το καλύτερο ποσοστό επιτυχίας στο παράδειγμα αυτό. Επίσης, παρουσίασε μικρότερο κόστος σε συναρτησιακούς υπολογισμούς κατά μέσο όρο από την BBP ενώ ο μέσος όρος των υπολογισμών κλίσεως ήταν περίπου ο ίδιος.

3.4.3 Αναγνώριση Κεφαλαίων Γραμμάτων

Για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων της Αγγλικής αλφαβήτου (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα $35 - 30 - 26$ τεχνητό νευρωνικό δίκτυο (1830 βάρη και 56 μεροληψίες). Το κρυφό επίπεδο του νευρωνικού δικτύου είναι βασισμένο σε νευρώνες με λογιστική συνάρτηση ενεργοποίησης, ενώ το επίπεδο εξόδου είναι βασισμένο σε νευρώνες γραμμικής συνάρτησης ενεργοποίησης. Η συνθήκη τερματισμού είναι $E \leq 0.1$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 2000. Για τις μεθόδους σταθερού ρυθμού εκπαίδευσης (BP και BPM) ο ρυθμός εκπαίδευσης ορίστηκε στην προεπιλεγμένη τιμή 0.01. Η αρχικοποίηση των βαρών έγινε με την τεχνική των Nguyen και Widrow, αλλά στη συνέχεια τα βάρη μεταξύ του κρυφού επιπέδου και του επιπέδου της εξόδου πολλαπλασιάστηκαν με 0.01. Τα αποτελέσματα των αριθμητικών πειραμάτων παρουσιάζονται στον Πίνακα 3.3.

Αλγόριθμος	Min	Max	μ	σ	Επιτυχία
BP	1044	1998	1548.4	199.7	75.6%
BPM	1199	1899	1560.9	183.5	3.6%
ABP	1230	1999	1776.5	157.5	16.9%
BBP	101	362	178.9	37.5	100.0%
	64	222	116.1	22.1	
ABB	103	350	168.7	32.2	100.0%
	68	215	108.2	20.3	

Πίνακας 3.3: Συγκριτικά αποτελέσματα για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων

Από τους κλασικούς αλγόριθμους εκπαίδευσης, ο BP παρουσιάζει την καλύτερη επίδοση τόσο σε ποσοστό επιτυχίας όσο και σε υπολογισμούς της τιμής της συνάρτησης σφάλματος. Ειδικά η BPM παρουσιάζει πολύ μικρό ποσοστό επιτυχίας. Ο αλγόριθμος BBP παρουσιάζει εξαιρετικό ποσοστό επιτυχίας, αφού συγκλίνει σε όλες τις προσομοιώσεις. Επιπλέον η απόδοση του σε συναρτησιακούς υπολογισμούς και υπολογισμούς κλίσεως είναι εξαιρετική. Ωστόσο, ο προτεινόμενος αλγόριθμος, ABB, επίσης συγκλίνει σε όλες τις προσομοιώσεις, αλλά παρουσιάζει καλύτερη επίδοση τόσο στους υπολογισμούς της τιμής της συνάρτησης σφάλματος όσο και στους υπολογισμούς κλίσεως. Αξίζει να σημειωθεί ότι ο μέγιστος αριθμός συναρτησιακών υπολογισμών που χρειάστηκαν οι BBP και ABB είναι μικρότερος από 1000 το οποίο σημαίνει ότι για αυτούς τους αλγόριθμους είναι δυνατό να μειώσουμε το όριο των συναρτησιακών υπολογισμών χωρίς να επηρεαστεί το ποσοστό επιτυχίας. Σε αυτή την περίπτωση όμως, κανένας από τους κλασικούς αλγόριθμους δεν θα συνέκλινε, αφού ο ελάχιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι μεγαλύτερος από 1000.

3.4.4 Προσέγγιση Συνεχούς Τριγωνομετρικής Συνάρτησης

Για το πρόβλημα της προσέγγισης μιας συνεχούς τριγωνομετρικής συνάρτησης (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα $1 - 15 - 1$ τεχνητό νευρωνικό δίκτυο (30 βάρη και 16 μεροληψίες). Το κρυφό επίπεδο του νευρωνικού δικτύου είναι βασισμένο σε νευρώνες με λογιστική συνάρτηση ενεργοποίησης, ενώ το επίπεδο εξόδου είναι βασισμένο σε νευρώνες γραμμικής συνάρτησης ενεργοποίησης. Η συνθήκη τερματισμού είναι $E \leq 0.1$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 1000. Για τις μεθόδους σταθερού ρυθμού εκπαίδευσης (BP και BPM) ο ρυθμός εκπαίδευσης ορίστηκε στην προεπιλεγμένη τιμή 0.01. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 3.4.

Αλγόριθμος	Min	Max	μ	σ	Επιτυχία
BP	318	985	710.4	176.3	12.0%
BPM	323	996	688.2	172.6	11.9%
ABP	173	996	627.1	221.5	26.7%
BBP	30	995	238.2	183.7	93.3%
	25	599	149.6	108.6	
ABB	30	988	225.8	183.1	93.9%
	25	612	150.1	215.2	

Πίνακας 3.4: Συγκριτικά αποτελέσματα για το πρόβλημα της προσέγγισης συνεχούς τριγωνομετρικής συνάρτησης.

Σε αυτό το πρόβλημα, ο αλγόριθμος εκπαίδευσης ABP είχε την καλύτερη επίδοση μεταξύ των κλασικών αλγορίθμων εκπαίδευσης σε ποσοστό επιτυχίας και σε μέσο όρο υπολογισμών των τιμών της συνάρτησης σφάλματος. Ο αλγόριθμος BBP απέδωσε πολύ καλύτερα όσον αφορά το ποσοστό επιτυχίας αφού η διαφορά του σε ποσοστό επιτυχίας από τον αλγόριθμο ABP ήταν πάνω από 60%. Επιπλέον, είχε πολύ μικρότερο κόστος σε συναρτησιακούς υπολογισμούς και υπολογισμούς κλίσεως. Ο προτεινόμενος αλγόριθμος εκπαίδευσης ABB παρουσίασε το καλύτερο ποσοστό επιτυχίας στο παράδειγμα αυτό. Επίσης, παρουσίασε μικρότερο κόστος σε συναρτησιακούς υπολογισμούς κατά μέσο όρο από την BBP ενώ ο μέσος όρος των υπολογισμών κλίσεως ήταν περίπου ο ίδιος.

3.5 Συμπεράσματα

Σε αυτό το κεφάλαιο αναπτύχθηκε ένας νέος ρυθμός εκπαίδευσης προσεγγίζοντας μια νέα εξίσωση χορδής, που προτάθηκε από τον Zhang [90], η οποία χρησιμοποιεί πληροφορία παραγώγων και συναρτησιακών τιμών. Στην συνέχεια, αναπτύχθηκε ένας νέος

αλγόριθμος εκπαίδευσης ο οποίος χρησιμοποιεί ως κατεύθυνση αναζήτησης την κατεύθυνση της πιο απότομης καθόδου. Ο αλγόριθμος αυτός χρησιμοποιεί έναν κατάλληλο μηχανισμό επιλογής του ρυθμού εκπαίδευσης ώστε να επιλέγει κάθε φορά τον καλύτερο ρυθμό εκπαίδευσης (μεταξύ του νέου ρυθμού εκπαίδευσης και του ρυθμού εκπαίδευσης των Barzilai και Borwein). Επιπλέον χρησιμοποιείται μια μη μονότονη τεχνική γραμμικής αναζήτησης για την αποδοχή του ρυθμού εκπαίδευσης και αποδείχθηκε η ολική σύγκλιση του αλγορίθμου.

Στη συνέχεια αξιολογήθηκε η απόδοση του νέου αλγορίθμου συγκρίνοντας τον με διάφορους γνωστούς αλγόριθμους της ίδιας τάξης. Για τον σκοπό αυτό χρησιμοποιήθηκαν γνωστά προβλήματα εκπαίδευσης τεχνητών νευρωνικών δικτύων. Τα αποτελέσματα των πειραμάτων έδειξαν ότι ο προτεινόμενος αλγόριθμος υπερέχει τόσο σε ποσοστά επιτυχίας όσο και σε ταχύτητα σύγκλισης. Πιστεύουμε ότι ο συγκεκριμένος αλγόριθμος μπορεί να βελτιωθεί περαιτέρω χρησιμοποιώντας καταλληλότερο μηχανισμό επιλογής του ρυθμού εκπαίδευσης. Επίσης, χρησιμοποιώντας μια πιο αποδοτική τεχνική γραμμικής αναζήτησης είμαστε βέβαιοι ότι η ταχύτητα του αλγορίθμου θα βελτιωθεί περισσότερο.

Νέα Οικογένεια Αλγορίθμων Εκπαίδευσης Συζυγών Κλίσεων

Σε αυτό το κεφάλαιο παρουσιάζονται μερικοί νέοι αποτελεσματικοί αλγόριθμοι εκπαίδευσης οι οποίοι βασίζονται στις μεθόδους βελτιστοποίησης συζυγών κλίσεων. Στους υπάρχοντες αλγόριθμους εκπαίδευσης συζυγών κλίσεων προστίθεται ένας νέος αλγόριθμος εκπαίδευσης που βασίζεται στην μέθοδο συζυγών κλίσεων του Perry [60]. Η μέθοδος του Perry έχει αποδειχθεί ότι είναι μια πολύ αποτελεσματική μέθοδος στο πλαίσιο της βελτιστοποίησης χωρίς περιορισμούς, αλλά δεν είχε εφαρμοστεί ποτέ στην περιοχή της εκπαίδευσης των τεχνητών νευρωνικών δικτύων. Επιπλέον, προτείνεται νέες μέθοδοι συζυγών κλίσεων, που ονομάζονται κλιμακωτές μέθοδοι συζυγών κλίσεων, και προκύπτουν από τις ίδιες αρχές που προέρχονται οι γνωστές μέθοδοι συζυγών κλίσεων των Hestenes-Stiefel, Fletcher-Reeves, Polak-Ribière και Perry. Αυτή η κατηγορία μεθόδων βασίζεται στη φασματική παράμετρο κλιμάκωσης που προτάθηκε από τους Barzilai και Borwein [5]. Η φασματική παράμετρος κλιμάκωσης περιέχει πληροφορία δευτέρας τάξεως χωρίς να απαιτείται ο υπολογισμός του Εσσιανού πίνακα. Επιπλέον, ενσωματώνεται στους αλγόριθμους εκπαίδευσης συζυγών κλίσεων μια αποδοτική τεχνική γραμμικής αναζήτησης που βασίζεται στις συνθήκες του Wolfe και στην διασφαλισμένη κυβική παρεμβολή [77]. Ακόμη, η παράμετρος του αρχικού ρυθμού εκπαίδευσης, που τροφοδοτείται στον αλγόριθμο της γραμμικής αναζήτησης, προσαρμόζεται αυτόματα σε κάθε επανάληψη σύμφωνα με ένα κλειστό τύπο που προτάθηκε στις εργασίες [77] και [80]. Στη συνέχεια, εφαρμόζεται μια αποτελεσματική διαδικασία επανεκκίνησης έτσι ώστε να βελτιωθούν περαιτέρω οι αλγόριθμοι εκπαίδευσης συζυγών κλίσεων και να αποδειχθεί η ολική τους σύγκλιση. Τέλος, παρουσιάζονται τα πειραματικά αποτελέσματα για διάφορα προβλήματα εκπαίδευσης.

4.1 Αλγόριθμοι Εκπαίδευσης Συζυγών Κλίσεων

Οι μέθοδοι συζυγών κλίσεων είναι μια πολύ σημαντική κατηγορία μεθόδων για την ελαχιστοποίηση συνεχών συναρτήσεων, ειδικότερα όταν η διάσταση είναι μεγάλη [57]. Μπορούν να θεωρηθούν ως μέθοδοι συζυγών κατευθύνσεων ή εκτροπής της κλίσης, οι οποίες μπορούν να κατηγοριοποιηθούν μεταξύ της μεθόδου της πιο απότομης καθόδου και της μεθόδου του Newton. Το κύριο πλεονέκτημά τους είναι ότι δεν απαιτούν την αποθήκευση πινάκων, όπως η μέθοδος Newton, και ότι είναι σχεδιασμένες ώστε να συγκλίνουν ταχύτερα από την μέθοδο της πιο απότομης καθόδου.

Οι μέθοδοι συζυγών κλίσεων συγκλίνουν το πολύ σε n επαναλήψεις για τετραγωνικά προβλήματα βελτιστοποίησης χωρίς περιορισμούς στο $\mathbb{R}^n$ όταν χρησιμοποιούνται ακριβείς γραμμικές αναζητήσεις. Ωστόσο, εφαρμόζονται και στην περίπτωση των μη τετραγωνικών προβλημάτων, αφού οι συνεχείς συναρτήσεις παρουσιάζουν τετραγωνική συμπεριφορά στην περιοχή του βέλτιστου. Σε αυτές τις περιπτώσεις, για να βελτιωθεί ο ρυθμός σύγκλισης, ο αλγόριθμος επανεκκινείται κάθε n επαναλήψεις. Πολλές τροποποιήσεις της στρατηγικής των συζυγών κλίσεων έχουν προταθεί στην βιβλιογραφία βασισμένες στην χαλάρωση των τετραγωνικών υποθέσεων ή της ακριβούς γραμμικής αναζήτησης [25], [56]. Επίσης, έχουν προταθεί πολλές συνθήκες επανεκκίνησης του αλγορίθμου των συζυγών κλίσεων [17], [40], [66].

Η μέθοδος συζυγών κλίσεων δημιουργεί μια κατεύθυνση αναζήτησης η οποία είναι αμοιβαία συζυγής με τις προηγούμενες κατευθύνσεις αναζήτησης σε σχέση με ένα δοθέν θετικά ορισμένο πίνακα H , και βρίσκει το βέλτιστο σημείο σε αυτή την κατεύθυνση χρησιμοποιώντας μια τεχνική γραμμικής αναζήτησης. Δύο κατευθύνσεις αναζήτησης d_i και d_j λέγεται ότι είναι αμοιβαία συζυγής ως προς τον H αν ικανοποιείται η ακόλουθη συνθήκη

$$d_i^T H d_j = 0 \quad \text{όπου } i \neq j. \quad (4.1)$$

Με άλλα λόγια, η επόμενη κατεύθυνση αναζήτησης υπολογίζεται ως ένας γραμμικός συνδυασμός της προηγούμενης κατεύθυνσης και της τρέχουσας κλίσης με τέτοιο τρόπο ώστε τα βήματα ελαχιστοποίησης σε όλες τις προηγούμενες κατευθύνσεις να μην παρεμβάλλονται. Η κατεύθυνση αναζήτησης μπορεί να καθοριστεί ως

$$d_k = \begin{cases} -g_k, & \text{για } k = 0 \\ -g_k + \beta_k d_{k-1}, & \text{για } k > 0. \end{cases} \quad (4.2)$$

όπου β_k είναι μια υπό καθορισμό παράμετρος έτσι ώστε η κατεύθυνση d_k να γίνει η k -οστή συζυγής κατεύθυνση. Υπάρχουν πολλοί τρόποι για τον υπολογισμό του β_k . Ο καθένας δημιουργεί μια ξεχωριστή μη γραμμική μέθοδο συζυγών κλίσεων η οποία έχει την δική της ιδιότητα σύγκλισης και αριθμητική απόδοση. Πολλοί τύποι για τον υπολογισμό της παραμέτρου β_k έχουν προταθεί στην βιβλιογραφία. Οι πιο αξιοσημείωτοι

και ευρέως χρησιμοποιούμενοι στην περιοχή της εκπαίδευσης των τεχνητών νευρωνικών δικτύων είναι οι ακόλουθοι:

- Hestenes–Stiefel (HS) [33]

$$\beta_k^{HS} = \frac{g_k^T y_k}{d_{k-1}^T y_k} \quad (4.3)$$

- Fletcher–Reeves (FR) [26]

$$\beta_k^{FR} = \frac{g_k^T g_k}{g_{k-1}^T g_{k-1}} \quad (4.4)$$

- Polak–Ribière (PR) [65]

$$\beta_k^{PR} = \frac{g_k^T y_k}{g_{k-1}^T g_{k-1}} \quad (4.5)$$

όπου $y_k = g_k - g_{k-1}$. Επίσης, δηλώνουμε εδώ για μελλοντική αναφορά ότι $s_k = w_k - w_{k-1}$.

Στην περιοχή της αριθμητικής βελτιστοποίησης χωρίς περιορισμούς, ο Perry στην εργασία [60] πρότεινε μια νέα αποδοτική μέθοδο συζυγών κλίσεων η οποία υπερτερεί σημαντικά από τις προαναφερθείσες κλασικές μεθόδους συζυγών κλίσεων. Η μέθοδος αυτή δεν έχει χρησιμοποιηθεί ποτέ στην περιοχή των τεχνητών νευρωνικών δικτύων. Επομένως, η μέθοδος συζυγών κλίσεων του Perry παρουσιάζεται στη συνέχεια, προσαρμοσμένη στην περιοχή της εκπαίδευσης των τεχνητών νευρωνικών δικτύων.

Ο Perry παρατήρησε στην παραπάνω εργασία ότι χρησιμοποιώντας την επιλογή των Hestenes–Stiefel για την παράμετρο β_k η κατεύθυνση d_k , για $k > 0$, στην εξίσωση (4.2) μπορεί να επαναδιατυπωθεί ως εξής:

$$d_k = -\left[I - \frac{s_k y_k^T}{y_k^T s_k}\right] g_k \equiv -D_k^{HS} g_k \quad (4.6)$$

Σημειώνουμε ότι ο πίνακας D_k^{HS} είναι μια προσέγγιση του αντιστρόφου του Εσσιανού πίνακα, αλλά όχι συμμετρικός. Ως εκ τούτου, δεν αποτελεί μια ψευδό–Newton χωρίς μνήμη ανανέωση του πίνακα [47]. Πιο κατάλληλα, αν ο D_k υποδηλώνεται ως μια προσέγγιση του αντιστρόφου του Εσσιανού πίνακα, η συνθήκη ψευδό–Newton απαιτεί ότι $D_k y_k = s_k$ ή βάσει συμμετρίας ότι

$$y_k^T D_k = s_k^T \quad (4.7)$$

Εκτός από την μη συμμετρία, η εξίσωση (4.7) αποφέρει ότι $y_k^T D_k^{HS} = 0$. Ο Perry [60] σημειώνει ότι κάτω από μη ακριβείς γραμμικές αναζητήσεις, είναι πιο σκόπιμο να επιλέξουμε ότι $d_k = -D_k^P g_k$, όπου η προσέγγιση του αντίστροφου Εσσιανού πίνακα D_k^P επιλέγεται να ικανοποιεί την συνθήκη ψευδό–Newton (4.7) αντί να ικανοποιεί απλά

την H_k -συζυγία (4.1). Έτσι, πρέπει να ισχύει $y_k^T d_k = -y_k^T D_k^P g_k = -s_k^T g_k$. Αντικαθιστώντας αυτό στην εξίσωση (4.2) ο Perry δίνει την δική του επιλογή για το β_k και την αντίστοιχη κατεύθυνση αναζήτησης d_k . Δηλαδή,

$$\beta_k^P = \frac{(y_k - s_k)^T g_k}{y_k^T d_{k-1}} \quad (4.8)$$

$$d_k = -\left[I - \frac{s_k y_k^T}{y_k^T s_k} - \frac{s_k s_k^T}{y_k^T s_k}\right] g_k \equiv -D_k^P g_k \quad (4.9)$$

Σημειώνεται ότι ένας επιπλέον όρος προστέθηκε στον πίνακα D_k^{HS} για να παραχθεί η προσέγγιση D_k^P και η επιλογή του Perry είναι η ίδια με την επιλογή των Hestenes–Stiefel/Polak–Ribière αν χρησιμοποιηθούν ακριβείς γραμμικές αναζητήσεις. Κάτω από αυτές τις συνθήκες, η τροποποίηση τίθεται σε ισχύ όταν $s_k^T g_k \neq 0$.

Συνοψίζοντας τα προηγούμενα, σε αυτή την παράγραφο κατασκευάστηκε μια πλήρης κατηγορία αλγορίθμων εκπαίδευσης συζυγών κλίσεων. Αποτελείται από τους ήδη γνωστούς αλγορίθμους εκπαίδευσης νευρωνικών δικτύων HS, FR και PR και του νέου αλγορίθμου εκπαίδευσης του Perry που προτάθηκε προηγουμένως. Στη συνέχεια γίνεται αναφορά σε αυτή την κατηγορία ως η κλασική κατηγορία αλγορίθμων εκπαίδευσης συζυγών κλίσεων.

4.2 Κλιμακωτοί Αλγόριθμοι Εκπαίδευσης Συζυγών Κλίσεων

Σε αυτή την παράγραφο, λαμβάνοντας υπόψη την εργασία των Birgin και Martinez [9], υποτίθεται ότι μια πιο γενική μορφή της κατεύθυνσης αναζήτησης συζυγών κλίσεων δίνεται από την εξίσωση

$$d_k = \begin{cases} -g_k, & \text{για } k = 0 \\ -\vartheta_k g_k + \beta_k d_{k-1}, & \text{για } k > 0. \end{cases} \quad (4.10)$$

όπου ϑ_k και β_k είναι υπό καθορισμό παράμετροι έτσι ώστε η κατεύθυνση d_k να γίνει η k -οστή συζυγής κατεύθυνση. Σημειώνεται ότι αν $\vartheta_k = 1$ τότε η κατεύθυνση αναζήτησης (4.10) ανάγεται στην κλασική κατεύθυνση συζυγών κλίσεων (4.2).

Ας υποθεθεί τώρα ότι η κατεύθυνση d_k είναι H_k -συζυγής ως προς την κατεύθυνση d_{k-1} , επιβάλλοντας την συνθήκη συζυγίας (4.1), δηλαδή $d_{k-1}^T H_k d_k = 0$, όπου H_k είναι ο Εσσιανός πίνακας της συνάρτησης σφάλματος E στο σημείο w_k . Χρησιμοποιώντας μια τετραγωνική προσέγγιση της συνάρτησης σφάλματος E στο σημείο w_k μαζί με την εξίσωση

$$w_{k+1} = w_k + \alpha_k d_k, \quad k = 0, 1, 2, \dots \quad (4.11)$$

έχουμε ότι $y_k = \alpha_{k-1} H_k d_{k-1}$. Επομένως η συνθήκη συζυγίας απαιτεί ότι

$$d_k^T y_k = 0 \quad (4.12)$$

Αντικαθιστώντας την (4.10) στην (4.12) έχουμε

$$\beta_k = \frac{\vartheta_k g_k^T y_k}{d_{k-1}^T y_k} \quad (4.13)$$

Παρατηρείται ότι στην περίπτωση που το $\vartheta_k = 1$ ο τύπος (4.13) ανάγεται στην επιλογή των Hestenes-Stiefel (4.3).

Αν υποτεθεί ότι εκτελούνται ακριβείς γραμμικές αναζητήσεις, τότε ικανοποιείται η εξίσωση $g_k^T d_{k-1} = 0$. Μετά από μερικούς απλούς υπολογισμούς, ο τύπος (4.13) μετατρέπεται ως

$$\beta_k = \frac{\vartheta_k g_k^T y_k}{\vartheta_{k-1} g_{k-1}^T g_{k-1}} \quad (4.14)$$

ο οποίος είναι η επιλογή των Polak–Ribière (4.5) όταν το $\vartheta_k = \vartheta_{k-1} = 1$.

Επιπλέον, όταν η συνάρτηση σφάλματος E είναι τετραγωνική, όλες οι κατευθύνσεις που παράγονται από την εξίσωση (4.10) χρησιμοποιώντας τον τύπο (4.14) είναι μεταξύ τους συζυγείς, και οι κλίσεις της E σε διαφορετικές επαναλήψεις είναι μεταξύ τους ορθογώνιες. Δηλαδή ισχύει ότι $g_k^T g_{k-1} = 0$. Έτσι, για τετραγωνικές συναρτήσεις και χρησιμοποιώντας ακριβείς γραμμικές αναζητήσεις ο τύπος (4.14) μετατρέπεται στον τύπο

$$\beta_k = \frac{\vartheta_k g_k^T g_k}{\vartheta_{k-1} g_{k-1}^T g_{k-1}} \quad (4.15)$$

ο οποίος είναι η επιλογή των Fletcher–Reeves (4.4) όταν το $\vartheta_k = \vartheta_{k-1} = 1$.

Αν επιβληθεί ότι η γενική μορφή των κατευθύνσεων αναζήτησης συζυγών κλίσεων που δίνεται από την εξίσωση (4.10), χρησιμοποιώντας την γενίκευση του τύπου HS για το β_k (4.13), ικανοποιεί την συμμετρική ψευδό-Newton εξίσωση (4.7) παρά την συνθήκη H_k -συζυγίας (4.1), τότε καθορίζεται η γενικευμένη επιλογή του Perry η οποία δίνεται από την εξίσωση

$$\beta_k = \frac{(\vartheta_k y_k - s_k)^T g_k}{y_k^T d_{k-1}} \quad (4.16)$$

Σημειώνεται ότι για $\vartheta_k = 1$ ο τύπος (4.16) ανάγεται στην κλασική επιλογή του Perry για το β_k που δίνεται από τον τύπο (4.8).

4.2.1 Παράμετρος Κλιμάκωσης

Στο σημείο αυτό, ο στόχος είναι να βρεθεί μια κατάλληλη επιλογή για την παράμετρο κλιμάκωσης ϑ_k έτσι ώστε να βελτιωθεί η απόδοση και η ταχύτητα σύγκλισης των κλασικών αλγορίθμων εκπαίδευσης συζυγών κλίσεων. Μια καλή και αποδοτική επιλογή για

την παράμετρο κλιμάκωσης είναι το φασματικό βήμα που προτάθηκε από τους Barzilai και Borwein [5], δηλαδή

$$\vartheta_k = \frac{s_k^T s_k}{s_k^T y_k} \quad (4.17)$$

Οι Barzilai και Borwein πρότειναν αυτό το βήμα έτσι ώστε να σταθεροποιήσουν και να επιταχύνουν την μέθοδο της πιο απότομης καθόδου. Τα πειραματικά αποτελέσματα έδειξαν ότι το φασματικό βήμα βελτιώνει σημαντικά την απόδοση της μεθόδου της πιο απότομης καθόδου. Αργότερα, ο Raydan στην εργασία [69] συνδύασε το φασματικό βήμα με μια μη μονότονη στρατηγική γραμμικής αναζήτησης [29] και παρήγαγε μια εύρωστη μέθοδο ελαχιστοποίησης που μπορεί να συγκριθεί με τις κλασικές μεθόδους συζυγών κλίσεων. Στις εργασίες [63] και [64], η μέθοδος του Raydan μεταφέρθηκε επιτυχώς στην περιοχή των τεχνητών νευρωνικών δικτύων με πολύ καλά αποτελέσματα.

Η επιτυχία του φασματικού βήματος έρχεται από το γεγονός ότι το ϑ_k είναι προσεγγίσεις του πηλίκου Rayleigh για κάποια τιμή $1/\lambda_i$. Έχοντας στο μυαλό την εξίσωση χορδής, $B_k s_k = y_k$, μπορούμε να επαληθευτεί ότι το πηλίκο Rayleigh που δίνεται από την εξίσωση

$$R_q = s_k^T \left[\int_0^1 \nabla^2 E(w_k + \tau s_k) d\tau \right] s_k / s_k^T s_k$$

αντιστοιχεί στον μέσο όρο του Εσσιανού πίνακα

$$\int_0^1 \nabla^2 E(w_k + t s_k) dt$$

στο ευθύγραμμο τμήμα μεταξύ του w_k και w_{k+1} . Επομένως το R_q είναι μια καλή προσέγγιση της ιδιοτιμής λ_i για την οποία το s_k είναι το αντίστοιχο ιδιοδιάνυσμα. Επομένως είναι προφανές ότι το βήμα (5.9) σχετίζεται με τις ιδιοτιμές του Εσσιανού πίνακα στην τιμή του ελαχίστου και όχι με την συναρτησιακή τιμή. Αυτό σημαίνει ότι το φασματικό βήμα περιέχει πληροφορία δευτέρας τάξεως χωρίς να απαιτείται ο υπολογισμός του Εσσιανού πίνακα.

Η χρήση του φασματικού βήματος ϑ_k (5.9) ως παράμετρος κλιμάκωσης στην γενικευμένη κατεύθυνση αναζήτησης συζυγών κλίσεων (4.10) και στους γενικευμένους τύπους για το β_k , δηλαδή για τους τύπους (4.13), (4.15), (4.14) και (4.16), οδηγεί σε μια νέα κατηγορία αλγόριθμων εκπαίδευσης συζυγών κλίσεων. Το όνομα της κατηγορίας αυτής είναι κλιμακωτοί αλγόριθμοι εκπαίδευσης συζυγών κλίσεων Self-Scaled Conjugate Gradient (S-SCG).

4.3 Γενικός Αλγόριθμος Εκπαίδευσης

Σε αυτή την παράγραφο δίνεται ένας γενικός αλγόριθμος εκπαίδευσης τεχνητών νευρωνικών δικτύων που βασίζεται στις μεθόδους συζυγών κλίσεων. Πριν παρουσιαστεί η

διαδικασία εκπαίδευσης, γίνεται μια λεπτομερής παρουσίαση όσον αφορά τις συνθήκες επανεκκίνησης, τις τεχνικές γραμμικής αναζήτησης και τον αρχικό ρυθμό εκπαίδευσης.

4.3.1 Διαδικασίες Επανεκκίνησης

Μια τροποποίηση που χρησιμοποιείται συχνά στους μη γραμμικούς αλγόριθμους συζυγών κλίσεων είναι η επανεκκίνηση του αλγόριθμου κάθε n επαναλήψεις θέτοντας την παράμετρο $\beta_k = 0$ στην εξίσωση (4.10), δηλαδή εφαρμόζοντας μια επανάληψη χρησιμοποιώντας την κατεύθυνση της πιο απότομης καθόδου. Η επανεκκίνηση υπηρετεί το σκοπό της περιοδικής ανανέωσης του αλγόριθμου, διαγράφοντας παλιά πληροφορία που δεν είναι πλέον χρήσιμη. Έχει αποδειχθεί ένα ισχυρό θεωρητικό αποτέλεσμα όσον αφορά την επανεκκίνηση. Πιο συγκεκριμένα, οδηγεί σε τετραγωνική σύγκλιση n -βημάτων, δηλαδή

$$\|w_{k+n} - w\| = O(\|w_k - w^*\|^2) \quad (4.18)$$

Το αποτέλεσμα αυτό δεν προκαλεί έκπληξη. Ας θεωρήσουμε ότι η συνάρτηση E είναι κυρτή τετραγωνική στην περιοχή της υπό αναζήτηση λύσης, αλλά μη τετραγωνική στο υπόλοιπο πεδίο ορισμού της. Υποθέτοντας ότι ο αλγόριθμος συγκλίνει στην υπό αναζήτηση λύση, οι επαναλήψεις θα οδηγηθούν τελικά στην τετραγωνική περιοχή. Κάποια στιγμή ο αλγόριθμος θα επανεκκινηθεί σε αυτή την περιοχή και από το σημείο αυτό η συμπεριφορά του θα είναι ίδια με αυτή της γραμμικής μεθόδου συζυγών κλίσεων. Δηλαδή, ο αλγόριθμος θα τερματίσει σε πεπερασμένο αριθμό επαναλήψεων μέσα στις επόμενες n επαναλήψεις από την επανεκκίνηση. Η επανεκκίνηση είναι σημαντική γιατί η ιδιότητα του πεπερασμένου τερματισμού της γραμμικής μεθόδου συζυγών κλίσεων ισχύει μόνο όταν η αρχική κατεύθυνση αναζήτησης d_0 είναι ίση με την αρνητική κλίση.

Ακόμη και αν η συνάρτηση E δεν είναι ακριβώς τετραγωνική στην περιοχή της λύσης, το θεώρημα του Taylor υποδηλώνει ότι μπορεί να προσεγγιστεί από μια τετραγωνική συνάρτηση, δεδομένου ότι είναι ομαλή. Επομένως, ενώ δεν αναμένεται πλέον τερματισμός μέσα σε n επαναλήψεις από την επανεκκίνηση, δεν προκαλεί έκπληξη ότι γίνεται σημαντική πρόοδος για την εύρεση της λύσης, όπως φαίνεται από την σχέση (4.18).

Ωστόσο η τεχνική της επανεκκίνησης λαμβάνοντας υπόψιν τον αριθμό των επαναλήψεων ενδέχεται να μην είναι επαρκής στην πράξη, αφού οι μη γραμμικοί αλγόριθμοι συζυγών κλίσεων χρησιμοποιούνται κυρίως για την επίλυση προβλημάτων μεγάλης διάστασης n . Σε τέτοιου είδους προβλήματα μπορεί ο αλγόριθμος να μην επανεκκινηθεί καθόλου αφού συνήθως βρίσκεται μια προσεγγιστική λύση σε λιγότερες από n επαναλήψεις. Έτσι οι μη γραμμικές μέθοδοι συζυγών κλίσεων υλοποιούνται με στρατηγικές επανεκκίνησης που χρησιμοποιούν άλλους παράγοντες.

Η πιο ευρέως χρησιμοποιούμενη διαδικασία επανεκκίνησης προτάθηκε από τον Powell [66]. Η διαδικασία επανεκκίνησης του Powell ελέγχει εάν υπάρχει πολύ μικρή

ορθογωνιότητα μεταξύ της τρέχουσας και της προηγούμενης κλίσης. Ειδικότερα, αν η συνθήκη

$$|g_k^T g_{k-1}| \geq 0.2 \|g_k\|^2$$

ικανοποιείται, τότε η κατεύθυνση αναζήτησης συζυγών κλίσεων επανεκκινείται με την κατεύθυνση της πιο απότομης καθόδου $-g_k$.

Ο Shanno, στην εργασία [76], πρότεινε μια διαδικασία ελέγχου γωνίας για να προσδιορίσει πότε θα πρέπει να επανεκκινείται η μέθοδος συζυγών κλίσεων με την κατεύθυνση της πιο απότομης καθόδου. Ο έλεγχος βασίζεται στην εγγύηση ότι το συνημίτονο της γωνίας μεταξύ της κατεύθυνσης αναζήτησης συζυγών κλίσεων και της αρνητικής κλίσης είναι εντός ενός σταθερού πολλαπλασίου του συνημιτόνου μεταξύ της κατεύθυνσης αναζήτησης συζυγών κλίσεων των FR και της αρνητικής κλίσης.

Μια άλλη διαδικασία επανεκκίνησης, πιο απλή από την προηγούμενη που προσδιορίστηκε από τον Shanno, προτάθηκε από τους Birgin και Martinez στην εργασία [9]. Ο αλγόριθμος συζυγών κλίσεων επανεκκινείται όταν η γωνία μεταξύ της τρέχουσας κατεύθυνσης αναζήτησης συζυγών κλίσεων d_k και της κλίσης g_k δεν είναι αρκετά οξεία. Δηλαδή, αν η συνθήκη

$$d_k^T g_k \leq -10^{-3} \|d_k\| \|g_k\| \quad (4.19)$$

ικανοποιείται, τότε ο αλγόριθμος επανεκκινείται χρησιμοποιώντας ως κατεύθυνση αναζήτησης την φασματική κατεύθυνση αναζήτησης $-\theta_k g_k$.

Πιο εξελιγμένες διαδικασίες επανεκκίνησης έχουν προταθεί στη βιβλιογραφία όπου η κατεύθυνση επανεκκίνησης είναι διαφορετική από την κατεύθυνση της πιο απότομης καθόδου [7], [22], [54] και [66]. Σε αυτές τις διαδικασίες επανεκκίνησης, η κατεύθυνση επανεκκίνησης d_k ορίζεται χρησιμοποιώντας μια επανάληψη τριών όρων βασισμένων στην κλίση g_k , την προηγούμενη κατεύθυνση συζυγών κλίσεων d_{k-1} και μια τρίτη κατεύθυνση που περιέχει προηγούμενη πληροφορία για την συμπεριφορά της συνάρτησης σφάλματος. Ωστόσο, εδώ χρησιμοποιείται η διαδικασία επανεκκίνησης που προτάθηκε από τους Birgin και Martinez στην εργασία [9], η οποία είναι μια απλή διαδικασία, έτσι ώστε να δειχθεί η ισχύς των προτεινόμενων αλγορίθμων συζυγών κλίσεων χωρίς να ενισχύονται με εξελιγμένες τεχνικές επανεκκίνησης.

4.3.2 Διαδικασίες Γραμμικής Αναζήτησης

Οι αλγόριθμοι εκπαίδευσης συζυγών κλίσεων για να έχουν την ιδιότητα της ολικής σύγκλισης, ο ρυθμός εκπαίδευσης πρέπει να καθορίζεται από μια μονοδιάστατη γραμμική αναζήτηση κατά μήκος της κατεύθυνσης αναζήτησης συζυγών κλίσεων d_k [56]. Πολλές τεχνικές γραμμικής αναζήτησης έχουν προταθεί στην βιβλιογραφία για χρήση με τους αλγόριθμους εκπαίδευσης συζυγών κλίσεων.

Ενίοτε, τεχνικές γραμμικής αναζήτησης χωρίς την χρήση παραγώγων έχουν χρησιμοποιηθεί για τον καθορισμό της παραμέτρου του ρυθμού εκπαίδευσης. Η πιο δημοφιλής τεχνική από αυτού του είδους γραμμικής αναζήτησης είναι ο αλγόριθμος της χρυσής τομής και η αναζήτηση Fibonacci [47]. Η διαφορά σε αυτές τις δύο τεχνικές εντοπίζεται στον τρόπο που παράγονται οι δοκιμαστικές παράμετροι του ρυθμού εκπαίδευσης. Ο Brent στην εργασία [12] πρότεινε μια τεχνική γραμμικής αναζήτησης η οποία είναι συνδυασμός της αναζήτησης της χρυσής τομής και μια τετραγωνικής προσέγγισης. Η αναζήτηση του Brent επιχειρεί να συνδυάσει τις καλύτερες ιδιότητες και από τις δύο προσεγγίσεις και είναι ένας αλγόριθμος που δεν απαιτεί την χρήση παραγώγων.

Ωστόσο, τεχνικές γραμμικής αναζήτησης που κάνουν χρήση της πληροφορίας πρώτης τάξεως, δηλαδή της κλίσης της συνάρτησης σφάλματος, χρησιμοποιούνται πολύ συχνά με τους αλγόριθμους εκπαίδευσης συζυγών κλίσεων. Πιο συγκεκριμένα, ο ρυθμός εκπαίδευσης που επιτυγχάνεται από μια τεχνική γραμμικής αναζήτησης πρέπει να ικανοποιεί της συνθήκες του Wolfe που δίνονται από τις παρακάτω σχέσεις

$$E(w_k + \alpha_k d_k) - E(w_k) \leq c_1 \alpha_k \nabla E(w_k)^T d_k \quad (4.20)$$

$$\nabla E(w_k + \alpha_k d_k)^T d_k \geq c_2 \nabla E(w_k)^T d_k \quad (4.21)$$

όπου $0 < c_1 \leq c_2 < 1$. Η πρώτη συνθήκη, που δίνεται από την σχέση (4.20), επιτρέπει οποιοδήποτε σημείο να γίνει αποδεκτό αν βελτιώνει ικανοποιητικά την συναρτησιακή τιμή αναφορικά με την προηγούμενη συναρτησιακή τιμή. Η δεύτερη συνθήκη, που δίνεται από την σχέση (4.21) δεν επιτρέπει στον ρυθμό εκπαίδευσης να γίνει πολύ μικρός και εξασφαλίζει ότι ο παρονομαστής της φασματικής παραμέτρου (5.9) είναι καλά ορισμένος και πάντοτε θετικός, αφού υποδηλώνει ότι $s_k^T y_k > 0$.

Πολλοί αλγόριθμοι γραμμικής αναζήτησης που χρησιμοποιούν της συνθήκες του Wolfe, (4.20) και (4.21), έχουν προταθεί στην βιβλιογραφία. Οι Nocedal και Wright [57] περιέγραψαν μια τεχνική γραμμικής αναζήτησης που βασίζεται στην προσέγγιση της γνωστής συνάρτησης σφάλματος και των τιμών της κλίσης της. Στο βιβλίο [75] περιγράφεται μια υβριδική τεχνική γραμμικής αναζήτησης που βασίζεται στην διχοτόμηση και την κυβική παρεμβολή. Ο Χαραλάμπους, στην εργασία [16], ανέπτυξε μια τεχνική γραμμικής αναζήτησης για χρήση με τους αλγόριθμους εκπαίδευσης συζυγών κλίσεων. Όπως η προηγούμενη μέθοδος, εφαρμόζει μια υβριδική αναζήτηση. Χρησιμοποιεί μια κυβική παρεμβολή μαζί με μια μορφή τμηματοποίησης. Οι Shanno και Phua, στην εργασία [77], πρότειναν μια αποδοτική τεχνική γραμμικής αναζήτησης για χρήση μαζί με τον αλγόριθμο CONMIN. Αυτός ο αλγόριθμος γραμμικής αναζήτησης βασίζεται σε μια τεχνική εγκλωβισμού που περιγράφεται στην εργασία [78] και μια διασφαλισμένη κυβική παρεμβολή όμοια με αυτή που χρησιμοποιείται στην εργασία [18].

Όλες οι παραπάνω τεχνικές γραμμικής αναζήτησης έχουν χρησιμοποιηθεί μαζί με τους αλγόριθμους εκπαίδευσης συζυγών κλίσεων από πολλούς ερευνητές για την επίλυση διαφόρων προβλημάτων. Ωστόσο δεν είναι σαφές ποια απ' όλες αποδίδει καλύτερα γενικά. Δηλαδή, διαφορετικοί συνδυασμοί των αλγορίθμων εκπαίδευσης συζυγών κλίσεων με τεχνικές γραμμικής αναζήτησης είναι κατάλληλοι για κάθε πρόβλημα. Εδώ

επιλέχθηκε να χρησιμοποιηθεί η τεχνική γραμμικής αναζήτησης που χρησιμοποιείται στον αλγόριθμο CONMIN.

4.3.3 Αρχικός Ρυθμός Εκπαίδευσης

Η επιλογή της παραμέτρου του αρχικού ρυθμού εκπαίδευσης για τους αλγόριθμους εκπαίδευσης συζυγών κλίσεων είναι πολύ σημαντική για την επιτυχημένη και γρήγορη σύγκλισή τους. Μια δημοφιλής στρατηγική είναι να υποθέσουμε ότι η πρώτης τάξεως αλλαγή στην συνάρτηση στο σημείο w_k θα είναι η ίδια με του προηγούμενου βήματος [57]. Με άλλα λόγια, η αρχική εκτίμηση του ρυθμού εκπαίδευσης α_k επιλέγεται να είναι τέτοια ώστε $\alpha_k g_k^T d_k = \alpha_{k-1} g_{k-1}^T d_{k-1}$. Δηλαδή η αρχική παράμετρος του ρυθμού εκπαίδευσης δίνεται από το τύπο

$$\alpha_k = \alpha_{k-1} \frac{g_{k-1}^T d_{k-1}}{g_k^T d_k}$$

Μια άλλη επιλογή για τον αρχικό ρυθμό εκπαίδευσης προτάθηκε στο βιβλίο [25]. Εάν είναι διαθέσιμη μια εκτίμηση $\Delta E = E(w_{k-1}) - E(w_k)$, τότε είναι δυνατό να παρεμβάλλουμε μια τετραγωνική συνάρτηση στο $g_k^T d_k$ και ΔE , η οποία δίνει την ακόλουθη εκτίμηση

$$\alpha_k = \frac{-2\Delta E}{g_k^T d_k}$$

Στην πρώτη επανάληψη του αλγορίθμου εκπαίδευσης, το ΔE πρέπει να οριστεί από τον χρήστη. Στη συνέχεια, η επιλογή του $\Delta E = \max(E(w_{k-1}) - E(w_k), \varepsilon)$, όπου ε είναι ένας μικρός θετικός αριθμός, λειτουργεί αρκετά καλά, που σημαίνει ότι η μείωση από την προηγούμενη επανάληψη διασφαλίζεται.

Μια δημοφιλής επιλογή του αρχικού ρυθμού εκπαίδευσης που αποδεδειγμένα συνεργάζεται επιτυχώς με της μεθόδους συζυγών κλίσεων έχει προταθεί από τον Shanno στον αλγόριθμο CONMIN [77]. Ο αρχικός ρυθμός εκπαίδευσης του Shanno δίνεται από τον τύπο

$$\alpha_k = \begin{cases} \frac{1}{\|g_0\|_2}, & \text{αν } k = 0; \\ \frac{\alpha_{k-1} \|d_k\|}{\|d_{k-1}\|}, & \text{διαφορετικά.} \end{cases} \quad (4.22)$$

Εδώ χρησιμοποιείται ο αρχικός ρυθμός εκπαίδευσης του Shanno.

4.3.4 Περιγραφή του Αλγορίθμου Εκπαίδευσης

Σε αυτό το σημείο δίνεται μια λεπτομερής διατύπωση του κλιμακωτού αλγορίθμου εκπαίδευσης συζυγών κλίσεων. Σημειώνεται ότι για $\nu_k = 1$, ο αλγόριθμος ανάγεται στον

κλασικό αλγόριθμο εκπαίδευσης συζυγών κλίσεων. Ο αλγόριθμος εκπαίδευσης δέχεται ως είσοδο το διάνυσμα των αρχικών βαρών w_0 , το κριτήριο τερματισμού ϵ , τον μέγιστο αριθμό επαναλήψεων μ , τις παραμέτρους c_1 και c_2 που χρειάζονται για τον αλγόριθμο γραμμικής αναζήτησης και επιστρέφει το σημείο ελαχίστου w_* και το ελάχιστο της συνάρτησης σφάλματος $E(w_*)$.

Αλγόριθμος 5 Ο κλιμακωτός αλγόριθμος εκπαίδευσης συζυγών κλίσεων.

```

SCG( $w_0, \epsilon, \mu, c_1, c_2, w_*, E(w_*)$ )
1: Για  $k := 0$  έως  $\mu$  Κάνε
2:    $E_k = E(w_k), g_k = \nabla E(w_k)$ .
3:   Αν  $E_k \leq \epsilon$  Τότε
4:     Επέστρεψε  $w_* = w_k, E(w_*) = E(w_k)$ .
5:   Τέλος
6:   Αν  $k := 0$  Τότε
7:      $d_k = -g_k$ .
8:      $\alpha_k = 1/\|g_k\|$ .
9:   Αλλιώς
10:     $\vartheta_k = \frac{s_k^T s_k}{s_k^T y_k}$ .
11:    Υπολόγισε το  $\beta_k$  επιλέγοντας έναν από τους τύπους (4.13), (4.14), (4.15) και (4.16).
12:     $d = -\vartheta_k g_k + \beta_k d_{k-1}$ .
13:    Αν  $d^T g_k \leq -10^{-3} \|d_k\| \|g_k\|$  Τότε
14:       $d_k = d$ .
15:    Αλλιώς
16:       $d_k = -\vartheta_k g_k$ .
17:    Τέλος
18:     $\alpha_k = \frac{\alpha_{k-1} \|d_k\|}{\|d_{k-1}\|}$ .
19:  Τέλος
20:  LineSearch. Υπολόγισε ένα ρυθμό εκπαίδευσης  $\alpha_k$  που να ικανοποιεί τις συνθήκες του Wolfe (4.20) και (4.21).
21:   $w_{k+1} = w_k + \alpha_k d_k$ .
22: Τέλος
23: Επέστρεψε Το δίκτυο δεν εκπαιδεύτηκε.

```

Αρχικά, ο αλγόριθμος υπολογίζει την τιμή της συνάρτησης σφάλματος $E(w_k)$ και την κλίση της $g_k = -\nabla E(w_k)$ για το τρέχον σημείο w_k . Στα βήματα 3-5, ελέγχεται το κριτήριο τερματισμού. Εάν ικανοποιείται, τότε ο αλγόριθμος τερματίζει επιστρέφοντας το σημείο ελαχίστου w_* και το ελάχιστο της συνάρτησης σφάλματος $E(w_*)$, διαφορετικά συνεχίζει στο επόμενο βήμα. Το βήμα 6 ελέγχει αν ο αλγόριθμος βρίσκεται στην πρώτη επανάληψη του. Εάν αληθεύει, τότε λαμβάνεται ως κατεύθυνση αναζήτησης η αρνητική κλίση, δηλαδή $d_k = -g_k$, ο αρχικός ρυθμός εκπαίδευσης υπολογίζεται από τον τύπο $\alpha_k = 1/\|g_k\|$ και ο αλγόριθμος συνεχίζει στο βήμα 20. Διαφορετικά, υπο-

λογίζεται η παράμετρος κλιμάκωσης από τον τύπο (5.9) στο βήμα 10. Στο βήμα 11 υπολογίζεται η παράμετρος β_k επιλέγοντας έναν από τους τύπους (4.13), (4.14), (4.15) ή (4.16). Στη συνέχεια, στο βήμα 12, υπολογίζεται η γενική κατεύθυνση αναζήτησης συζυγών κλίσεων χρησιμοποιώντας τον τύπο (4.10). Στο βήμα 13, ελέγχεται το κριτήριο επανεκκίνησης για την κατεύθυνση αναζήτησης d_k . Αν η συνθήκη ικανοποιείται, τότε η γενική κατεύθυνση αναζήτησης συζυγών κλίσεων γίνεται αποδεκτή, διαφορετικά, η κατεύθυνση αναζήτησης επανεκκινείται χρησιμοποιώντας την φασματική κατεύθυνση αναζήτησης $d_k = -\theta_k g_k$. Η αρχική παράμετρος του ρυθμού εκπαίδευσης υπολογίζεται στο βήμα 18. Στο βήμα 20, εφαρμόζεται μια τεχνική γραμμικής αναζήτησης για τον ρυθμό εκπαίδευσης α_k . Στο βήμα 21, υπολογίζεται το νέο διάνυσμα των βαρών και ο αλγόριθμος επαναλαμβάνεται μέχρι να ικανοποιηθεί το κριτήριο τερματισμού ή να επιτευχθεί ο μέγιστος αριθμός επαναλήψεων.

4.3.5 Ολική Σύγκλιση

Σε αυτή την παράγραφο αποδεικνύεται η ολική σύγκλιση του Αλγορίθμου 6. Για τον σκοπό αυτό θα χρησιμοποιηθεί το Θεώρημα του Zoutendijk.

Θεώρημα 4.1. Θεωρούμε το επαναληπτικό σχήμα της μορφής (4.11), όπου d_k είναι μια καθοδική κατεύθυνση και το α_k ικανοποιεί τις συνθήκες του Wolfe (4.20) και (4.21). Υποθέτουμε ότι η συνάρτηση σφάλματος E είναι κάτω φραγμένη στο $\mathbb{R}^n$ και ότι η E είναι συνεχώς διαφορίσιμη σε μια γειτονιά $\mathcal{N}$ του συνόλου $\mathcal{L} = \{w \in \mathbb{R}^n : E(w) \leq E(w_0)\}$, όπου w_0 είναι το αρχικό σημείο του επαναληπτικού σχήματος. Υποθέτουμε επίσης ότι η κλίση ∇E είναι Lipschitz συνεχής στο $\mathcal{N}$, δηλαδή, ότι υπάρχει μια σταθερά $L > 0$ τέτοια ώστε

$$\|\nabla E(w) - \nabla E(\tilde{w})\| \leq L\|w - \tilde{w}\|, \quad \text{για όλα τα } w, \tilde{w} \in \mathcal{N}. \quad (4.23)$$

Τότε έχουμε

$$\sum_{k \geq 0} \cos^2 \theta_k \|\nabla E_k\|^2 < \infty \quad (4.24)$$

όπου

$$\cos \theta_k = \frac{-\nabla E_k^T d_k}{\|\nabla E_k\| \|d_k\|} \quad (4.25)$$

Απόδειξη. Από τις σχέσεις (4.21) και (4.11) έχουμε ότι

$$(\nabla E_{k+1} - \nabla E_k)^T d_k \geq (c_2 - 1) \nabla E_k^T d_k$$

ενώ η συνθήκη Lipschitz υποδηλώνει ότι

$$(\nabla E_{k+1} - \nabla E_k)^T d_k \leq \alpha_k L \|d_k\|^2$$

συνδυάζοντας τις δύο αυτές σχέσεις έχουμε ότι

$$\alpha_k \geq \frac{c_2 - 1}{L} \frac{\nabla E_k^T d_k}{\|d_k\|^2}$$

Αντικαθιστώντας αυτή την ανισότητα στην πρώτη συνθήκη του Wolfe (4.20), έχουμε ότι

$$E_{k+1} \leq E_k - c_1 \frac{1 - c_2}{L} \frac{(\nabla E_k^T d_k)^2}{\|d_k\|^2}.$$

Χρησιμοποιώντας την σχέση (4.25), μπορούμε να επαναδιατυπώσουμε την προηγούμενη σχέση ως

$$E_{k+1} \leq E_k - c \cos^2 \theta_k \|\nabla E_k\|^2.$$

όπου $c = c_1(1 - c_2)/L$. Αθροίζοντας την παραπάνω έκφραση για όλους τους δείκτες που είναι μικρότεροι ή ίση του k , έχουμε ότι

$$E_{k+1} \leq E_0 - c \sum_{j=0}^k \cos^2 \theta_j \|\nabla E_j\|^2. \quad (4.26)$$

Εφόσον η E είναι κάτω φραγμένη, έχουμε ότι η διαφορά $E_{k+1} - E_0$ είναι μικρότερη από μια θετική σταθερά για όλα τα k . Έτσι παίρνοντας το όριο για την σχέση (4.26) έχουμε ότι

$$\sum_{k \geq 0} \cos^2 \theta_k \|\nabla E_k\|^2 < \infty$$

το οποίο ολοκληρώνει την απόδειξη. $\square$

Παρατήρηση 4.1. Το παραπάνω θεώρημα μπορεί να εφαρμοστεί στον Αλγόριθμο 6 αφού η συνάρτηση σφάλματος E ικανοποιεί τις απαιτήσεις του θεωρήματος. Πιο συγκεκριμένα, η συνάρτηση σφάλματος E είναι κάτω φραγμένη ($E(w) > 0$) ως άθροισμα τετραγώνων. Επιπλέον, για τα τεχνητά νευρωνικά δίκτυα που χρησιμοποιούν συνεχείς συναρτήσεις ενεργοποίησης με συνεχείς παραγώγους (υπερβολική εφαπτομένη, λογιστική συνάρτηση), όπως στην περίπτωση που μελετάμε, η συνάρτηση σφάλματος E είναι επίσης συνεχώς διαφορίσιμη.

Χρησιμοποιώντας το αποτέλεσμα του Θεωρήματος του Zoutendijk μπορεί να αποδείχθει την ολική σύγκλιση για όλους τους αλγόριθμους οι οποίοι επανεκκινούνται περιοδικά θέτοντας την παράμετρο $\beta_k = 0$. Αν $k_1, k_2, \dots$ υποδηλώνουν τις επαναλήψεις στις οποίες λαμβάνουν χώρα οι επανεκκινήσεις, τότε από την συνθήκη του Zoutendijk (4.24) έχουμε ότι

$$\sum_{k=k_1, k_2, \dots} \|\nabla E_k\|^2 < \infty \quad (4.27)$$

Εάν δεν επιτραπούν περισσότερες από $\bar{n}$ επαναλήψεις μεταξύ δύο διαδοχικών επανεκκινήσεων, η ακολουθία $\{k_j\}_{j=1}^\infty$ θα είναι μη πεπερασμένη, και από την σχέση (4.27)

λαμβάνεται ότι το $\lim_{j \rightarrow \infty} \|\nabla E_{k_j}\| = 0$. Δηλαδή, μια υπακολουθία των κλίσεων της συνάρτησης σφάλματος E συγκλίνει στο 0, ή ισοδύναμα,

$$\liminf_{k \rightarrow \infty} \|\nabla E_k\| = 0 \quad (4.28)$$

Το αποτέλεσμα αυτό ισχύει για όλους τους αλγορίθμους συζυγών κλίσεων που ενσωματώνουν τεχνικές επανεκκίνησης και επομένως για τον Αλγόριθμο 6.

4.4 Πειραματικά Αποτελέσματα

Σε αυτή την παράγραφο, αξιολογείται η απόδοση των προτεινόμενων αλγορίθμων εκπαίδευσης και συγκρίνονται με τους κλασικούς αλγόριθμους εκπαίδευσης συζυγών κλίσεων των Hestenes-Stiefel (HS) [33], Fletcher-Reeves (FR) [26] και Polak-Ribière (PR) [65]. Επίσης, συμπεριλαμβάνονται στα πειράματα οι αλγόριθμοι εκπαίδευσης OSS [6], BFGS περιορισμένης μνήμης με $m = 1$ (LN) [2, 46], GB1 και GB2 [21]. Όλοι οι προαναφερθέντες αλγόριθμοι εκπαίδευσης απαιτούν μια στρατηγική γραμμικής αναζήτησης για να έχουν την ιδιότητα της ολικής σύγκλισης. Για να είναι δίκαιη η σύγκριση, οι αλγόριθμοι εκπαίδευσης OSS, LN, GB1 και GB2 εφοδιάστηκαν με την ίδια στρατηγική γραμμική αναζήτησης και κριτήριο επανεκκίνησης που χρησιμοποιήθηκε στους αλγόριθμους εκπαίδευσης συζυγών κλίσεων (κλασικών και κλιμακωτών). Οι αλγόριθμοι υλοποιήθηκαν στο υπολογιστικό πρόγραμμα Matlab χρησιμοποιώντας το παρεχόμενο πακέτο για τα νευρωνικά δίκτυα. Για τους αλγόριθμους εκπαίδευσης συζυγών κλίσεων (κλασικούς και κλιμακωτούς) και για της ψευδό-Newton μεθόδους, οι τιμές των παραμέτρων c_1 και c_2 έχουν καθοριστεί στις 10^{-4} και 0.5, αντίστοιχα, για όλα τα πειράματα έτσι ώστε να επαληθευθεί ότι οι προτεινόμενοι αλγόριθμοι αποδίδουν αρκετά καλά σε διάφορα προβλήματα χωρίς να χρειάζεται ρύθμιση αυτών των παραμέτρων. Επιπλέον, για την αρχικοποίηση των βαρών και των μεροληψιών εφαρμόστηκε η τεχνική που προτάθηκε από τους Nguyen και Widrow [55] για όλους τους αλγόριθμους (εκτός κι αν αναφερθεί διαφορετικά για κάποιο πρόβλημα).

Η επιτυχία και η ταχύτητα σύγκλισης των προτεινόμενων αλγορίθμων εκπαίδευσης αξιολογείται αναφορικά με τους προαναφερθέντες αλγόριθμους εκπαίδευσης χρησιμοποιώντας τα ακόλουθα πολύ γνωστά προβλήματα:

1. Το πρόβλημα του Αποκλειστικού-EITE.
2. Το πρόβλημα της ισοτιμίας των 3-bit.
3. Το πρόβλημα του $N - M - N$ κωδικοποιητή/αποκωδικοποιητή (όπου $M = \log_2 N$, $N = 4, 8, 16, 32, 64, 128, 256$).
4. Το πρόβλημα της αναγνώρισης των κεφαλαίων γραμμάτων της Αγγλικής αλφαβήτου.

5. Το πρόβλημα της αναγνώρισης αριθμών.

Επιπλέον αξιολογείται η απόδοση των αλγορίθμων εκπαίδευσης χρησιμοποιώντας μερικά πραγματικά προβλήματα από την αποθήκη δεδομένων μηχανικής μάθησης του UCI [53]. Τα προβλήματα αυτά είναι η ταξινόμηση των φυτών της οικογένειας Iris, η οπτική αναγνώριση χειρόγραφων αριθμών και η αναγνώριση φωνηέντων. Για τα προβλήματα αυτά αξιολογείται και η ικανότητα γενίκευσης των αλγορίθμων εκπαίδευσης.

Για την ανάλυση της στατιστικής σημασίας των αποτελεσμάτων, χρησιμοποιήθηκε η μέθοδος ελέγχου Wilcoxon paired two sided sign rank test [89], [27]. Αυτή είναι μια μη παραμετρική μέθοδος που θεωρείται εναλλακτική λύση για τη μέθοδο ελέγχου paired t-test. Ο έλεγχος αυτός υποθέτει ότι υπάρχει πληροφορία στο μέγεθος της διαφοράς μεταξύ αντιστοιχισμένων παρατηρήσεων, καθώς επίσης και στα πρόσημα. Η μηδενική υπόθεση του ελέγχου είναι ότι η διαφορά μεταξύ των αντιστοιχισμένων παρατηρήσεων προέρχεται από μια κατανομή της οποίας ο διάμεσος είναι μηδέν. Οι παρατηρήσεις στα πειράματά μας είναι η ταχύτητα σύγκλισης (αριθμός υπολογισμών συναρτησιακών τιμών και υπολογισμών κλίσεων) και ο έλεγχος πραγματοποιήθηκε με επίπεδο σημαντικότητας 5%. Δηλαδή αν η τιμή p-value είναι μικρότερη από 0.05 τότε η μηδενική υπόθεση μπορεί να απορριφθεί. Οι στατιστικώς σημαντικές περιπτώσεις σημειώνονται με το σύμβολο (+) ενώ οι στατιστικώς ασήμαντες περιπτώσεις με το σύμβολο (-). Στα πειράματά μας συγκρίθηκαν οι παρατηρήσεις του αλγορίθμου εκπαίδευσης SPerry έναντι εκείνων των αλγορίθμων που είχαν χειρότερη απόδοση από αυτόν. Ο λόγος για τον οποίο επιλέχθηκε ο αλγόριθμος εκπαίδευσης SPerry είναι ότι ο αλγόριθμος αυτός ήταν ο αποδοτικότερος στα περισσότερα προβλήματα που χρησιμοποιήθηκαν.

Για κάθε πρόβλημα, παρατίθεται ένας πίνακας ο οποίος συνοψίζει την απόδοση των αλγορίθμων για προσομοιώσεις που έφτασαν σε λύση εντός ενός προκαθορισμένου ορίου υπολογισμών της συνάρτησης σφάλματος που ορίζεται σε κάθε πρόβλημα. Οι αναφερόμενες στους πίνακες παράμετροι είναι: *Min* ο ελάχιστος αριθμός επαναλήψεων, *Max* ο μέγιστος αριθμός επαναλήψεων, *μ* η μέση τιμή των επαναλήψεων, *σ* η τυπική απόκλιση, και *Επιτυχία* ο αριθμός των πετυχημένων προσομοιώσεων από ένα σύνολο 1000 δοκιμών (εκτός κι αν αναφερθεί διαφορετικά για κάποιο πρόβλημα). Σε περίπτωση που ένας αλγόριθμος αποτύχει να συγκλίνει μέσα στο προκαθορισμένο όριο των υπολογισμών της συνάρτησης σφάλματος, θεωρείται ότι απέτυχε να εκπαιδεύσει το τεχνητό νευρωνικό δίκτυο και οι επαναλήψεις της συγκεκριμένης προσομοίωσης δεν περιλαμβάνονται στην στατιστική ανάλυση του αλγορίθμου. Για τα προβλήματα που χρησιμοποιήθηκαν για την αξιολόγηση της γενίκευσης των αλγορίθμων εκπαίδευσης, προστίθεται μια επιπλέον παράμετρος στους πίνακες, που ονομάζεται *Γενίκευση*, η οποία παρουσιάζει το ποσοστό της μέσης γενίκευσης για πετυχημένες προσομοιώσεις των αλγορίθμων εκπαίδευσης.

Σε αυτό το σημείο, πρέπει να επισημανθεί ότι γενικά μια επανάληψη ενός αλγορίθμου εκπαίδευσης σημαίνει ότι όλα τα πρότυπα εκπαίδευσης έχουν παρουσιαστεί στο τεχνητό νευρωνικό δίκτυο. Στους κλασικούς αλγορίθμους, όπως η BP, κάθε επανάληψη αντιστοιχεί σε ένα συναρτησιακό υπολογισμό της συνάρτησης σφάλματος και έναν

υπολογισμό της κλίσης (παράγωγοι πρώτης τάξεως). Στους αλγορίθμους εκπαίδευσης συζυγών κλίσεων και στις ψευδό-Newton μεθόδους λαμβάνουν χώρα περισσότεροι συναρτησιακοί υπολογισμοί και υπολογισμοί κλίσεων ανά επανάληψη, γεγονός που οφείλεται στην παρουσία της τεχνικής γραμμικής αναζήτησης. Εξαιτίας της φύσεως της γραμμικής αναζήτησεως που χρησιμοποιείται στους αλγορίθμους αυτούς, απαιτείται ο ίδιος αριθμός συναρτησιακών υπολογισμών και υπολογισμών κλίσεων ανά επανάληψη. Δηλαδή μετά την εκτέλεση της διαδικασίας της γραμμικής αναζήτησης σε μια επανάληψη του αλγορίθμου εκπαίδευσης ο ίδιος αριθμός συναρτησιακών υπολογισμών και υπολογισμών κλίσεων λαμβάνει χώρα (για παράδειγμα, στην k επανάληψη του αλγορίθμου εκπαίδευσης η διαδικασία της γραμμικής αναζήτησης χρειάστηκε r συναρτησιακούς υπολογισμούς και υπολογισμούς κλίσεων).

Στην στατιστική ανάλυση των αλγορίθμων εκπαίδευσης, εκτός από τους συναρτησιακούς υπολογισμούς και τους υπολογισμούς κλίσεων, υπολογίστηκε και ο χρόνος της κεντρικής μονάδας επεξεργασίας (CPU) σε δευτερόλεπτα (sec) που χρειάστηκε κάθε αλγόριθμος εκπαίδευσης για κάθε ένα από τα προβλήματα. Στην συνέχεια, απεικονίζονται οι μέσοι όροι των χρόνων της CPU σε ραβδογράμματα για κάθε αλγόριθμο εκπαίδευσης σε κάθε πρόβλημα. Επιπλέον, δημιουργήθηκαν γραφήματα τα οποία απεικονίζουν τον μέσο όρο των καμπυλών εκπαίδευσης για τους αλγορίθμους εκπαίδευσης για κάποια από τα προβλήματα. Στον άξονα x απεικονίζεται ο μέσος όρος του αθροίσματος των υπολογισμών της συνάρτησης σφάλματος και της κλίσης της, ενώ στον άξονα y απεικονίζεται η μέση τιμή της συνάρτησης σφάλματος. Ο άξονας y είναι σε λογαριθμική κλίμακα έτσι ώστε οι διαφορές μεταξύ των αλγορίθμων εκπαίδευσης να γίνονται εύκολα αντιληπτές. Σε αυτά τα γραφήματα λήφθηκαν υπόψιν μόνο οι προσομοιώσεις που έφτασαν στην λύση από ένα σύνολο 100 δοκιμών.

4.4.1 Αποκλειστικό-EITE

Για το πρόβλημα του αποκλειστικού-EITE (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα $2 - 2 - 1$ τεχνητό νευρωνικό δίκτυο (6 βάρη και 3 μεροληψίες). Το κρυφό επίπεδο του νευρωνικού δικτύου είναι βασισμένο σε νευρώνες με συνάρτηση ενεργοποίησης την υπερβολική εφαπτομένη, ενώ το επίπεδο εξόδου είναι βασισμένο σε νευρώνες με συνάρτηση ενεργοποίησης την γραμμική συνάρτηση. Η συνθήκη τερματισμού είναι $E \leq 0.01$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 1000. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 4.1 και στις Εικόνες 4.1 και 4.2.

Όλοι οι αλγόριθμοι εκπαίδευσης που δοκιμάστηκαν παρουσίασαν καλό ποσοστό επιτυχίας και καλή ταχύτητα σύγκλισης. Ο καλύτερος αλγόριθμος όσον αφορά το ποσοστό επιτυχίας ήταν ο SPerry. Ωστόσο, και οι υπόλοιποι αλγόριθμοι εκπαίδευσης παρουσίασαν ανταγωνιστικά ποσοστά επιτυχίας, με εξαίρεση τους GB1, GB2, FR, SFR και SPR που παρουσίασαν χαμηλότερα ποσοστά επιτυχίας (τουλάχιστον 2.6% λιγότερο από τον

Σχήμα 4.1: Μέσος χρόνος για το πρόβλημα του αποκλειστικού-EITE

Σχήμα 4.2: Καμπύλες εκπαίδευσης των αλγορίθμων εκπαίδευσης για το πρόβλημα του αποκλειστικού-EITE

αλγόριθμο SPerry). Όσον αφορά την ταχύτητα σύγκλισης, ο αλγόριθμος GB1 παρουσίασε τον ταχύτερο ρυθμό σύγκλισης. Οι υπόλοιπες μέθοδοι παρουσίασαν επίσης καλή ταχύτητα σύγκλισης, εκτός από τις μεθόδους FR και SFR (βλέπε τις Εικόνες 1.2 και 1.3). Συνολικά, οι κλιμακωτοί αλγόριθμοι συζυγών κλίσεων παρουσίασαν ελαφρώς καλύτερη απόδοση τόσο σε ποσοστό επιτυχίας όσο και σε ταχύτητα σύγκλισης σε σχέση με τους κλασικούς αλγόριθμους εκπαίδευσης συζυγών κλίσεων με κάποιες εξαιρέσεις (ο αλγόριθμος PR ήταν καλύτερος από τον SPR όσον αφορά το ποσοστό επιτυχίας, ενώ ο αλγόριθμος του Perry ήταν ελαφρώς γρηγορότερος από τον αλγόριθμο SPerry). Ο Πίνακας 4.1 παρουσιάζει, επίσης, τα αποτελέσματα του ελέγχου Wilcoxon sign rank test για τις μεθόδους που ήταν χειρότερες από τον αλγόριθμο εκπαίδευσης SPerry. Όπως είναι προφανές, η βελτίωση που παρουσιάζει ο αλγόριθμος SPerry σε σχέση με αυτούς τους αλγορίθμους είναι στατιστικά σημαντική όσον αφορά την ταχύτητα σύγκλισης.

Αλγόριθμος	min	max	μ	σ	Επιτυχία	
OSS	11	990	102.5	(+)	143.9	92.6%
GB1	8	993	81.8		158.3	90.9%
GB2	8	995	82.5		156.1	87.5%
LN	7	997	93.5		159.7	92.4%
HS	9	996	101.8	(+)	157.4	92.6%
SHS	9	975	101.4	(+)	153.7	92.9%
FR	8	961	163.8	(+)	198.2	86.8%
SFR	8	945	146.4	(+)	177.6	87.7%
PR	8	991	93.4		142.1	92.0%
SPR	8	946	87.3		131.1	90.0%
Perry	7	981	95.2		146.9	92.9%
SPerry	9	993	97.8		153.9	93.5%

Πίνακας 4.1: Συγκριτικά αποτελέσματα για το πρόβλημα του αποκλειστικού-EITE

4.4.2 Ισοτιμία των 3-bit

Για το πρόβλημα της ισοτιμίας των 3-bit (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα $3 - 2 - 1$ τεχνητό νευρωνικό δίκτυο (8 βάρη και 3 μεροληψίες). Το κρυφό επίπεδο του νευρωνικού δικτύου είναι βασισμένο σε νευρώνες με συνάρτηση ενεργοποίησης την υπερβολική εφαπτομένη, ενώ το επίπεδο εξόδου είναι βασισμένο σε νευρώνες με συνάρτηση ενεργοποίησης την γραμμική συνάρτηση. Η συνθήκη τερματισμού είναι $E \leq 0.01$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 1000. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 4.2 και στις Εικόνες 4.3 και 4.4.

Σε αυτό το πρόβλημα, ο αλγόριθμος εκπαίδευσης SPerry παρουσίασε το καλύτερο ποσοστό επιτυχίας από όλους τους υπόλοιπους αλγορίθμους. Επίσης, ήταν ο καλύτερος

Σχήμα 4.3: Μέσος χρόνος για το πρόβλημα της ισοτιμίας των 3-bit

Σχήμα 4.4: Καμπύλες εκπαίδευσης των αλγορίθμων εκπαίδευσης για το πρόβλημα της ισοτιμίας των 3-bit

αλγόριθμος όσον αφορά την ταχύτητα σύγκλισης σε σχέση με τους αλγόριθμους εκπαίδευσης συζυγών κλίσεων (κλασικών και κλιμακωτών). Οι αλγόριθμοι εκπαίδευσης GB1, GB2 και LN παρουσίασαν ταχύτερη σύγκλιση από τον αλγόριθμο SPerry, αλλά το ποσοστό επιτυχίας του ήταν αρκετά χαμηλότερο. Ο έλεγχος Wilcoxon sign rank test έδειξε ότι τα αποτελέσματα του αλγορίθμου εκπαίδευσης SPerry είναι στατιστικώς σημαντικά σε σχέση με τους άλλους αλγόριθμους που απέδωσαν χειρότερα. Οι αλγόριθμοι εκπαίδευσης FR και SFR είχαν πολύ κακή απόδοση τόσο σε ποσοστό επιτυχίας όσο και σε ταχύτητα σύγκλισης. Ωστόσο, ο αλγόριθμος εκπαίδευσης SFR ήταν βελτιωμένος σε σχέση με τον κλασικό αλγόριθμο FR. Η μέθοδος PR παρουσίασε το δεύτερο καλύτερο ποσοστό επιτυχίας σε αυτό το πρόβλημα. Ο αντίστοιχος κλιμακωτός αλγόριθμος εκπαίδευσης, SPR είχε σημαντικά χαμηλότερο ποσοστό επιτυχίας, ενώ ήταν ελάχιστα ταχύτερος. Ο αλγόριθμος εκπαίδευσης SHS απέδωσε καλύτερα σε ποσοστό επιτυχίας, αλλά συνέκλινε βραδύτερα από τον αλγόριθμο HS.

Αλγόριθμος	min	max	μ	σ	Επιτυχία
OSS	111	990	347.6 (+)	193.3	65.5%
GB1	42	987	191.1	174.4	71.2%
GB2	39	975	189.9	178.6	65.9%
LN	49	991	228.3	174.7	76.5%
HS	67	998	247.9 (+)	167.9	74.6%
SHS	67	995	253.7 (+)	178.1	74.9%
FR	97	992	468.9 (+)	248.2	23.9%
SFR	93	997	445.2 (+)	254.4	28.5%
PR	59	998	269.8 (+)	174.9	77.3%
SPR	59	993	267.1 (+)	163.6	71.5%
Perry	64	995	233.1 (+)	162.7	73.0%
SPerry	69	996	232.9	169.1	80.6%

Πίνακας 4.2: Συγκριτικά αποτελέσματα για το πρόβλημα της ισοτιμίας των 3-bit

4.4.3 $N - M - N$ Κωδικοποιητής/Αποκωδικοποιητής

Για το πρόβλημα του $N - M - N$ Κωδικοποιητή/Αποκωδικοποιητή (βλέπε Παράρτημα Α') χρησιμοποιήσαμε μια σειρά από επτά τεχνητά νευρωνικά δίκτυα. Ουσιαστικά, πραγματοποιήσαμε δοκιμές για $N = 4, 8, 16, 32, 64, 128$ και 256 . Οι νευρώνες του κρυφού επιπέδου, αλλά και του επιπέδου εξόδου είχαν ως συναρτήσεις ενεργοποίησης την λογιστική συνάρτηση. Στον Πίνακα 4.3, δίνονται οι προδιαγραφές των επτά προβλημάτων Κωδικοποίησης/Αποκωδικοποίησης. Η αρχικοποίηση των βαρών έγινε με την τεχνική των Nguyen και Widrow, αλλά στη συνέχεια τα βάρη μεταξύ του κρυφού επιπέδου και του επιπέδου της εξόδου πολλαπλασιάστηκαν με 0.01. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στους Πίνακες 4.4-4.10. Επίσης, η μέση τιμή της ταχύτητας εκπαίδευσης για όλους τους αλγορίθμους εκπαίδευσης απεικονίζεται στις Εικόνες

4.5-4.13.

Πρόβλημα	Δοκιμές	Όριο Επ.	Βάρη	Μεροληψίες	e.g.
4-2-4	1000	1000	16	6	0.1
8-3-8	1000	1000	48	11	0.1
16-4-16	1000	1000	128	20	0.1
32-5-32	1000	1000	320	37	0.1
64-6-64	1000	2000	768	70	0.1
128-7-128	100	10000	1792	135	0.1
256-8-256	10	50000	4096	264	0.1

Πίνακας 4.3: Προδιαγραφές των προβλημάτων $N - M - N$ Κωδικοποιητή/Αποκωδικοποιητή. Οι αναφερόμενες παράμετροι είναι: Δοκιμές, ο αριθμός των εξομοιώσεων, Όριο Επ., ο μέγιστος επιτρεπτός αριθμός συναρτησιακών υπολογισμών, Βάρη, ο αριθμός των συναπτικών βαρών, Μεροληψίες, ο αριθμός των μεροληψιών, και e, g , η συνθήκη τερματισμού.

Ο αλγόριθμος εκπαίδευσης SPerry ήταν ο καλύτερος όσον αφορά το ποσοστό επιτυχίας σε ολόκληρη την οικογένεια των προβλημάτων Κωδικοποιητή/Αποκωδικοποιητή, με εξαίρεση το πρόβλημα 8-3-8 όπου ο αλγόριθμος εκπαίδευσης LN είχε ελάχιστα καλύτερο ποσοστό επιτυχίας. Αξίζει να σημειωθεί ότι καθώς η διάσταση του προβλήματος μεγαλώνει, ο αλγόριθμος εκπαίδευσης SPerry διατηρεί τα υψηλά ποσοστά επιτυχίας, ενώ το ποσοστό επιτυχίας των υπολοίπων αλγορίθμων εκπαίδευσης μειώνεται δραστικά. Ο αλγόριθμος SPerry παρουσιάζει πολύ καλή ταχύτητα σύγκλισης. Ειδικότερα στις υψηλές διαστάσεις του προβλήματος του Κωδικοποιητή/Αποκωδικοποιητή η ταχύτητα σύγκλισης του αλγορίθμου ήταν μακράν η ταχύτερη. Σε μερικές περιπτώσεις παρουσίασε χαμηλότερη ταχύτητα σύγκλισης σε σχέση με τον αλγόριθμο εκπαίδευσης LN (για τα προβλήματα 4-2-4 και 16-4-16) αλλά η διαφορά ήταν πολύ μικρή ενώ συνέκλινε σε περισσότερες προσομοιώσεις. Οι κλιμακωτοί αλγόριθμοι εκπαίδευσης συζυγών κλίσεων είχαν καλύτερη απόδοση τόσο σε ποσοστό επιτυχίας όσο και σε ταχύτητα σύγκλισης στις περισσότερες περιπτώσεις σε σχέση με τους αντίστοιχους κλασικούς αλγορίθμους εκπαίδευσης συζυγών κλίσεων. Επίσης, αξίζει να σημειωθεί ότι στα προβλήματα Κωδικοποιητή/Αποκωδικοποιητή 128-7-128 και 256-8-256 οι αλγόριθμοι εκπαίδευσης OSS, GB1, GB2, FR και SFR απέτυχαν να συγκλίνουν σε όλες τις εξομοιώσεις που επιχειρήθηκαν. Είναι πιθανό ότι μια αύξηση του μέγιστου αριθμού επιτρεπόμενων συναρτησιακών υπολογισμών θα είχε ως αποτέλεσμα να συγκλίνουν σε κάποιες εξομοιώσεις. Ωστόσο, το υπολογιστικό τους κόστος θα ήταν σημαντικά υψηλότερο από τους υπόλοιπους αλγορίθμους εκπαίδευσης και ειδικότερα τον SPerry. Στους Πίνακες 4.4-4.10 παρουσιάζονται τα αποτελέσματα του ελέγχου Wilcoxon sign rank test. Τα αποτελέσματα του αλγορίθμου εκπαίδευσης SPerry όσον αφορά τους υπολογισμούς της συνάρτησης σφάλματος και της κλίσεως της ήταν στατιστικά σημαντικά σε σχέση με τους αλγορίθμους που απέδωσαν χειρότερα. Μόνο στην περίπτωση του προβλήματος 8-3-8 τα αποτελέσματα του αλγορίθμου εκπαίδευσης SPerry δεν ήταν στατιστικώς σημαντικά συγκρινόμενα με τα αποτελέσματα του αλγορίθμου εκπαίδευσης LN.

Αλγόριθμος	min	max	μ	σ	Επιτυχία
OSS	16	965	117.9 (+)	107.1	92.0%
GB1	16	999	98.5 (+)	95.9	94.6%
GB2	16	991	111.6 (+)	109.1	93.1%
LN	10	659	51.7	53.1	99.7%
HS	11	657	58.2 (+)	52.8	99.8%
SHS	11	659	58.1 (+)	52.9	99.8%
FR	12	995	129.7 (+)	176.1	95.4%
SFR	12	864	78.7 (+)	101.5	99.3%
PR	13	379	63.9 (+)	49.9	100.0%
SPR	13	674	59.7 (+)	47.5	100.0%
Perry	13	965	99.8 (+)	98.9	97.7%
SPerry	12	656	51.9	45.1	100.0%

Πίνακας 4.4: Συγκριτικά αποτελέσματα για το πρόβλημα του 4 – 2 – 4 Κωδικοποιητή/Αποκωδικοποιητή

Σχήμα 4.5: Μέσος χρόνος για το πρόβλημα του 4 – 2 – 4 Κωδικοποιητή / Αποκωδικοποιητή

Σχήμα 4.6: Καμπύλες εκπαίδευσης των αλγορίθμων εκπαίδευσης για το πρόβλημα του 4 – 2 – 4 Κωδικοποιητή / Αποκωδικοποιητή

Αλγόριθμος	min	max	μ	σ	Επιτυχία
OSS	57	998	245.9 (+)	164.9	80.8%
GB1	59	975	220.9 (+)	162.3	86.8%
GB2	65	996	236.1 (+)	169.7	84.1%
LN	27	973	130.6 (–)	127.5	99.7%
HS	31	999	161.7 (+)	164.1	98.9%
SHS	31	974	158.4 (+)	155.7	98.9%
FR	24	995	226.2 (+)	205.9	77.1%
SFR	24	976	179.7 (+)	169.8	96.1%
PR	40	950	175.2 (+)	132.5	97.0%
SPR	40	944	161.7 (+)	117.6	97.1%
Perry	39	981	186.4 (+)	123.4	98.4%
SPerry	29	977	130.4	128.2	99.2%

Πίνακας 4.5: Συγκριτικά αποτελέσματα για το πρόβλημα του 8 – 3 – 8 Κωδικοποιητή/Αποκωδικοποιητή

Σχήμα 4.7: Μέσος χρόνος για το πρόβλημα του 8 – 3 – 8 Κωδικοποιητή / Αποκωδικοποιητή

Σχήμα 4.8: Καμπύλες εκπαίδευσης των αλγορίθμων εκπαίδευσης για το πρόβλημα του 8 – 3 – 8 Κωδικοποιητή / Αποκωδικοποιητή

Αλγόριθμος	min	max	μ	σ	Επιτυχία
OSS	114	997	383.1 (+)	170.1	57.7%
GB1	130	999	363.3 (+)	177.1	65.4%
GB2	117	995	389.4 (+)	187.6	59.6%
LN	83	931	271.2	168.8	96.2%
HS	89	998	412.9 (+)	244.1	89.6%
SHS	89	999	403.6 (+)	237.1	88.7%
FR	89	990	424.1 (+)	254.9	25.0%
SFR	82	994	394.5 (+)	245.8	59.1%
PR	128	997	359.1 (+)	183.9	84.6%
SPR	132	998	340.5 (+)	169.6	85.6%
Perry	91	997	313.8 (+)	163.3	90.5%
SPerry	73	996	273.9	193.8	98.6%

Πίνακας 4.6: Συγκριτικά αποτελέσματα για το πρόβλημα του 16 – 4 – 16 Κωδικοποιητή/Αποκωδικοποιητή

Σχήμα 4.9: Μέσος χρόνος για το πρόβλημα του 16 – 4 – 16 Κωδικοποιητή / Αποκωδικοποιητή

Αλγόριθμος	min	max	μ	σ	Επιτυχία
OSS	179	988	504.9 (+)	201.4	11.6%
GB1	197	999	524.7 (+)	192.8	10.5%
GB2	207	997	516.7 (+)	204.8	11.6%
LN	144	999	509.2 (+)	198.6	72.7%
HS	256	998	595.3 (+)	209.5	63.8%
SHS	225	999	591.7 (+)	206.8	62.9%
FR	220	901	528.5 (+)	234.2	1.5%
SFR	168	992	595.9 (+)	247.5	8.3%
PR	319	999	621.2 (+)	172.1	56.0%
SPR	327	995	617.6 (+)	164.9	59.3%
Perry	170	997	503.5 (+)	190.9	54.3%
SPerry	227	996	498.6	201.7	85.9%

Πίνακας 4.7: Συγκριτικά αποτελέσματα για το πρόβλημα του 32 – 5 – 32 Κωδικοποιητή/Αποκωδικοποιητή

Σχήμα 4.10: Μέσος χρόνος για το πρόβλημα του 32 – 5 – 32 Κωδικοποιητή / Αποκωδικοποιητή

Αλγόριθμος	min	max	μ		σ	Επιτυχία
OSS	445	1915	1088.1	(+)	431.4	3.6%
GB1	392	1977	1052.1	(+)	445.1	4.7%
GB2	367	1983	1209.7	(+)	516.4	3.7%
LN	344	1997	1144.3	(+)	397.1	43.9%
HS	636	1999	1310.5	(+)	358.8	51.3%
SHS	637	1999	1325.4	(+)	363.7	52.4%
FR	1504	1504	1504	(+)	0	0.1%
SFR	592	1999	1220.5	(+)	522.2	0.8%
PR	820	1982	1415.5	(+)	266.2	41.9%
SPR	861	1999	1476.7	(+)	287.1	32.3%
Perry	346	1988	1008.2		372.9	23.8%
SPerry	575	1996	1088.3		347.9	86.2%

Πίνακας 4.8: Συγκριτικά αποτελέσματα για το πρόβλημα του 64 – 6 – 64 Κωδικοποιητή/Αποκωδικοποιητή

Σχήμα 4.11: Μέσος χρόνος για το πρόβλημα του 64 – 6 – 64 Κωδικοποιητή / Αποκωδικοποιητή

Αλγόριθμος	min	max	μ		σ	Επιτυχία
OSS	-	-	-		-	-
GB1	-	-	-		-	-
GB2	-	-	-		-	-
LN	1698	9455	5619.5	(+)	2352.7	48%
HS	1756	9873	4263.9	(+)	2359.1	77%
SHS	1562	9544	4286.9	(+)	2251.1	80%
FR	-	-	-		-	-
SFR	-	-	-		-	-
PR	2465	9872	4572.3	(+)	2068.1	59%
SPR	1974	9804	6585.1	(+)	2661.2	31%
Perry	1459	7760	3680.9	(+)	1935.1	30%
SPerry	1211	9313	3117.1		1642.9	91%

Πίνακας 4.9: Συγκριτικά αποτελέσματα για το πρόβλημα του 128 – 7 – 128 Κωδικοποιητή/Αποκωδικοποιητή

Σχήμα 4.12: Μέσος χρόνος για το πρόβλημα του 128 – 7 – 128 Κωδικοποιητή / Αποκωδικοποιητή

Αλγόριθμος	min	max	μ	σ	Επιτυχία	
OSS	-	-	-	-	-	
GB1	-	-	-	-	-	
GB2	-	-	-	-	-	
LN	10377	46780	29146.2	(+)	13255.1	50%
HS	10882	26464	17909.8	(+)	6432.4	60%
SHS	6018	42443	23497	(+)	14148.1	60%
FR	-	-	-	-	-	-
SFR	-	-	-	-	-	-
PR	49003	49003	49003	(+)	0	10%
SPR	48826	48826	48826	(+)	0	10%
Perry	9309	13094	11210.3	(+)	1892.6	30%
SPerry	3975	9728	6015.9		2090.5	70%

Πίνακας 4.10: Συγκριτικά αποτελέσματα για το πρόβλημα του 256 – 8 – 256 Κωδικοποιητή/Αποκωδικοποιητή

Σχήμα 4.13: Μέσος χρόνος για το πρόβλημα του 256 – 8 – 256 Κωδικοποιητή / Αποκωδικοποιητή

Σχήμα 4.14: Μέσος χρόνος για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων

4.4.4 Αναγνώριση Κεφαλαίων Γραμμάτων

Για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων της Αγγλικής αλφαβήτου (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα 35 – 30 – 26 τεχνητό νευρωνικό δίκτυο (1830 βάρη και 56 μεροληψίες). Το κρυφό επίπεδο του νευρωνικού δικτύου είναι βασισμένο σε νευρώνες με λογιστική συνάρτηση ενεργοποίησης, ενώ το επίπεδο εξόδου είναι βασισμένο σε νευρώνες γραμμικής συνάρτησης ενεργοποίησης. Η συνθήκη τερματισμού είναι $E \leq 0.1$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 2000. Η αρχικοποίηση των βαρών έγινε με την τεχνική των Nguyen και Widrow, αλλά στη συνέχεια τα βάρη μεταξύ του κρυφού επιπέδου και του επιπέδου της εξόδου πολλαπλασιάστηκαν με 0.01. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 4.11 και στις Εικόνες 4.14 και 4.15.

Σε αυτό το πρόβλημα όλοι οι αλγόριθμοι απέδωσαν πολύ καλά όσον αφορά το ποσοστό επιτυχίας. Κατάφεραν να συγκλίνουν σε όλες τις εξομοιώσεις που πραγματοποιήθηκαν, εκτός από τους αλγορίθμους εκπαίδευσης OSS, PR και SPR. Ωστόσο, οι αλγόριθμοι εκπαίδευσης που δεν συνέκλιναν σε όλες τις εξομοιώσεις παρουσίασαν ικανοποιητικά ποσοστά επιτυχίας. Οι ψευδο-Newton αλγόριθμοι εκπαίδευσης χρειάστηκαν αρκετούς υπολογισμούς της συνάρτησης σφάλματος και της κλίσεως της για να συγκλίνουν, με εξαίρεση τον αλγόριθμο εκπαίδευσης LN ο οποίος ήταν ανταγωνιστικός με τους αλγορίθμους συζυγών κλίσεων. Ο καλύτερος αλγόριθμος από τους κλασικούς αλγορίθμους εκπαίδευσης συζυγών κλίσεων ήταν ο προτεινόμενος αλγόριθμος Perry. Οι

Σχήμα 4.15: Καμπύλες εκπαίδευσης των αλγορίθμων εκπαίδευσης για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων

Αλγόριθμος	min	max	μ	σ	Επιτυχία
OSS	1033	1999	1517.3 (+)	241.1	90.0%
GB1	888	1320	1034.7 (+)	72.7	100.0%
GB2	899	1192	1020.1 (+)	65.5	100.0%
LN	230	495	345.9 (+)	48.8	100.0%
HS	195	533	355.7 (+)	65.9	100.0%
SHS	234	561	359.5 (+)	66.9	100.0%
FR	266	553	390.1 (+)	61.7	100.0%
SFR	273	593	398.1 (+)	62.8	100.0%
PR	293	1651	612.7 (+)	338.1	99.3%
SPR	293	1952	557.3 (+)	371.9	93.5%
Perry	219	534	344.3 (+)	59.2	100.0%
SPerry	188	569	329.8	67.1	100.0%

Πίνακας 4.11: Συγκριτικά αποτελέσματα για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων

κλιμακωτοί αλγόριθμοι εκπαίδευσης συζυγών κλίσεων είχαν την ίδια συμπεριφορά με τα κλασικά τους ανάλογα. Οι αλγόριθμοι εκπαίδευσης SHS και SFR χρειάστηκαν ελαφρώς περισσότερους συναρτησιακούς υπολογισμούς και υπολογισμούς κλίσεων από τα κλασικά τους ανάλογα, ενώ οι αλγόριθμοι εκπαίδευσης SPerry και SPR χρειάστηκαν λιγότερους συναρτησιακούς υπολογισμούς και υπολογισμούς κλίσεων από τους αλγόριθμους εκπαίδευσης Perry και PR, αντίστοιχα. Ο καλύτερος αλγόριθμος εκπαίδευσης όσον αφορά το ποσοστό επιτυχίας και στην ταχύτητα σύγκλισης ήταν ο αλγόριθμος SPerry. Επίσης, όπως φαίνεται από τον έλεγχο Wilcoxon sign rank test, τα αποτελέσματα του αλγορίθμου SPerry ήταν στατιστικά σημαντικά σε σχέση με τα αποτελέσματα των αλγορίθμων που ήταν χειρότεροι από αυτόν όσον αφορά την ταχύτητα σύγκλισης.

4.4.5 Αναγνώριση Αριθμών

Για το πρόβλημα της αναγνώρισης αριθμών (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα $64 - 6 - 10$ τεχνητό νευρωνικό δίκτυο (444 βάρη και 16 μεροληψίες). Το κρυφό επίπεδο και το επίπεδο εξόδου του νευρωνικού δικτύου είναι βασισμένα σε νευρώνες με λογιστική συνάρτηση ενεργοποίησης. Η συνθήκη τερματισμού είναι $E \leq 0.001$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 2000. Τα αρχικά βάρη του νευρωνικού δικτύου αρχικοποιήθηκαν με τυχαίους αριθμούς μέσα από το κλειστό διάστημα $[-1, 1]$. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 4.12 και στις Εικόνες 4.16 και 4.17.

Σε αυτό το πρόβλημα, ο αλγόριθμος εκπαίδευσης SPerry παρουσίασε την καλύτερη επίδοση τόσο σε ποσοστό επιτυχίας όσο και σε ταχύτητα σύγκλισης. Επιπλέον, τα αποτελέσματα του αλγορίθμου εκπαίδευσης SPerry όσον αφορά την ταχύτητα σύγκλισης ήταν στατιστικά σημαντικά συγκρινόμενα με τους υπόλοιπους αλγορίθμους εκπαίδευσης εκτός από τον αλγόριθμο LN, όπως φαίνεται από τον έλεγχο Wilcoxon sign rank test. Συνέκλινε σε όλες τις εξομοιώσεις που πραγματοποιήθηκαν με λιγότερους υπολογισμούς τιμών συνάρτησης σφάλματος και υπολογισμούς κλίσεων. Κι άλλοι αλγόριθμοι εκπαίδευσης, όπως οι HS, SHS και SPR, κατάφεραν να συγκλίνουν σε όλες τις εξομοιώσεις, αλλά ήταν σημαντικά πιο δαπανηροί σε υπολογισμούς. Οι κλιμακωτοί αλγόριθμοι εκπαίδευσης συζυγών κλίσεων παρουσιάστηκαν βελτιωμένοι σε σχέση με τα κλασικά τους ανάλογα. Οι αλγόριθμοι εκπαίδευσης SFR, SPR και SPerry συνέκλιναν σε περισσότερες εξομοιώσεις ενώ παρουσίασαν ταχύτερη σύγκλιση σε σχέση με τους αλγόριθμους εκπαίδευσης FR, PR και Perry. Ο αλγόριθμος εκπαίδευσης SHS συνέκλινε σε όλες τις εξομοιώσεις όπως και ο αλγόριθμος HS, αλλά όσον αφορά την ταχύτητα σύγκλισης ήταν ελαφρώς χειρότερος. Οι αλγόριθμοι εκπαίδευσης ψευδό-Newton, με εξαίρεση τον αλγόριθμο LN που ήταν ανταγωνιστικός με τους αλγόριθμους εκπαίδευσης συζυγών κλίσεων, παρουσίασαν χαμηλότερα ποσοστά επιτυχίας σε αυτό το πρόβλημα και απαιτούσαν περισσότερους συναρτησιακούς υπολογισμούς και υπολογισμούς κλίσεων για να συγκλίνουν.

Σχήμα 4.16: Μέσος χρόνος για το πρόβλημα της αναγνώρισης αριθμών

Σχήμα 4.17: Καμπύλες εκπαίδευσης των αλγορίθμων εκπαίδευσης για το πρόβλημα της αναγνώρισης αριθμών

Αλγόριθμος	min	max	μ		σ	Επιτυχία
OSS	47	19422	1021.3	(+)	2455.8	81.5%
GB1	45	19432	1080.4	(+)	2303.5	84.6%
GB2	45	19900	1192.7	(+)	2760.3	83.5%
LN	36	15732	245.5	(-)	644.3	98.7%
HS	62	3287	244.9	(+)	247.1	100.0%
SHS	62	3475	246.4	(+)	266.1	100.0%
FR	50	17442	1382.5	(+)	2144.4	93.7%
SFR	50	9425	376.6	(+)	613.4	98.9%
PR	61	19025	428.1	(+)	940.5	99.5%
SPR	61	11644	312.9	(+)	530.1	100.0%
Perry	34	19901	2033.4	(+)	3766.9	90.8%
SPerry	45	1754	182.2		162.6	100.0%

Πίνακας 4.12: Συγκριτικά αποτελέσματα για το πρόβλημα της αναγνώρισης αριθμών

4.4.6 Ταξινόμηση των Φυτών της Οικογένειας Iris

Για το πρόβλημα της ταξινόμησης των φυτών της οικογένειας Iris (βλέπε Παράρτημα Α') χρησιμοποιήσαμε την μέθοδο της διασταυρωμένης επικύρωσης από 10 μέρη. Τα δεδομένα χωρίστηκαν σε 10 υποσύνολα το καθένα από τα οποία είχαν 5 παρουσίες από κάθε κατηγορία. Τα σύνολα εκπαίδευσης περιέχουν 135 πρότυπα, 45 από κάθε κατηγορία. Τα υπόλοιπα 15 πρότυπα χρησιμοποιήθηκαν για τον έλεγχο της γενίκευσης των αλγορίθμων εκπαίδευσης. Κάθε φορά ένα από τα 10 υποσύνολα χρησιμοποιούνταν για την εξέταση της γενίκευσης ενώ τα υπόλοιπα για την εκπαίδευση του δικτύου. Για το πρόβλημα αυτό χρησιμοποιήσαμε ένα 4–4–3 τεχνητό νευρωνικό δίκτυο (28 βάρη και 7 μεροληψίες). Το κρυφό επίπεδο και το επίπεδο εξόδου του νευρωνικού δικτύου είναι βασισμένα σε νευρώνες με λογιστική συνάρτηση ενεργοποίησης. Η συνθήκη τερματισμού είναι $E \leq 0.02$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 1000. Τα αρχικά βάρη του νευρωνικού δικτύου αρχικοποιήθηκαν με τυχαίους αριθμούς μέσα από το κλειστό διάστημα $[-1, 1]$. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 4.13 και την Εικόνα 4.18.

Όσον αφορά τους κλασικούς αλγόριθμους εκπαίδευσης συζυγών κλίσεων, οι αλγόριθμοι PR και Perry παρουσίασαν την καλύτερη επίδοση αφού συνέκλιναν σε όλες τις εξομοιώσεις που διεξήχθησαν, ακολουθούμενοι από τους αλγόριθμους HS και FR. Όσον αφορά την ταχύτητα σύγκλισης, ο αλγόριθμος FR συνέκλινε γρηγορότερα σε αυτό το πρόβλημα ακολουθούμενος από τους αλγόριθμους Perry, HS και PR. Οι κλιμακωτοί αλγόριθμοι εκπαίδευσης συζυγών κλίσεων παρουσιάστηκαν βελτιωμένοι σε κάποιες περιπτώσεις έναντι των κλασικών τους αναλόγων. Πιο συγκεκριμένα, οι αλγόριθμοι εκπαίδευσης SPerry και SHS απέδωσαν παρόμοια με τα κλασικά τους ανάλογα. Παρουσίασαν το ίδιο ποσοστό επιτυχίας, αλλά ο αλγόριθμος SPerry ήταν ταχύτερος σε σχέση με τον αλγόριθμο Perry, ενώ ο αλγόριθμος SHS ήταν οριακά πιο αργός από τον

Σχήμα 4.18: Μέσος χρόνος για το πρόβλημα της ταξινόμησης των φυτών της οικογένειας Iris

Αλγόριθμος	min	max	μ		σ	Επιτυχία	Γενίκευση
OSS	18	979	154.5	(+)	92.4	95.6%	96.8%
GB1	14	910	51.8	(+)	37.6	99.2%	96.7%
GB2	15	823	53.8	(+)	36.8	99.2%	96.8%
LN	17	307	51.9	(+)	19.9	99.9%	96.7%
HS	17	488	55.2	(+)	28.1	99.8%	96.8%
SHS	17	505	55.3	(+)	29.6	99.8%	96.8%
FR	18	504	48.9	(+)	31.4	99.5%	96.8%
SFR	18	780	49.8	(+)	48.6	99.8%	96.7%
PR	22	360	71.5	(+)	27.7	100.0%	96.9%
SPR	22	590	77.9	(+)	41.9	99.6%	96.9%
Perry	15	192	50.5	(+)	17.2	100.0%	96.9%
SPerry	19	237	48.6		16.8	100.0%	96.9%

Πίνακας 4.13: Συγκριτικά αποτελέσματα για το πρόβλημα της ταξινόμησης των φυτών της οικογένειας Iris

αλγόριθμο HS. Ο αλγόριθμος SFR παρουσιάστηκε καλύτερος σε ποσοστό επιτυχίας σε σχέση με τον αλγόριθμο FR, αλλά με ελάχιστα παραπάνω συναρτησιακούς υπολογισμούς και υπολογισμούς κλίσεων. Ο αλγόριθμος εκπαίδευσης SPR παρουσιάστηκε χειρότερος από τον αλγόριθμο PR. Οι αλγόριθμοι εκπαίδευσης ψευδό-Newton απέδωσαν επίσης καλά και ήταν ανταγωνιστικοί με τους αλγόριθμους εκπαίδευσης συζυγών κλίσεων με εξαίρεση τον αλγόριθμο OSS. Όσον αφορά την γενίκευση των αλγορίθμων εκπαίδευσης, οι αλγόριθμοι PR, Perry, SPR και SPerry παρουσίασαν την καλύτερη επίδοση. Οι υπόλοιποι αλγόριθμοι εκπαίδευσης παρουσίασαν επίσης καλό ποσοστό γενίκευσης, αλλά ήταν χαμηλότερο από το ποσοστό των προαναφερθέντων αλγορίθμων. Συνοψίζοντας την απόδοση των αλγορίθμων εκπαίδευσης για το πρόβλημα της ταξινόμησης των φυτών της οικογένειας Iris, ο καλύτερος αλγόριθμος εκπαίδευσης ήταν ο SPerry αφού παρουσίασε ένα εξαιρετικό ποσοστό επιτυχίας, με τη γρηγορότερη ταχύτητα σύγκλισης και εξαιρετικό ποσοστό γενίκευσης. Επίσης, ο έλεγχος Wilcoxon sign rank test έδειξε ότι τα αποτελέσματα του αλγορίθμου εκπαίδευσης SPerry είναι στατιστικά σημαντικά έναντι των υπολοίπων αλγορίθμων όσον αφορά την ταχύτητα σύγκλισης.

4.4.7 Οπτική Αναγνώριση Χειρόγραφων Αριθμών

Για το πρόβλημα της οπτικής αναγνώρισης χειρόγραφων αριθμών (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα 64-32-10 τεχνητό νευρωνικό δίκτυο (2368 βάρη και 42 μεροληψίες). Το κρυφό επίπεδο και το επίπεδο εξόδου του νευρωνικού δικτύου είναι βασισμένα σε νευρώνες με λογιστική συνάρτηση ενεργοποίησης. Η εκπαίδευση του τεχνητού νευρωνικού δικτύου ολοκληρώνεται όταν το δίκτυο έχει σφάλμα ταξινόμησης του συνόλου εκπαίδευσης λιγότερο από 2%. Ο μέγιστος επιτρεπτός αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 10000 και πραγματοποιήθηκαν 100 εξομοιώσεις. Τα αρχικά βάρη του νευρωνικού δικτύου αρχικοποιήθηκαν με τυχαίους αριθμούς μέσα από το κλειστό διάστημα $[-1, 1]$. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 4.14 και την Εικόνα 4.19.

Σε αυτό το πρόβλημα, οι αλγόριθμοι εκπαίδευσης SHS, SPR και SPerry παρουσίασαν τα ίδια εξαιρετικά ποσοστά επιτυχίας σε σχέση με τα κλασικά τους ανάλογα, αλλά είχαν ταχύτερη σύγκλιση. Από την άλλη μεριά, ο αλγόριθμος SFR όχι μόνο παρουσίασε ταχύτερη σύγκλιση από τον αλγόριθμο FR, αλλά είχε και μεγαλύτερο ποσοστό επιτυχίας. Όσον αφορά τους αλγόριθμους εκπαίδευσης ψευδό-Newton, οι αλγόριθμοι OSS, GB1 και GB2 είχαν κακή απόδοση σε αυτό το πρόβλημα, αφού παρουσίασαν χαμηλά ποσοστά επιτυχίας και αργή σύγκλιση. Από την άλλη μεριά, ο αλγόριθμος εκπαίδευσης LN απέδωσε εξαιρετικά αφού συνέκλινε σε όλες τις εξομοιώσεις που πραγματοποιήθηκαν με την ταχύτερη σύγκλιση. Όσον αφορά τα αποτελέσματα γενίκευσης των αλγορίθμων εκπαίδευσης, οι αλγόριθμοι ψευδό-Newton παρουσίασαν τα χαμηλότερα ποσοστά γενίκευσης. Ο αλγόριθμος εκπαίδευσης SHS παρουσίασε υψηλότερο μέσο ποσοστό γενίκευσης από τον αλγόριθμο HS. Οι υπόλοιποι κλιμακωτοί αλγόριθμοι εκπαίδευσης συζυγών κλίσεων είχαν την ίδια επίδοση όσον αφορά την γενίκευση με τα κλασικά τους ανάλογα. Οι

Σχήμα 4.19: Μέσος χρόνος για το πρόβλημα της οπτικής αναγνώρισης χειρόγραφων αριθμών

Αλγόριθμος	min	max	μ	σ	Επιτυχία	Γενίκευση
OSS	357	2126	979.5 (+)	566.8	12%	94.7%
GB1	239	7184	1599.4 (+)	2306.9	12%	94.6%
GB2	213	1093	441.5 (+)	323.1	10%	94.8%
LN	77	1259	161.4	174.2	100%	94.8%
HS	65	1235	235.1 (+)	250.3	99%	95.2%
SHS	65	1169	233.4 (+)	249.9	99%	95.3%
FR	75	4001	435.4 (+)	818.8	97%	94.9%
SFR	65	3157	349.3 (+)	518.1	99%	94.9%
PR	102	9755	429.5 (+)	1135.6	100%	95.3%
SPR	120	5554	417.5 (+)	790.2	100%	95.3%
Perry	80	2251	244.4 (+)	299.8	100%	95.3%
SPerry	66	1059	197.9	212.3	100%	95.3%

Πίνακας 4.14: Συγκριτικά αποτελέσματα για το πρόβλημα της οπτικής αναγνώρισης χειρόγραφων αριθμών

Αλγόριθμος	min	max	μ		σ	Επιτυχία	Γενίκευση
OSS	890	2334	1113.5	(+)	327.7	21%	55.4%
GB1	674	1612	1037.5	(+)	323.1	8%	49.1%
GB2	746	2708	1392.5	(+)	891.4	6%	49.2%
LN	238	2730	511.8	(+)	380.5	99%	53.5%
HS	288	1629	447.2	(+)	156.6	99%	55.3%
SHS	312	1790	456.1	(+)	189.7	100%	55.3%
FR	263	4166	955.2	(+)	759.9	83%	51.9%
SFR	275	4039	899.7	(+)	635.2	98%	52.1%
PR	476	4502	981.1	(+)	650.6	100%	55.5%
SPR	460	3730	928.3	(+)	555.3	100%	55.6%
Perry	279	981	444.6	(+)	110.1	98%	55.4%
SPerry	298	665	404.9		78.4	100%	55.4%

Πίνακας 4.15: Συγκριτικά αποτελέσματα για το πρόβλημα της αναγνώρισης φωνηέντων

αλγόριθμοι εκπαίδευσης PR, SPR, Perry και SPerry παρουσίασαν το υψηλότερο μέσο ποσοστό γενίκευσης. Συνολικά, ο αλγόριθμος εκπαίδευσης SPerry απέδωσε καλύτερα από τους υπόλοιπους αλγορίθμους σε αυτό το πρόβλημα αφού συνέκλινε σε όλες τις εξομοιώσεις με την δεύτερη καλύτερη ταχύτητα σύγκλισης (κοντά στην επίδοση του αλγορίθμου LN) και το καλύτερο μέσο ποσοστό γενίκευσης. Επίσης ο έλεγχος Wilcoxon sign rank test έδειξε ότι τα αποτελέσματα του αλγορίθμου εκπαίδευσης SPerry είναι στατιστικά σημαντικά έναντι των υπολοίπων αλγορίθμων που απέδωσαν χειρότερα από αυτόν όσον αφορά την ταχύτητα σύγκλισης.

4.4.8 Αναγνώριση Φωνηέντων

Για το πρόβλημα της αναγνώρισης φωνηέντων (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα 10-88-11 τεχνητό νευρωνικό δίκτυο (1856 βάρη και 99 μεροληψίες). Το κρυφό επίπεδο και το επίπεδο εξόδου του νευρωνικού δικτύου είναι βασισμένα σε νευρώνες με λογιστική συνάρτηση ενεργοποίησης. Η εκπαίδευση του τεχνητού νευρωνικού δικτύου ολοκληρώνεται όταν το δίκτυο έχει σφάλμα ταξινόμησης του συνόλου εκπαίδευσης λιγότερο από 5%. Ο μέγιστος επιτρεπτός αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 5000 και πραγματοποιήθηκαν 100 εξομοιώσεις. Τα αρχικά βάρη του νευρωνικού δικτύου αρχικοποιήθηκαν με τυχαίους αριθμούς μέσα από το κλειστό διάστημα $[-0.3, 0.3]$. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 4.15 και την Εικόνα 4.20.

Σε αυτό το πρόβλημα, οι κλιμακωτοί αλγόριθμοι εκπαίδευσης συζυγών κλίσεων παρουσίασαν βελτιωμένη απόδοση σε σχέση με τα κλασικά τους ανάλογα. Ο αλγόριθμος εκπαίδευσης SHS ήταν ελάχιστα πιο αργός σε σχέση με τον αλγόριθμο HS, αλλά είχε μεγαλύτερο ποσοστό επιτυχίας. Ο αλγόριθμος εκπαίδευσης SPR παρουσίασε το ίδιο

Σχήμα 4.20: Μέσος χρόνος για το πρόβλημα της αναγνώρισης φωνηέντων

ποσοστό επιτυχίας με τον αλγόριθμο PR, αλλά συνέκλινε ταχύτερα. Οι αλγόριθμοι εκπαίδευσης SFR και SPerry ήταν καλύτεροι από τα κλασικά τους ανάλογα τόσο σε ποσοστά επιτυχίας όσο και σε ταχύτητα σύγκλισης. Ο αλγόριθμος εκπαίδευσης SPerry παρουσίασε την ταχύτερη σύγκλιση, και όπως φάνηκε από τον έλεγχο Wilcoxon sign rank test, τα αποτελέσματα ήταν στατιστικά σημαντικά σε σχέση με τους υπόλοιπους αλγορίθμους σε αυτό το πρόβλημα. Από τους αλγόριθμους εκπαίδευσης ψευδό-Newton, μόνο ο αλγόριθμος LN ήταν ανταγωνιστικός με τους αλγόριθμους εκπαίδευσης συζυγών κλίσεων. Όσον αφορά την γενίκευση των αλγορίθμων εκπαίδευσης, ο αλγόριθμος SPR παρουσίασε τα καλύτερα μέσο ποσοστό γενίκευσης ακολουθούμενος από τον αλγόριθμο PR. Οι αλγόριθμοι εκπαίδευσης Perry και SPerry παρουσίασαν επίσης καλά αποτελέσματα γενίκευσης. Ο αλγόριθμος εκπαίδευσης OSS παρουσίασε την ίδια ικανότητα γενίκευσης με τους αλγορίθμους Perry και SPerry, αλλά συνέκλινε μόνο σε λίγες εξομοιώσεις με σημαντικά χαμηλότερη ταχύτητα σύγκλισης. Οι αλγόριθμοι εκπαίδευσης SHS και SHS παρουσίασαν επαρκή απόδοση γενίκευσης. Ο αλγόριθμος εκπαίδευσης SFR παρουσίασε καλύτερο ποσοστό μέσης γενίκευσης από το κλασικό του ανάλογο. Ωστόσο το ποσοστό γενίκευσης τους ήταν σημαντικά χαμηλότερο από αυτό των προαναφερθέντων αλγορίθμων εκπαίδευσης και είναι συγκρίσιμο μόνο με αυτό του αλγορίθμου LN. Όσον αφορά τους αλγόριθμους GB1 και GB2 η συνολική τους απόδοση ήταν χειρότερη συγκρινόμενη με αυτή των υπολοίπων αλγορίθμων.

4.5 Συμπεράσματα

Σε αυτό το κεφάλαιο παρουσιάστηκαν νέοι αλγόριθμοι εκπαίδευσης συζυγών κλίσεων. Εκτός από τους υπάρχοντες αλγόριθμους εκπαίδευσης HS, FR και PR, προτάθηκε η μέθοδος του Perry [60] για χρήση στην εκπαίδευση των τεχνητών νευρωνικών δικτύων. Στη συνέχεια, παράχθηκε μια νέα κατηγορία αλγορίθμων εκπαίδευσης συζυγών κλίσεων που ονομάζονται κλιμακωτοί αλγόριθμοι εκπαίδευσης συζυγών κλίσεων και βασίζονται στο φασματικό βήμα των Barzilai και Borwein [5]. Επιπλέον ενσωματώθηκε στους αλγόριθμους εκπαίδευσης συζυγών κλίσεων μια αποτελεσματική τεχνική γραμμικής αναζήτησης που βασίζεται στις συνθήκες του Wolfe και σε μια διασφαλισμένη κυβική παρεμβολή [77]. Επιπλέον, ο αρχικός ρυθμός εκπαίδευσης, που τροφοδοτεί την διαδικασία της γραμμικής αναζήτησης, προσαρμόζεται αυτόματα σε κάθε επανάληψη σύμφωνα με ένα κλειστό τύπο που προτάθηκε στις εργασίες [77, 80]. Τέλος, εφαρμόστηκε μια αποδοτική διαδικασία επανεκκίνησης η οποία βελτιώνει επιπλέον την αποτελεσματικότητα των αλγορίθμων εκπαίδευσης συζυγών κλίσεων.

Αναφέρθηκαν πειραματικά αποτελέσματα για 14 προβλήματα εκπαίδευσης. Τα πειράματα που διεξήχθησαν έδειξαν ότι οι προτεινόμενοι αλγόριθμοι εκπαίδευσης απέδωσαν καλά και ότι ήταν ανταγωνιστικοί με τους υπάρχοντες αλγόριθμους εκπαίδευσης. Στις περισσότερες περιπτώσεις παρουσίασαν καλύτερα ποσοστά επιτυχίας και ταχύτητα σύγκλισης. Η ικανότητα γενίκευσης του τεχνητού νευρωνικού δικτύου είναι ένα πολύ δύσκολο θέμα. Γενικά, η γενίκευση του δικτύου επηρεάζεται από πολλές παραμέτρους. Εξαρτάται από το μέγεθος του χώρου των βαρών (αρχιτεκτονική του δικτύου), το μέγεθος του συνόλου δεδομένων και ο ενδεχόμενος θόρυβος που μπορεί να περιέχουν, την αρχικοποίηση των βαρών και από τον αλγόριθμο εκπαίδευσης που χρησιμοποιείται κατά την διάρκεια της εκπαίδευσης. Είναι γνωστό ότι η γρήγορη εκπαίδευση μπορεί να οδηγήσει σε φτωχά αποτελέσματα γενίκευσης. Στα πειράματα που διεξήχθησαν, η αρχιτεκτονική του δικτύου, τα δεδομένα και η μέθοδος αρχικοποίησης των βαρών του δικτύου ήταν σταθερά για όλους τους αλγόριθμους εκπαίδευσης. Η μόνη διαφορά ήταν οι αλγόριθμοι εκπαίδευσης. Αποδείχθηκε μέσω των πειραματικών αποτελεσμάτων ότι παρόλο που οι προτεινόμενοι αλγόριθμοι εκπαίδευσης συγκλίνουν τις περισσότερες φορές ταχύτερα από τους υπόλοιπους αλγόριθμους εκπαίδευσης, παρουσιάζουν και καλά αποτελέσματα γενίκευσης. Αυτή η συμπεριφορά μπορεί να εξηγηθεί από το γεγονός ότι οι κλιμακωτοί αλγόριθμοι εκπαίδευσης συζυγών κλίσεων κάνουν καλή χρήση της διαθέσιμης πληροφορίας κατά την αναζήτηση του ελαχίστου στο χώρο των βαρών. Αποφεύγουν τις περισσότερες φορές τα τοπικά ελάχιστα και συγκλίνουν σε «καλά» σημεία λύσης.

Ο προτεινόμενος αλγόριθμος εκπαίδευσης SPerry ήταν ο καλύτερος αλγόριθμος. Παρουσίασε τα καλύτερα αποτελέσματα σχεδόν σε όλα τα προβλήματα. Επιπλέον, τα αποτελέσματα ήταν ισχυρά στατιστικά σημαντικά όσον αφορά την ταχύτητα σύγκλισης, αφού η p -τιμή που επέστρεψε ο έλεγχος Wilcoxon sign rank test ήταν κοντά στο 0 στις περισσότερες περιπτώσεις. Είναι ωστόσο απαραίτητο να ερευνηθεί περαιτέρω η απόδοση

των κλιμακωτών αλγορίθμων εκπαίδευσης συζυγών κλίσεων σε άλλα προβλήματα εκπαίδευσης έτσι ώστε να διερευνηθούν κατά το δυνατόν πληρέστερα τα πλεονεκτήματά τους και να αποκαλυφθούν τυχόν περιορισμοί. Παρόλα αυτά, είναι σημαντικό να αναφερθεί ότι οι κλιμακωτοί αλγόριθμοι εκπαίδευσης συζυγών κλίσεων αποτελούν βελτίωση έναντι των κλασικών αλγορίθμων εκπαίδευσης συζυγών κλίσεων και ότι είναι ανταγωνιστικοί με υπάρχοντες αλγόριθμους εκπαίδευσης της ίδιας κατηγορίας.

Υπάρχει μια μεγάλη δυνατότητα να βελτιωθεί περαιτέρω η απόδοση των προτεινόμενων κλιμακωτών αλγορίθμων συζυγών κλίσεων υιοθετώντας νέες ιδέες από την περιοχή της βελτιστοποίησης χωρίς περιορισμούς. Πιο συγκεκριμένα, νέες τεχνικές γραμμικής αναζήτησης, που βασίζονται στην μη μονότονη στρατηγική [29], μπορούν να χρησιμοποιηθούν με τους προτεινόμενους αλγόριθμους. Αυτή η στρατηγική χρησιμοποιήθηκε στην εργασία [64] για την βελτίωση της απόδοσης των αλγορίθμων εκπαίδευσης που βασίζονται στην κατεύθυνση της πιο απότομης καθόδου. Από την άλλη μεριά, ο έλεγχος της γωνίας που χρησιμοποιούν οι προτεινόμενοι αλγόριθμοι εκπαίδευσης ως συνθήκη επανεκκίνησης μπορεί να αντικατασταθεί με πιο ισχυρές διαδικασίες επανεκκίνησης [66]. Επίσης, η παράμετρος κλιμάκωσης που χρησιμοποιείται στους προτεινόμενους αλγόριθμους μπορεί να ενισχυθεί ή ακόμη και να αντικατασταθεί από άλλες επιλογές [81]. Τέλος, άλλοι αλγόριθμοι εκπαίδευσης που έχουν προταθεί για την εκπαίδευση τεχνητών νευρωνικών δικτύων (π.χ. [37]) μπορούν να συμπεριληφθούν μέσα στην οικογένεια των κλιμακωτών αλγορίθμων εκπαίδευσης συζυγών κλίσεων, με ενδεχόμενη καλύτερη απόδοση.

Μη Μονότονοι Αλγόριθμοι Εκπαίδευσης Συζυγών Κλίσεων

Στο προηγούμενο κεφάλαιο παρουσιάστηκε μια νέα κατηγορία αλγορίθμων εκπαίδευσης που βασίζονται στην κλιμακωτή κατεύθυνση συζυγών κλίσεων. Σε αυτό το κεφάλαιο απομονώνεται και τροποποιείται ο κλιμακωτός αλγόριθμος του Perry. Πιο συγκεκριμένα, διατηρούνται μερικά από τα χαρακτηριστικά του αλγορίθμου εκπαίδευσης, όπως είναι η παράμετρος κλιμάκωσης που βασίζεται στο φασματικό βήμα των Barzilai και Borwein [5], η συνθήκη επανεκκίνησης και ο αρχικός ρυθμός εκπαίδευσης, αλλά εφαρμόζεται μια διαφορετική τεχνική γραμμικής αναζήτησης η οποία βασίζεται στις μη μονότονες συνθήκες του Wolfe. Επίσης προτείνεται ένας νέος αρχικός ρυθμός εκπαίδευσης για χρήση με τον κλιμακωτό αλγόριθμο εκπαίδευσης συζυγών κλίσεων ο οποίος φαίνεται να είναι αποδοτικότερος από τον αρχικό ρυθμό εκπαίδευσης που προτάθηκε από τον Shanno [77] όταν χρησιμοποιείται σε συνδυασμό με την μη μονότονη τεχνική γραμμικής αναζήτησης. Στην συνέχεια παρουσιάζονται τα πειραματικά αποτελέσματα για διάφορα προβλήματα εκπαίδευσης. Τέλος, εκπαιδεύεται ένα πολυεπίπεδο εμπρόσθια τροφοδοτούμενο τεχνητό νευρωνικό δίκτυο με τον προτεινόμενο αλγόριθμο για το πρόβλημα της ταξινόμησης καρκινικών κυττάρων του εγκεφάλου και συγκρίνεται η απόδοσή του με αυτή ενός πιθανοτικού τεχνητού νευρωνικού δικτύου.

5.1 Ο Αλγόριθμος του Perry

Το πρόβλημα της εκπαίδευσης τεχνητών νευρωνικών δικτύων δίνεται από την παρακάτω σχέση

$$\min E(w), w \in \mathbb{R}^n \quad (5.1)$$

όπου η E είναι η συνάρτηση σφάλματος, που δίνεται από την εξίσωση (2.2). Ο αλγόριθμος της κλασικής μεθόδου συζυγών κλίσεων για την επίλυση του προβλήματος (5.1) είναι ένα επαναληπτικό σχήμα της μορφής

$$w_{k+1} = w_k - \alpha_k d_k \quad (5.2)$$

όπου d_k είναι η κατεύθυνση αναζήτησης και α_k είναι ο ρυθμός εκπαίδευσης. Ο ρυθμός εκπαίδευσης α_k μπορεί να καθοριστεί από μια τεχνική γραμμικής αναζήτησης έτσι ώστε η $E(w_k + \alpha_k d_k)$ να ελαχιστοποιείται κατά μήκος της κατεύθυνσης d_k , όταν τα w_k και d_k είναι σταθερά.

Ο κλασικός αλγόριθμος εκπαίδευσης συζυγών κλίσεων ξεκινά την διαδικασία της ελαχιστοποίησης με μια αρχική εκτίμηση του w_0 και μια αρχική κατεύθυνση

$$d_0 = -\nabla E(w_0) = -g_0$$

Κάθε κατεύθυνση d_{k+1} επιλέγεται έτσι ώστε να είναι γραμμικός συνδυασμός της κατεύθυνσης της πιο απότομης καθόδου $-g_{k+1}$ και της προηγούμενης κατεύθυνσης d_k . Επομένως έχουμε

$$d_{k+1} = -g_{k+1} + \beta_k d_k \quad (5.3)$$

όπου η βαθμωτή παράμετρος β_k είναι υπό καθορισμό απαιτώντας ότι η κατευθύνσεις d_k και d_{k+1} πρέπει να ικανοποιούν την συνθήκη της συζυγίας. Υπάρχουν πολλοί τύποι για την παράμετρο β_k . Ένας από αυτούς είναι ο τύπος που πρότεινε ο Perry [60] και δίνεται παρακάτω.

$$\beta_k = \frac{(y_k - s_k)^T g_{k+1}}{s_k^T y_k} \quad (5.4)$$

όπου

$$s_k = w_{k+1} - w_k \text{ και } y_k = g_{k+1} - g_k \quad (5.5)$$

5.2 Μη Μονότονος Κλιμακωτός Αλγόριθμος Εκπαίδευσης του Perry

Λαμβάνοντας υπόψιν την εργασία των Birgin και Martinez [9], υποτίθεται ότι μια πιο γενική μορφή της κατεύθυνσης αναζήτησης συζυγών κλίσεων δίνεται από την εξίσωση

$$d_{k+1} = -\vartheta_k g_{k+1} + \beta_k s_k \quad (5.6)$$

όπου $g_k = \nabla E(w_k)$ και $d_0 = -g_0$.

Ας υποθέσουμε ότι $E(w) \in \mathbb{C}^2$ και ότι ο Εσσιανός πίνακας $H \equiv \nabla^2 E(w)$ είναι θετικά ορισμένος. Αυτό υποδηλώνει ότι $y_k \neq 0$. Έτσι το σημείο ελαχίστου w_* ικανοποιεί την σχέση

$$w_* = w_{k+1} + d_*$$

όπου

$$Hd_* = -g_{k+1}$$

Πολλαπλασιάζοντας την προηγούμενη σχέση με το διάνυσμα s_k^T , έχουμε ότι

$$s_k^T Hd_* = -s_k^T g_{k+1}$$

Επομένως

$$y_k^T d_* = -s_k^T g_{k+1}$$

Έτσι το υπερεπίπεδο

$$\mathcal{H}_k \equiv \{d \in \mathbb{R}^n | y_k^T d = -s_k^T g_{k+1}\}$$

περιέχει την βέλτιστη προσάυξηση d_* , η οποία δίνει ότι $w_* = w_{k+1} + d_*$. Παρατηρούμε ότι η μηδενική κατεύθυνση $d = 0$ ανήκει στο υπερεπίπεδο $\mathcal{H}$ μόνο αν ισχύει ότι $s_k^T g_{k+1} = 0$ το οποίο είναι αντίθετο με την αρχική υπόθεσή μας.

Από τα παραπάνω, είναι φυσιολογικό να επιβάλλουμε για την κατεύθυνση αναζήτησης d_{k+1} ότι

$$d_{k+1} \in \mathcal{H}_k \tag{5.7}$$

Τότε, από την σχέση (5.6), έχουμε ότι

$$\beta_k = \frac{(\vartheta_k y_k - s_k)^T g_{k+1}}{s_k^T y_k} \tag{5.8}$$

Για $\vartheta_k = 1$ ο παραπάνω τύπος για την τιμή του β_k συμπίπτει με τον τύπο (5.4) που πρότεινε ο Perry στην εργασία [60].

Στόχος είναι να βρεθεί μια κατάλληλη επιλογή για την παράμετρο κλιμάκωσης ϑ_k έτσι ώστε να βελτιωθεί η απόδοση και η ταχύτητα σύγκλισης του κλασικού αλγόριθμου εκπαίδευσης του Perry. Μια καλή και αποδοτική επιλογή για την παράμετρο κλιμάκωσης είναι το φασματικό βήμα που προτάθηκε από τους Barzilai και Borwein [5], δηλαδή

$$\vartheta_k = \frac{s_k^T s_k}{s_k^T y_k} \tag{5.9}$$

Οι ιδιότητες και οι λόγοι για την επιλογή του βήματος των Barzilai και Borwein αναφέρονται στα προηγούμενα κεφάλαια και γι' αυτό τον λόγο δεν αναφέρονται εδώ.

Όπως αναφέρθηκε προηγουμένως, οι αλγόριθμοι συζυγών κλίσεων για να έχουν την ιδιότητα της ολικής σύγκλισης πρέπει η παράμετρος του ρυθμού εκπαίδευσης να καθορίζεται από μια τεχνική μονοδιάστατης γραμμικής αναζήτησης κατά μήκος της κατεύθυνσης συζυγών κλίσεων d_k . Στον αλγόριθμο που περιγράφεται, ο ρυθμός εκπαίδευσης που λαμβάνεται από την τεχνική της γραμμικής αναζήτησης πρέπει να ικανοποιεί τις μη μονότονες συνθήκες του Wolfe που δίνονται από τις παρακάτω εξισώσεις

$$E(w_k + \alpha_k d_k) - \max_{0 \leq j \leq M} E(w_{k-j}) \leq c_1 \alpha_k \nabla E(w_k)^T d_k \quad (5.10)$$

$$\nabla E(w_k + \alpha_k d_k)^T d_k \geq c_2 \nabla E(w_k)^T d_k \quad (5.11)$$

όπου $0 < c_1 \leq c_2 < 1$ και M είναι ένας θετικός ακέραιος αριθμός, μη μονότονος ορίζοντας εκπαίδευσης. Η πρώτη συνθήκη (5.10) επιτρέπει σε οποιοδήποτε σημείο να γίνει αποδεκτό αν βελτιώνει ικανοποιητικά την συναρτησιακή τιμή αναφορικά με τη μεγαλύτερη από τις $M + 1$ (ή k αν $k \leq M$) πιο πρόσφατες συναρτησιακές τιμές. Η δεύτερη συνθήκη (5.11) διασφαλίζει ότι ο παρονομαστής της φασματικής κλιμακωτής παραμέτρου ότι είναι πάντα καλά ορισμένος και θετικός, αφού υποδηλώνει ότι $s_k^T y_k > 0$. Και οι δυο συνθήκες επιτρέπουν μια αύξηση στην τιμή της συνάρτησης σφάλματος χωρίς να επηρεάζει την ιδιότητα της ολικής σύγκλισης όπως έχει αποδειχθεί στην εργασία [30].

Μια απλή τεχνική γραμμικής αναζήτησης που χρησιμοποιείται για την ρύθμιση της παραμέτρου του ρυθμού εκπαίδευσης α_k είναι η μείωση του ρυθμού εκπαίδευσης κατά έναν παράγοντα $1/q$, όπου $q > 1$, έτσι ώστε να ικανοποιούνται οι συνθήκες (5.10) και (5.11). Η επιλογή του q δεν είναι σημαντική για την επιτυχή εκπαίδευση, αλλά έχει άμεση επίδραση στον αριθμό των υπολογισμών της συνάρτησης σφάλματος και της κλίσεως της. Μικρές τιμές της παραμέτρου q μπορεί να οδηγήσουν σε πολλές επαναλήψεις της διαδικασίας της γραμμικής αναζήτησης ενώ μεγάλες τιμές μπορεί να οδηγήσουν σε αύξηση των επαναλήψεων του αλγορίθμου γενικά. Στην βιβλιογραφία προτείνεται συνήθως η τιμή $q = 2$, η οποία λειτουργεί τις περισσότερες φορές αποτελεσματικά και αποδοτικά. Αυτή η στρατηγική της οπίσθιας ιχνηλάτησης διασφαλίζει ότι ο ρυθμός εκπαίδευσης ελαττώνεται έτσι ώστε να ικανοποιείται η συνθήκη (5.10). Πιο συγκεκριμένα, η ελάττωση του ρυθμού εκπαίδευσης δίνεται από τον τύπο $\tilde{\alpha}_k = 2^{-r} \alpha_k$, όπου $\tilde{\alpha}_k$ είναι ο ρυθμός εκπαίδευσης μετά από r υποδιαίρεσεις που απαιτούνται για να ικανοποιείται η συνθήκη (5.10). Επιπλέον, για να αποφευχθεί η άσκοπη ελάττωση της παραμέτρου του ρυθμού εκπαίδευσης, επιβάλλεται ότι ο ρυθμός εκπαίδευσης πρέπει να ικανοποιεί την συνθήκη (5.11).

Μια δημοφιλής επιλογή του αρχικού ρυθμού εκπαίδευσης που αποδεδειγμένα συνεργάζεται επιτυχώς με της μεθόδους συζυγών κλίσεων έχει προταθεί από τον Shanno στον αλγόριθμο CONMIN [77]. Ο αρχικός ρυθμός εκπαίδευσης του Shanno δίνεται

από τον τύπο

$$\alpha_k = \begin{cases} \frac{1}{\|g_0\|}, & \text{αν } k = 0; \\ \frac{\alpha_{k-1}\|d_k\|}{\|d_{k-1}\|}, & \text{διαφορετικά.} \end{cases} \quad (5.12)$$

όπου d_k είναι η κατεύθυνση αναζήτησης συζυγών κλίσεων, d_{k-1} είναι η προηγούμενη κατεύθυνση και α_{k-1} είναι ο προηγούμενος ρυθμός εκπαίδευσης. Ο αρχικός ρυθμός μάθησης α_0 είναι $1/\|g_0\|$ όπου g_0 είναι η αρχική κατεύθυνση της πιο απότομης καθόδου.

Μια άλλη αποτελεσματική επιλογή του ρυθμού εκπαίδευσης, που προτάθηκε στην εργασία [41], δίνεται από τον παρακάτω τύπο

$$\alpha_k = \begin{cases} \frac{1}{\|g_0\|}, & \text{αν } k = 0; \\ \frac{2^{-(r+1)}\|d_{k-1}\|}{\|d_k - d_{k-1}\|}, & \text{διαφορετικά.} \end{cases} \quad (5.13)$$

όπου r είναι ο αριθμός των υποδιαιρέσεων που χρειάζεται ο προηγούμενος ρυθμός εκπαίδευσης ώστε να γίνει αποδεκτός από τις μη μονότονες συνθήκες του Wolfe (5.10) και (5.11).

Η κατεύθυνση αναζήτησης συζυγών κλίσεων d_{k+1} που δίνεται από τον τύπο (5.6) μερικές φορές αποτυγχάνει να είναι κατεύθυνση καθόδου. Για να αποφευχθούν αυτές οι παθολογικές καταστάσεις ο αλγόριθμος εκπαίδευσης επανεκκινείται. Πολλές τεχνικές επανεκκίνησης έχουν προταθεί στην βιβλιογραφία. Στον προτεινόμενο αλγόριθμο εκπαίδευσης, ο αλγόριθμος επανεκκινείται με την φασματική κατεύθυνση της πιο απότομης καθόδου που δίνεται από τον παρακάτω τύπο

$$d_{k+1} = -\eta_k g_{k+1} \quad (5.14)$$

Το κριτήριο που χρησιμοποιείται για να αποφασιστεί αν η κατεύθυνση d_{k+1} είναι καθοδική, δίνεται από τον ακόλουθο τύπο

$$d_{k+1}^T g_{k+1} \leq -10^{-3} \|d_{k+1}\| \|g_{k+1}\| \quad (5.15)$$

Πιο συγκεκριμένα, αν ισχύει η σχέση (5.15), γίνεται αποδεκτή η κατεύθυνση αναζήτησης που δίνεται από την σχέση (5.6), διαφορετικά η γωνία μεταξύ της κατεύθυνσης d_{k+1} και της κλίσης g_{k+1} δεν είναι αρκετά οξεία, οπότε επανεκκινείται ο αλγόριθμος με την κατεύθυνση που δίνεται από την σχέση (5.14).

Οι μη μονότονες συνθήκες του Wolfe (5.10) και (5.11) μαζί με την συνθήκη επανεκκίνησης (5.15) είναι επαρκείς για να αποδείξουν την ολική σύγκλιση του αλγορίθμου κάτω από λογικές υποθέσεις. Αν η κλίση της συνάρτησης σφάλματος E , g_k , είναι Lipschitz συνεχής και η E είναι κάτω φραγμένη μπορεί να αποδειχθεί ότι

$$\liminf_{k \rightarrow \infty} \|g_k\| = 0$$

Αυτό υποδηλώνει ότι υπάρχει μια υπακολουθία των κλίσεων της συνάρτησης σφάλματος E που συγκλίνει στο 0 [29, 56].

Σε αυτό το σημείο δίνεται μια λεπτομερής διατύπωση του μη μονότονου κλιμακωτού αλγορίθμου εκπαίδευσης του Perry. Σημειώνεται ότι για $\vartheta_k = 1$, ο αλγόριθμος συμπίπτει με τον κλασικό αλγόριθμο εκπαίδευσης συζυγών κλίσεων του Perry. Ο αλγόριθμος εκπαίδευσης δέχεται ως είσοδο το διάνυσμα των αρχικών βαρών w_0 , το κριτήριο τερματισμού ϵ , τον μέγιστο αριθμό επαναλήψεων μ , τις παραμέτρους c_1 , c_2 και M που χρειάζονται για τον αλγόριθμο γραμμικής αναζήτησης και επιστρέφει το σημείο ελαχίστου w_* και το ελάχιστο της συνάρτησης σφάλματος $E(w_*)$.

Αλγόριθμος 6 Ο κλιμακωτός αλγόριθμος εκπαίδευσης συζυγών κλίσεων του Perry.

NMSP($w_0, \epsilon, \mu, c_1, c_2, M, w_*, E(w_*)$)

- 1: **Για** $k := 0$ **έως** μ **Κάνε**
 - 2: $E_k = E(w_k)$, $g_k = \nabla E(w_k)$.
 - 3: **Αν** $E_k \leq \epsilon$ **Τότε**
 - 4: **Επέστρεψε** $w_* = w_k$, $E(w_*) = E(w_k)$.
 - 5: **Τέλος**
 - 6: **Αν** $k := 0$ **Τότε**
 - 7: $d_k = -g_k$.
 - 8: $\alpha_k = 1/\|g_k\|$.
 - 9: **Αλλιώς**
 - 10: $\vartheta_k = \frac{s_k^T s_k}{s_k^T y_k}$.
 - 11: Υπολόγισε το $\beta_k = \frac{(\vartheta_k y_k - s_k)^T g_k}{s_k^T y_k}$.
 - 12: $d = -\vartheta_k g_k + \beta_k s_k$.
 - 13: **Αν** $d^T g_k \leq -10^{-3} \|d_k\| \|g_k\|$ **Τότε**
 - 14: $d_k = d$.
 - 15: **Αλλιώς**
 - 16: $d_k = -\vartheta_k g_k$.
 - 17: **Τέλος**
 - 18: Υπολόγισε τον ρυθμό εκπαίδευσης α_k χρησιμοποιώντας τον τύπο (5.12) ή τον τύπο (5.13).
 - 19: **Τέλος**
 - 20: LineSearch. Υπολόγισε ένα ρυθμό εκπαίδευσης α_k που να ικανοποιεί τις μη μονότονες συνθήκες του Wolfe (5.10) και (5.11).
 - 21: $w_{k+1} = w_k + \alpha_k d_k$.
 - 22: **Τέλος**
 - 23: **Επέστρεψε** Το δίκτυο δεν εκπαιδεύτηκε.
-

Παρατήρηση 5.1. Όταν ο παραπάνω αλγόριθμος εκπαίδευσης χρησιμοποιεί τον ρυθμό εκπαίδευσης που δίνεται από τον τύπο (5.12), τότε ονομάζεται NMSP1, ενώ όταν χρησιμοποιεί τον ρυθμό εκπαίδευσης που δίνεται από τον τύπο (5.13), ονομάζεται NMSP2.

5.3 Πειραματικά Αποτελέσματα

Στην παράγραφο αυτή αξιολογείται η επίδοση των προτεινόμενων αλγορίθμων εκπαίδευσης τεχνητών νευρωνικών δικτύων NMSP1 [80] και NMSP2 [41] και συγκρίνονται με την επίδοση των ακόλουθων γνωστών αλγορίθμων εκπαίδευσης:

1. Τον αλγόριθμο της οπισθοδρομικής διάδοσης του σφάλματος με σταθερό ρυθμό εκπαίδευσης (BP) [74].
2. Τον αλγόριθμο της οπισθοδρομικής διάδοσης του σφάλματος με σταθερό ρυθμό εκπαίδευσης και ορμή (BPM) [74].
3. Τον αλγόριθμο της οπισθοδρομικής διάδοσης του σφάλματος με προσαρμοστικό ρυθμό εκπαίδευσης και ορμή (ABP) [84].

Για τους αλγόριθμους εκπαίδευσης NMSP1 και NMSP2 οι παράμετροι M , c_1 και c_2 έχουν τις τιμές 10, 10^{-4} και 0.5, αντίστοιχα. Οι τιμές των παραπάνω παραμέτρων είναι οι ίδιες για όλα τα πειράματα που πραγματοποιήσαμε. Και οι δύο αλγόριθμοι χρησιμοποιούν την ίδια μη μονότονη στρατηγική γραμμικής αναζήτησης. Επιπλέον, για την αρχικοποίηση των βαρών και των μεροληψιών εφαρμόστηκε η τεχνική που προτάθηκε από τους Nguyen και Widrow [55] για όλους τους αλγόριθμους.

Τα προβλήματα που χρησιμοποιήθηκαν για την αξιολόγηση της ταχύτητας σύγκλισης των προτεινόμενων αλγορίθμων εκπαίδευσης τεχνητών νευρωνικών δικτύων NMSP1 και NMSP2 σε σχέση με τους προαναφερθέντες αλγόριθμους εκπαίδευσης είναι τα ακόλουθα:

1. Το πρόβλημα του Αποκλειστικού-EITE.
2. Το πρόβλημα της ισοτιμίας των 3-bit.
3. Το πρόβλημα της αναγνώρισης των κεφαλαίων γραμμάτων της Αγγλικής αλφαβήτου.
4. Το πρόβλημα της προσέγγισης της συνεχούς τριγωνομετρικής συνάρτησης.

Για κάθε πρόβλημα, παρατίθεται ένας πίνακας ο οποίος συνοψίζει την απόδοση των αλγορίθμων για προσομοιώσεις που έφτασαν σε λύση εντός ενός προκαθορισμένου ορίου υπολογισμών της συνάρτησης σφάλματος που ορίζεται σε κάθε πρόβλημα. Οι αναφερόμενες στους πίνακες παράμετροι είναι: Min ο ελάχιστος αριθμός επαναλήψεων, Max ο μέγιστος αριθμός επαναλήψεων, μ η μέση τιμή των επαναλήψεων, σ η τυπική απόκλιση, και $Επιτυχία$ ο αριθμός των πετυχημένων προσομοιώσεων από ένα σύνολο 1000 δοκιμών. Σε περίπτωση που ένας αλγόριθμος αποτύχει να συγκλίνει μέσα στο

προκαθορισμένο όριο των υπολογισμών της συνάρτησης σφάλματος, θεωρείται ότι απέτυχε να εκπαιδεύσει το τεχνητό νευρωνικό δίκτυο και οι επαναλήψεις της συγκεκριμένης προσομοίωσης δεν περιλαμβάνονται στην στατιστική ανάλυση του αλγορίθμου.

Σε αυτό το σημείο, πρέπει να επισημανθεί ότι γενικά μια επανάληψη ενός αλγορίθμου εκπαίδευσης σημαίνει ότι όλα τα πρότυπα εκπαίδευσης έχουν παρουσιαστεί στο τεχνητό νευρωνικό δίκτυο. Στους κλασικούς αλγορίθμους, όπως η BP, κάθε επανάληψη αντιστοιχεί σε ένα συναρτησιακό υπολογισμό της συνάρτησης σφάλματος και έναν υπολογισμό της κλίσης (παράγωγοι πρώτης τάξεως). Στους προτεινόμενους αλγορίθμους εκπαίδευσης συζυγών κλίσεων λαμβάνουν χώρα περισσότεροι συναρτησιακοί υπολογισμοί και υπολογισμοί κλίσεων ανά επανάληψη, γεγονός που οφείλεται στην παρουσία της τεχνικής της μη μονότονης γραμμικής αναζήτησης. Εξαιτίας της φύσεως της γραμμικής αναζήτησης που χρησιμοποιείται στους αλγορίθμους αυτούς, απαιτείται ο ίδιος αριθμός συναρτησιακών υπολογισμών και υπολογισμών κλίσεων ανά επανάληψη. Δηλαδή μετά την εκτέλεση της διαδικασίας της γραμμικής αναζήτησης σε μια επανάληψη του αλγορίθμου εκπαίδευσης ο ίδιος αριθμός συναρτησιακών υπολογισμών και υπολογισμών κλίσεων λαμβάνει χώρα.

Επιπλέον αξιολογείται η γενίκευση του αλγορίθμου εκπαίδευσης NMSP2 χρησιμοποιώντας ένα πραγματικό πρόβλημα ταξινόμησης καρκινικών κυττάρων εγκεφάλου από δεδομένα που προέρχονται από το Τμήμα Παθολογίας του Πανεπιστημιακού Νοσοκομείου Πατρών και από το Τμήμα Παθολογίας του Γενικού Αντικαρκινικού Νοσοκομείου Πειραιά Μεταξάς. Τα αποτελέσματα που επιτυγχάνονται με την εκπαίδευση του πολυεπίπεδου εμπρόσθια τροφοδοτούμενου τεχνητού νευρωνικού δικτύου με τον προτεινόμενο αλγόριθμο συγκρίνονται με αυτά που επιτυγχάνονται από την εκπαίδευση ενός πιθανοτικού τεχνητού νευρωνικού δικτύου [42].

5.3.1 Αποκλειστικό-ΕΙΤΕ

Για το πρόβλημα του αποκλειστικού-ΕΙΤΕ (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα $2 - 2 - 1$ τεχνητό νευρωνικό δίκτυο (6 βάρη και 3 μεροληψίες). Το κρυφό επίπεδο του νευρωνικού δικτύου είναι βασισμένο σε νευρώνες με συνάρτηση ενεργοποίησης την υπερβολική εφαιπτομένη, ενώ το επίπεδο εξόδου είναι βασισμένο σε νευρώνες με συνάρτηση ενεργοποίησης την γραμμική συνάρτηση. Η συνθήκη τερματισμού είναι $E \leq 0.01$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 1000. Για τις μεθόδους σταθερού ρυθμού εκπαίδευσης (BP και BPM) ο ρυθμός εκπαίδευσης ορίστηκε στην τιμή 0.1, αντί της προεπιλεγμένης τιμής 0.01, έτσι ώστε να επιταχυνθεί η σύγκλιση τους, αφού με την προεπιλεγμένη τιμή συγκλίνουν πολύ αργά σε αυτό το πρόβλημα. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 5.1.

Ο αλγόριθμος BPM απέδωσε καλύτερα όσον αφορά το ποσοστό επιτυχίας σε σχέση με τον αλγόριθμο BP, αλλά χρειάστηκε περισσότερες επαναλήψεις. Ο αλγόριθμος ABP, είχε το ίδιο ποσοστό επιτυχίας με τον αλγόριθμο BP αλλά είχε σημαντικά μικρότερο

Αλγόριθμος	min	max	μ	σ	Επιτυχία
BP	28	992	94.5329	114.441	74.5%
BPM	23	905	111.821	136.753	76.1%
ABP	18	835	46.5517	73.1005	74.5%
NMSP1	28	991	133.136	170.856	87.3%
NMSP2	16	993	115.201	173.146	90.9%

Πίνακας 5.1: Συγκριτικά αποτελέσματα για το πρόβλημα του αποκλειστικού-EITE

μέσο αριθμό επαναλήψεων. Ο αλγόριθμος NMSP1 απέδωσε αρκετά καλά έχοντας πολύ καλύτερο ποσοστό επιτυχίας σε σχέση με τους κλασικούς αλγορίθμους εκπαίδευσης αλλά ήταν δαπανηρότερος. Ο αλγόριθμος εκπαίδευσης NMSP2 παρουσιάστηκε πιο βελτιωμένος από τον αλγόριθμο NMSP1 τόσο σε ποσοστό επιτυχίας όσο και σε ταχύτητα σύγκλισης.

5.3.2 Ισοτιμία των 3-bit

Για το πρόβλημα της ισοτιμίας των 3-bit (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα 3 – 2 – 1 τεχνητό νευρωνικό δίκτυο (8 βάρη και 3 μεροληψίες). Το κρυφό επίπεδο του νευρωνικού δικτύου είναι βασισμένο σε νευρώνες με συνάρτηση ενεργοποίησης την υπερβολική εφαπτομένη, ενώ το επίπεδο εξόδου είναι βασισμένο σε νευρώνες με συνάρτηση ενεργοποίησης την γραμμική συνάρτηση. Η συνθήκη τερματισμού είναι $E \leq 0.01$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 1000. Για τις μεθόδους σταθερού ρυθμού εκπαίδευσης (BP και BPM) ο ρυθμός εκπαίδευσης ορίστηκε στην τιμή 0.1, αντί της προεπιλεγμένης τιμής 0.01, έτσι ώστε να επιταχυνθεί η σύγκλιση τους, αφού με την προεπιλεγμένη τιμή συγκλίνουν πολύ αργά σε αυτό το πρόβλημα. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 5.2.

Αλγόριθμος	min	max	μ	σ	Επιτυχία
BP	-	-	-	-	0.0%
BPM	185	996	492.716	197.034	53.5%
ABP	431	982	596.344	128.165	47.4%
NMSP1	171	986	429.5	175.026	73.8%
NMSP2	112	992	323.792	174.252	79.4%

Πίνακας 5.2: Συγκριτικά αποτελέσματα για το πρόβλημα της ισοτιμίας των 3-bit

Ο αλγόριθμος εκπαίδευσης BP απέτυχε να συγκλίνει μέσα στο όριο των υπολογισμών των τιμών της συνάρτησης σφάλματος σε όλες τις προσομοιώσεις. Αξίζει να αναφερθεί ότι δοκιμάσαμε διαφορετικές τιμές για την τιμή της παραμέτρου του ρυθμού εκπαίδευσης, αλλά το αποτέλεσμα ήταν το ίδιο. Ο αλγόριθμος BPM απέδωσε αρκετά καλά, ενώ ήταν καλύτερος τόσο σε ποσοστό επιτυχίας όσο και σε υπολογισμούς τιμών

της συνάρτησης σφάλματος (και υπολογισμούς κλίσεως), αναφορικά με τον αλγόριθμο εκπαίδευσης ABP. Ο αλγόριθμος NMSP1 απέδωσε εξαιρετικά έχοντας πολύ καλύτερο ποσοστό επιτυχίας σε σχέση με τους κλασικούς αλγορίθμους εκπαίδευσης ενώ ήταν ταχύτερος από τον αλγόριθμο BPM που παρουσίασε την ταχύτερη σύγκλιση από τους κλασικούς αλγορίθμους. Ο αλγόριθμος εκπαίδευσης NMSP2 παρουσιάστηκε πιο βελτιωμένος από τον αλγόριθμο NMSP1 τόσο σε ποσοστό επιτυχίας όσο και σε ταχύτητα σύγκλισης μειώνοντας δραστικά τον υπολογισμό των τιμών της συνάρτησης σφάλματος και της κλίσης της.

5.3.3 Αναγνώριση Κεφαλαίων Γραμμάτων

Για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων της Αγγλικής αλφαβήτου (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα 35 – 30 – 26 τεχνητό νευρωνικό δίκτυο (1830 βάρη και 56 μεροληψίες). Το κρυφό επίπεδο του νευρωνικού δικτύου είναι βασισμένο σε νευρώνες με λογιστική συνάρτηση ενεργοποίησης, ενώ το επίπεδο εξόδου είναι βασισμένο σε νευρώνες γραμμικής συνάρτησης ενεργοποίησης. Η συνθήκη τερματισμού είναι $E \leq 0.1$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 2000. Για τις μεθόδους σταθερού ρυθμού εκπαίδευσης (BP και BPM) ο ρυθμός εκπαίδευσης ορίστηκε στην προεπιλεγμένη τιμή 0.01. Η αρχικοποίηση των βαρών έγινε με την τεχνική των Nguyen και Widrow, αλλά στη συνέχεια τα βάρη μεταξύ του κρυφού επιπέδου και του επιπέδου της εξόδου πολλαπλασιάστηκαν με 0.01. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 5.3.

Αλγόριθμος	min	max	μ	σ	Επιτυχία
BP	1100	1999	1548.22	197.517	75.6%
BPM	1246	1995	1579.5	175.82	4.8%
ABP	1234	1999	1785.87	156.965	36.1%
NMSP1	406	1210	707.732	131.425	100.0%
NMSP2	288	801	486.909	84.5928	100.0%

Πίνακας 5.3: Συγκριτικά αποτελέσματα για το πρόβλημα της αναγνώρισης κεφαλαίων γραμμάτων

Από τους κλασικούς αλγορίθμους εκπαίδευσης, ο BP παρουσιάζει την καλύτερη επίδοση τόσο σε ποσοστό επιτυχίας όσο και σε υπολογισμούς της τιμής της συνάρτησης σφάλματος. Ειδικά η BPM παρουσιάζει πολύ μικρό ποσοστό επιτυχίας. Ο αλγόριθμος NMSP1 παρουσιάζει εξαιρετικό ποσοστό επιτυχίας, αφού συγκλίνει σε όλες τις προσομοιώσεις. Επιπλέον η απόδοση του σε συναρτησιακούς υπολογισμούς και υπολογισμούς κλίσεως είναι εξαιρετική. Ο αλγόριθμος NMSP2, επίσης συγκλίνει σε όλες τις προσομοιώσεις, αλλά παρουσιάζει καλύτερη επίδοση όσον αφορά την ταχύτητα σύγκλισης. Αξίζει να σημειωθεί ότι ο μέγιστος αριθμός συναρτησιακών υπολογισμών που χρειάστηκε ο αλγόριθμος εκπαίδευσης NMSP2 είναι μικρότερος από 1000 το οποίο σημαίνει ότι για αυτόν τον αλγόριθμο είναι δυνατό να μειώσουμε το όριο των συναρτησιακών

υπολογισμών χωρίς να επηρεαστεί το ποσοστό επιτυχίας. Σε αυτή την περίπτωση όμως, κανένας από τους κλασικούς αλγόριθμους δεν θα συνέκλινε, αφού ο ελάχιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι μεγαλύτερος από 1000 ενώ ο αλγόριθμος NMSP1 θα είχε μικρότερο ποσοστό επιτυχίας.

5.3.4 Προσέγγιση Συνεχούς Τριγωνομετρικής Συνάρτησης

Για το πρόβλημα της προσέγγισης μιας συνεχούς τριγωνομετρικής συνάρτησης (βλέπε Παράρτημα Α') χρησιμοποιήσαμε ένα $1 - 10 - 1$ τεχνητό νευρωνικό δίκτυο (20 βάρη και 11 μεροληψίες). Το κρυφό επίπεδο του νευρωνικού δικτύου είναι βασισμένο σε νευρώνες με λογιστική συνάρτηση ενεργοποίησης, ενώ το επίπεδο εξόδου είναι βασισμένο σε νευρώνες γραμμικής συνάρτησης ενεργοποίησης. Η συνθήκη τερματισμού είναι $E \leq 0.1$ και ο μέγιστος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος είναι 1000. Για τις μεθόδους σταθερού ρυθμού εκπαίδευσης (BP και BPM) ο ρυθμός εκπαίδευσης ορίστηκε στην προεπιλεγμένη τιμή 0.01. Τα αποτελέσματα των εξομοιώσεων παρουσιάζονται στον Πίνακα 5.4.

Αλγόριθμος	min	max	μ	σ	Επιτυχία
BP	398	994	767.359	154.092	7.8%
BPM	389	995	760.506	157.334	7.7%
ABP	116	999	610.51	217.033	29.0%
NMSP1	112	998	486.212	179.492	70.4%
NMSP2	84	983	343.841	168.9	76.9%

Πίνακας 5.4: Συγκριτικά αποτελέσματα για το πρόβλημα της προσέγγισης συνεχούς τριγωνομετρικής συνάρτησης.

Σε αυτό το πρόβλημα, ο αλγόριθμος εκπαίδευσης ABP είχε την καλύτερη επίδοση μεταξύ των κλασικών αλγορίθμων εκπαίδευσης σε ποσοστό επιτυχίας και σε μέσο όρο υπολογισμών των τιμών της συνάρτησης σφάλματος. Ωστόσο το ποσοστό επιτυχίας του ήταν χαμηλό σε σχέση με το ποσοστό επιτυχίας των αλγορίθμων NMSP1 και NMSP2. Ο αλγόριθμος NMSP1 απέδωσε πολύ καλά όσον αφορά το ποσοστό επιτυχίας αφού η διαφορά του σε ποσοστό επιτυχίας από τον αλγόριθμο ABP ήταν πάνω από 40%. Επιπλέον, είχε πολύ μικρότερο κόστος σε συναρτησιακούς υπολογισμούς και υπολογισμούς κλίσεως. Ο αλγόριθμος εκπαίδευσης NMSP2 παρουσίασε το καλύτερο ποσοστό επιτυχίας στο παράδειγμα αυτό. Επίσης, παρουσίασε μικρότερο κόστος σε συναρτησιακούς υπολογισμούς κατά μέσο όρο από τον NMSP1 ενώ ο μέσος όρος των υπολογισμών κλίσεως ήταν αρκετά μικρότερος.

5.3.5 Ταξινόμηση Καρκινικών Κυττάρων Εγκεφάλου

Η ταξινόμηση των όγκων εγκεφάλου είναι ένα κρίσιμο βήμα για τον σχεδιασμό της αγωγής και της αντιμετώπισης της ασθένειας. Τα αστροκυτώματα του εγκεφάλου θεωρείται από τις πιο επιθετικές και δύσκολα ιάσιμες μορφές καρκίνου. Για να αποφασιστεί ο βαθμός της καρκινικής ανωμαλίας, παθολόγοι εξετάζουν τις βιοψίες με μικροσκόπιο. Σύμφωνα με το σύστημα του ΠΟΥ (Παγκόσμιο Οργανισμό Υγείας), κλινικοί ιατροί ταξινομούν τα αστροκυτώματα σε δυο κατηγορίες, χαμηλού και υψηλού βαθμού νεοπλασμάτων. Ωστόσο, η διαδικασία εξέτασης έχει αποδειχθεί ότι είναι υποκειμενική και ότι εξαρτάται σε μεγάλο βαθμό από την εμπειρία και τις ικανότητες του ειδικού. Προκειμένου να διευκολυνθούν οι παθολόγοι ώστε να μπορούν να πάρουν πιο αντικειμενικές αποφάσεις, εισηχθήσαν υπολογιστικά διαγνωστικά συστήματα, τα οποία όμως συνεπικουρούν τους ειδικούς και δεν τους αντικαθιστούν.

Πολλά υπολογιστικά συστήματα ταξινόμησης τα οποία βασίζονται στα δέντρα αποφάσεων, τα εμπρόσθια τροφοδοτούμενα νευρωνικά δίκτυα, την διακριτική ανάλυση, την ασαφή λογική, κλπ έχουν προταθεί στη βιβλιογραφία. Ωστόσο, τα συστήματα αυτά είναι πολύ δύσκολο να ενσωματωθούν στην καθημερινή κλινική διάγνωση εάν δεν χρησιμοποιούν κάποια καθιερωμένα πρότυπα όπως το σχήμα ταξινόμησης του ΠΟΥ και την διαδικασία χρώσης Αιματοξυλίνης - Εοσίνης (Hematoxylin-Eosin *H&E*) [67].

Στο πραγματικό αυτό πρόβλημα εκπαίδευσης, προετοιμάστηκαν 140 βιοψίες ακολουθώντας το πρωτόκολλο της χρώσης Αιματοξυλίνης - Εοσίνης. Οι βιοψίες προέρχονται από το Τμήμα Παθολογίας του Πανεπιστημιακού Νοσοκομείου Πατρών και από το Τμήμα Παθολογίας του Γενικού Αντικαρκινικού Νοσοκομείου Πειραιά Μεταξάς. Η ταξινόμηση πραγματοποιήθηκε από δύο παθολόγους, που ταξινόμησαν τους καρκινικούς όγκους ως χαμηλού και υψηλού βαθμού σύμφωνα με τον σύστημα του ΠΟΥ. 61 περιπτώσεις χαρακτηρίστηκαν ως χαμηλού βαθμού και 79 ως υψηλού βαθμού. Χρησιμοποιώντας ένα ειδικό φωτογραφικό μικροσκόπιο, που περιγράφεται στην εργασία [28], πάρθηκαν φωτογραφίες από περιοχές προκαθορισμένες από ειδικούς (Εικόνα 5.1). Στη συνέχεια εφαρμόστηκε ένας αλγόριθμος τμηματοποίησης ο οποίος διαχωρίζει τον πυρήνα από τον περιβάλλον ιστό έτσι ώστε να κωδικοποιηθεί ο κακοήθης όγκος σε ένα σύνολο από 40 ποσοτικά χαρακτηριστικά του πυρήνα. Τα χαρακτηριστικά του πυρήνα έχει αποδειχθεί ότι είναι ισχυροί περιγραφείς του βαθμού της καρκινικής ανωμαλίας [19].

Τα χαρακτηριστικά που εξήχθησαν μπορούν να κατηγοριοποιηθούν σε ένα σύνολο μορφολογικών χαρακτηριστικών που περιγράφουν το μέγεθος και το σχήμα του πυρήνα και σε ένα σύνολο από χαρακτηριστικά υφής που κωδικοποιούν την κατανομή και οργάνωση της χρωματίνης μέσα στον πυρήνα. Τα μορφολογικά χαρακτηριστικά περιλαμβάνουν μετρήσεις του εμβαδού, της στρογγυλότητας και της κοιλότητας. Για κάθε ένα από αυτά τα χαρακτηριστικά υπολογίστηκε η μέση τιμή, η τυπική απόκλιση, το εύρος, η λοξότητα, η κύρτωση και η μέγιστη τιμή. Τα χαρακτηριστικά υφής υπολογίστηκαν από το ιστόγραμμα του DNA, την συνύπαρξη και τους πίνακες διάρκειας εκτέλεσης. Μια πιο λεπτομερής περιγραφή του υπολογισμού αυτών των χαρακτηριστικών δίνεται στις

Σχήμα 5.1: Παραδείγματα χαμηλού (επάνω) και υψηλού βαθμού (κάτω) εγκεφαλικών αστροκυτωμάτων μαζί με την τμηματοποιημένη εικόνα.

εργασίες [82, 83].

Στην περίπτωση μας η διάσταση των διανυσμάτων των χαρακτηριστικών είναι πολύ μεγάλη (40 χαρακτηριστικά). Ωστόσο, οι συνιστώσες του διανύσματος είναι υψηλά συσχετισμένες, αφού κάποια χαρακτηριστικά όπως το εμβαδόν του πυρήνα και η περίμετρος του είναι ισχυρά συσχετισμένα. Επομένως είναι χρήσιμο να ελαττώσουμε τη διάσταση του διανύσματος χαρακτηριστικών αφού έτσι θα μειωθεί σημαντικά και ο υπολογιστικός χρόνος. Μια αποτελεσματική διαδικασία που πραγματοποιεί αυτήν τη διεργασία είναι η Ανάλυση Κύριων Συνιστωσών (PCA). Αυτή η διαδικασία έχει τρεις κύριες επιπτώσεις:

1. Ορθογωνοποιεί τις συνιστώσες των διανυσμάτων χαρακτηριστικών έτσι ώστε να μη συσχετίζονται μεταξύ τους.
2. Διατάσσει τις ορθογωνοποιημένες συνιστώσες, δηλαδή της κύριες συνιστώσες, έτσι ώστε αυτές με την μεγαλύτερη διαφοροποίηση να έρχονται πρώτες.
3. Εξαλείφει αυτές τις συνιστώσες που συνεισφέρουν ελάχιστα στην διαφοροποίηση των δεδομένων.

Στην πραγματικότητα, ο αλγόριθμος υπολογίζει καταρχήν τη μέση τιμή των διανυσμάτων χαρακτηριστικών και στη συνέχεια αφαιρεί από αυτά την μέση τιμή. Κατόπιν, υπολογίζεται ο πίνακας συνδιακύμανσης και εντοπίζονται οι ιδιοτιμές και τα ιδιοδιανύσματα του. Τα ιδιοδιανύσματα που αντιστοιχούν στις M μεγαλύτερες ιδιοτιμές διατηρούνται και τα διανύσματα χαρακτηριστικών προβάλλονται πάνω στα ιδιοδιανύσματα αυτά για να υπολογίσουν τις συνιστώσες των μετασχηματισμένων διανυσμάτων του M -διάστατου χώρου. Για το πρόβλημα αυτό, χρησιμοποιήσαμε την Ανάλυση Κύριων Συνιστωσών για να ελαττώσουμε την διάσταση του χώρου των χαρακτηριστικών από 40 σε 11. Καθένα από αυτά τα νέα 11 χαρακτηριστικά υπολογίστηκε ως το άθροισμα των κυρίων συνιστωσών που προέκυψαν από κάθε πρωτότυπο χαρακτηριστικό (Πίνακας 1 στην εργασία

[42]). Εν συνεχεία, τα διανύσματα εισόδου μετατρέπονται από τα αρχικά 40 χαρακτηριστικά στα 11 που προέρχονται από την Ανάλυση Κύριων Συνιστωσών. Τα διανύσματα αυτά είναι τώρα μη συσχετισμένα.

Για την εκπαίδευση των τεχνητών νευρωνικών δικτύων (εμπρόσθια τροφοδοτούμενων και πιθανοτικών) χρησιμοποιήθηκαν 91 πρότυπα (41 χαμηλού και 50 υψηλού βαθμού). Για την εξέταση της γενίκευσης χρησιμοποιήθηκαν τα υπόλοιπα πρότυπα (20 χαμηλού και 29 υψηλού βαθμού).

Για το πρόβλημα αυτό χρησιμοποιήσαμε δύο τεχνητά νευρωνικά δίκτυα. Το πρώτο ήταν ένα εμπρόσθια τροφοδοτούμενο νευρωνικό δίκτυο το οποίο εκπαιδεύτηκε με την μέθοδο NMSP2. Δοκιμάστηκαν πολλές αρχιτεκτονικές δικτύων αφού δεν ήταν γνωστό εκ των προτέρων ποια αρχιτεκτονική θα οδηγούσε στα καλύτερα αποτελέσματα. Η αρχιτεκτονική που χρησιμοποιήθηκε ήταν τριών επιπέδων και αποτελούνταν από το επίπεδο εισόδου (11 είσοδοι), το κρυφό επίπεδο με N νευρώνες όπου $N = 5, \dots, 21$ και το επίπεδο εξόδου με 2 νευρώνες. Το κρυφό επίπεδο και το επίπεδο εξόδου του νευρωνικού δικτύου είναι βασισμένο σε νευρώνες με λογιστική συνάρτηση ενεργοποίησης. Η εκπαίδευση του τεχνητού νευρωνικού δικτύου ολοκληρώνεται όταν το δίκτυο έχει σφάλμα ταξινόμησης του συνόλου εκπαίδευσης λιγότερο από 1%.

Ο δεύτερος ταξινομητής ήταν ένα πιθανοτικό νευρωνικό δίκτυο. Το δίκτυο αποτελείται από τέσσερα επίπεδα. Το πρώτο επίπεδο είναι το επίπεδο εισόδου το οποίο έχει ένα κόμβο για κάθε συνιστώσα του διανύσματος εισόδου. Το δεύτερο επίπεδο είναι το επίπεδο προτύπων το οποίο περιέχει ένα κόμβο για κάθε πρότυπο εκπαίδευσης. Κάθε κόμβος προτύπου σχηματίζει το γινόμενο του διανύσματος των βαρών με το δοθέν παράδειγμα προς ταξινόμηση, όπου τα βάρη που εισέρχονται στον κόμβο είναι από ένα συγκεκριμένο παράδειγμα. Στη συνέχεια το γινόμενο τροφοδοτείται σε μια συνάρτηση ενεργοποίησης που ακολουθεί την κανονική κατανομή

$$k(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)$$

και εισέρχεται στο επίπεδο άθροισης το οποίο λαμβάνει τις εξόδους από το επίπεδο προτύπων που σχετίζονται με μια τάξη. Το επίπεδο εξόδου έχει τόσους κόμβους όσες είναι και οι τάξεις.

Η καλύτερη απόδοση επιτεύχθηκε χρησιμοποιώντας το εμπρόσθια τροφοδοτούμενο νευρωνικό δίκτυο που εκπαιδεύτηκε με την μέθοδο NMSP2 με 7 νευρώνες στο κρυφό επίπεδο, το οποίο διέκρινε χαμηλού βαθμού από υψηλού βαθμού όγκους με ακρίβεια 92% (Πίνακας 5.5). Η ευαισθησία και η εξειδίκευση κυμάνθηκε στο 93.1% και 90.5%, αντίστοιχα. Το πιθανοτικό νευρωνικό δίκτυο είχε χαμηλότερα ποσοστά ταξινόμησης (83.8% εξειδίκευση, 91.5% ευαισθησία και 89.9% συνολική ακρίβεια). Αυτό μπορεί να εξηγηθεί από την πολύπλοκη πιθανοτική κατανομή που ακολουθούν τα δεδομένα που με το πιθανοτικό νευρωνικό δίκτυο προσεγγίζεται χρησιμοποιώντας ένα πυρήνα που ακολουθεί την κανονική κατανομή. Ωστόσο, αξίζει να αναφερθεί ότι το πιθανοτικό

νευρωνικό δίκτυο εκπαιδεύτηκε ταχύτερα από το εμπρόσθια τροφοδοτούμενο νευρωνικό δίκτυο, ενώ δεν χρειάστηκαν δοκιμές για την τοπολογία του.

Αλγόριθμος	Εξειδίκευση	Ευαισθησία	Συνολική Ακρίβεια
MLP	90.5%	93.1%	92.0%
PNN	83.3%	91.5%	89.9%

Πίνακας 5.5: Συγκριτικά αποτελέσματα για το πρόβλημα της Ταξινόμησης Καρκινικών Κυττάρων Εγκεφάλου μεταξύ του εμπρόσθια τροφοδοτούμενου νευρωνικού δικτύου με 7 νευρώνες (MLP) στο κρυφό επίπεδο και του πιθανοτικού νευρωνικού δικτύου (PNN).

Στο σημείο αυτό, αξίζει να αναφερθεί ότι το καλύτερο αποτέλεσμα ταξινόμησης δεν είναι αυτό το οποίο παρουσιάζει την βέλτιστη συνολική ακρίβεια. Το διαγνωστικό λάθος είναι πιο σημαντικό όταν ταξινομείτε λανθασμένα ένας καρκινικός όγκος υψηλού βαθμού ως χαμηλού βαθμού, αφού τέτοιου είδους λάθη οδηγούν σε λανθασμένη αγωγή με σοβαρές συνέπειες στην υγεία του ασθενή. Από την άλλη μεριά η λανθασμένη ταξινόμηση των καρκινικών όγκων χαμηλού βαθμού δεν είναι τόσο σημαντική. Επομένως, βέλτιστο αποτέλεσμα ταξινόμησης θεωρείται η μεγιστοποίηση της ακρίβειας αναγνώρισης των καρκινικών όγκων υψηλού βαθμού με το χαμηλότερο δυνατό σφάλμα στους καρκινικούς όγκους χαμηλού βαθμού [39]. Σύμφωνα με αυτή τη λογική θεωρήθηκε το εμπρόσθια τροφοδοτούμενο νευρωνικό δίκτυο με 7 νευρώνες στο κρυφό επίπεδο ως η βέλτιστη αρχιτεκτονική αφού είχε τον βέλτιστο συνδυασμό εξειδίκευσης και ευαισθησίας (Πίνακας 5.5). Η καλύτερη ευαισθησία επιτεύχθηκε με το εμπρόσθια τροφοδοτούμενο νευρωνικό δίκτυο με 8 νευρώνες στο κρυφό επίπεδο (96.5%), αλλά αυτή η αρχιτεκτονική δεν επιλέχθηκε ως βέλτιστη αφού το σχετικό κόστος στην ταξινόμηση των καρκινικών όγκων χαμηλού βαθμού ήταν σημαντικό (85.7%). Η Εικόνα 5.2 παρουσιάζει την απόδοση του εμπρόσθια τροφοδοτούμενου νευρωνικού δικτύου για διάφορες αρχιτεκτονικές. Η ίδια στρατηγική ακολουθήθηκε και για το πιθανοτικό νευρωνικό δίκτυο. Η καλύτερη απόδοση (Πίνακας 5.5) επιτεύχθηκε για $\sigma = 0.05$.

5.4 Συμπεράσματα

Σε αυτό το κεφάλαιο παρουσιάστηκαν δύο νέοι αλγόριθμοι εκπαίδευσης τεχνητών νευρωνικών δικτύων που βασίζονται στην μέθοδο συζυγών κλίσεων του Perry. Πιο συγκεκριμένα, συνδυάστηκε ο κλιμακωτός αλγόριθμος συζυγών κλίσεων του Perry με παράμετρο κλιμάκωσης το βήμα των Barzilai και Borwein με μια μη μονότονη στρατηγική γραμμικής αναζήτησης. Επίσης οι αλγόριθμοι εκπαίδευσης χρησιμοποιούν μια τεχνική επανεκκίνησης η οποία εξασφαλίζει και την ιδιότητα ολικής σύγκλισης. Η διαφορά των δύο προτεινόμενων αλγορίθμων είναι η παράμετρος του αρχικού ρυθμού εκπαίδευσης που τροφοδοτείται στην διαδικασία της γραμμικής αναζήτησης. Στον πρώτο αλγόριθμο ο αρχικός ρυθμός εκπαίδευσης προσαρμόζεται αυτόματα σε κάθε επανάληψη σύμφωνα

Σχήμα 5.2: Καλύτερες επιδόσεις για το εμπρόσθια τροφοδοτούμενο νευρωνικό δίκτυο για διαφορετικό αριθμό νευρώνων στο κρυφό επίπεδο.

με ένα κλειστό τύπο που προτάθηκε στις εργασίες [77], [80] ενώ στον δεύτερο αλγόριθμο χρησιμοποιείται ο ρυθμός εκπαίδευσης που προτάθηκε στην εργασία [41].

Τα πειραματικά αποτελέσματα έδειξαν ότι οι προτεινόμενοι αλγόριθμοι είναι αποδοτικότεροι από τους κλασικούς αλγορίθμους εκπαίδευσης, ενώ η τροποποίηση του αρχικού ρυθμού εκπαίδευσης που προτάθηκε στην εργασία [41] βελτίωσε περαιτέρω την απόδοση του μη μονότονου κλιμακωτού αλγορίθμου εκπαίδευσης συζυγών κλίσεων του Perry. Ωστόσο, η μη μονότονη τεχνική γραμμικής αναζήτησης που χρησιμοποιήθηκε στις εργασίες αυτές είναι σχετικά απλή. Πιστεύουμε ότι οι συγκεκριμένοι αλγόριθμοι μπορεί να βελτιωθούν περαιτέρω χρησιμοποιώντας πιο εξελιγμένες μη μονότονες τεχνικές γραμμικής αναζήτησης όπου ο ρυθμός εκπαίδευσης δεν θα επιτυγχάνεται από την απλή υποδιαίρεση του αρχικού ρυθμού εκπαίδευσης.

Μέρος ΙΙΙ

ΠΑΡΑΡΤΗΜΑΤΑ—ΒΙΒΛΙΟΓΡΑΦΙΑ

Προβλήματα Εκπαίδευσης Νευρωνικών Δικτύων

Σε αυτό το παράρτημα παρουσιάζονται τα προβλήματα εκπαίδευσης τεχνητών νευρωνικών δικτύων που χρησιμοποιήθηκαν σε αυτή τη διατριβή για την αξιολόγηση των νέων αλγορίθμων εκπαίδευσης ως προς την ταχύτητα, την αξιοπιστία και την γενίκευση.

Α'.1 Αποκλειστικό-ΕΙΤΕ (XOR)

Το Αποκλειστικό-ΕΙΤΕ είναι ένα πολύ δύσκολο πρόβλημα που χρησιμοποιείται ευρέως για την αξιολόγηση αλγορίθμων εκπαίδευσης τεχνητών νευρωνικών δικτύων. Η συνάρτηση του Αποκλειστικού-ΕΙΤΕ απεικονίζει δύο δυαδικές εισόδους σε μια δυαδική έξοδο. Στο Σχήμα Α'.1 απεικονίζονται στο επίπεδο τα πρότυπα εκπαίδευσης του Αποκλειστικού-ΕΙΤΕ. Όπως φαίνεται από το Σχήμα Α'.1, η συνάρτηση του Αποκλειστικού-ΕΙΤΕ δεν είναι γραμμικά διαχωρίσιμη. Γι' αυτό το λόγο απαιτείται τουλάχιστον ένα κρυφό επίπεδο νευρώνων για την επίλυση του προβλήματος. Το πρόβλημα αυτό είναι πολύ ευαίσθητο στην επιλογή των αρχικών βαρών, και η επιφάνεια σφάλματος παρουσιάζει πολλά τοπικά ελάχιστα με σχετικά μεγάλες συναρτησιακές τιμές.

Σχήμα Α'.1: Το πρόβλημα του Αποκλειστικού-ΕΙΤΕ στο επίπεδο. Από το σχήμα φαίνεται καθαρά ότι τα πρότυπα εκπαίδευσης δεν είναι γραμμικά διαχωρίσιμα.

Α'.2 Ισοτιμία των 3-bit

Το πρόβλημα της Ισοτιμίας των 3-bit μπορεί να θεωρηθεί ως μια γενίκευση του προβλήματος του αποκλειστικού-ΕΙΤΕ, το οποίο όμως είναι πολύ πιο δύσκολο. Η δυαδική συνάρτηση του προβλήματος απεικονίζει τρεις δυαδικές εισόδους σε μια δυαδική έξοδο. Η τιμή της εξόδου είναι 1 όταν ο αριθμός των bits με τιμή 1 στην είσοδο είναι περιττός, και 0 σε κάθε άλλη περίπτωση (περιττή ισοτιμία). Η επιφάνεια σφάλματος του προβλήματος αυτού παρουσιάζει πολλά τοπικά ελάχιστα όπου είναι εύκολο να παγιδευτούν οι αλγόριθμοι εκπαίδευσης. Στο Πίνακα Α'.1 απεικονίζονται οι 8 εισόδοι και οι αντίστοιχες εξόδοι του προβλήματος.

Είσοδος	Έξοδος
(0, 0, 0)	0
(0, 0, 1)	1
(0, 1, 0)	1
(0, 1, 1)	0
(1, 0, 0)	1
(1, 0, 1)	0
(1, 1, 0)	0
(1, 1, 1)	1

Πίνακας Α'.1: Τα πρότυπα εκπαίδευσης της Ισοτιμίας των 3-bit.

Α'.3 Αναγνώριση Κεφαλαίων Γραμμάτων

Το πρόβλημα της αναγνώρισης των κεφαλαίων γραμμάτων της Αγγλικής αλφαβήτου αποτελείται από 26 πίνακες διαστάσεως 7×5 . Οι πίνακες αυτοί έχουν κωδικοποιηθεί με δυαδικές τιμές και αποτελούν την είσοδο του δικτύου. Τα γράμματα Α, Β, C και D και οι δυαδικές τους κωδικοποιήσεις φαίνονται στο Σχήμα Α'.2. Οι έξοδοι έχουν κωδικοποιηθεί με διανύσματα 26 δυαδικών στοιχείων. Τα 25 στοιχεία του εκάστοτε διανύσματος εξόδου έχουν την δυαδική τιμή 0 και μόνο ένα έχει την δυαδική τιμή 1 και βρίσκεται στην αντίστοιχη θέση που έχει το γράμμα στην Αγγλική αλφάβητο.

Σχήμα Α'.2: Η κωδικοποίηση των γραμμάτων Α, Β, C και D και οι αντίστοιχες δυαδικές κωδικοποιήσεις.

Α'.4 Προσέγγιση Συνεχούς Τριγωνομετρικής Συνάρτησης

Το πρόβλημα της προσέγγισης μιας συνάρτησης είναι ένα από τα βασικά προβλήματα που καλούνται να λύσουν τα τεχνητά νευρωνικά δίκτυα. Στην περίπτωση αυτή το τεχνητό νευρωνικό δίκτυο καλείται να προσεγγίσει την συνεχή τριγωνομετρική συνάρτηση $f(x) = \sin(x) \cos(2x)$. Τα πρότυπα εισόδου είναι 20 ισαπέχουσες τιμές στο διάστημα $[0, 2\pi]$. Η έξοδος αποτελείται από την τιμή $f(x)$ για την εκάστοτε είσοδο x . Στο Σχήμα Α'.3 φαίνεται η γραφική παράσταση της συνάρτησης f και τα πρότυπα εκπαίδευσης.

Σχήμα Α'.3: Η γραφική παράσταση της $f(x) = \sin(x) \cos(2x)$ στο $[0, 2\pi]$. Τα 20 πρότυπα εκπαίδευσης απεικονίζονται πάνω στη γραφική παράσταση με κόκκινες κουκίδες.

Α'.5 $N-M-N$ Κωδικοποιητής/Αποκωδικοποιητής

Σε αυτό το πρόβλημα το τεχνητό νευρωνικό δίκτυο δέχεται M διαφορετικά πρότυπα εισόδου μήκους M bit όπου το καθένα έχει ένα μόνο bit με την τιμή 1 και όλα τα υπόλοιπα με την τιμή 0. Η έξοδος του δικτύου είναι ακριβώς η ίδια με την είσοδο. Αφού οι πληροφορίες της εισόδου περνούν από το κρυφό επίπεδο των M νευρώνων (όπου $M = \log_2 N$), το δίκτυο πρέπει να κατασκευάσει μια μοναδική κωδικοποίηση για κάθε ένα από τα N πρότυπα στους M κρυφούς νευρώνες και βάρη τέτοια ώστε να είναι δυνατή η κωδικοποίηση και η αποκωδικοποίηση. Το πρόβλημα αυτό, αν και

φαίνεται απλό, προσομοιάζει πραγματικά προβλήματα ταξινόμησης όπου μικρές αλλαγές στις εισόδους έχουν ως αποτέλεσμα μικρές αλλαγές στις εξόδους. Δύο παραδείγματα $N - M - N$ Κωδικοποιητή/Αποκωδικοποιητή για $N = 4$ και $N = 8$ δίνονται στους Πίνακες Α'.2 και Α'.3, αντίστοιχα.

Είσοδος	Έξοδος
(1, 0, 0, 0)	(1, 0, 0, 0)
(0, 1, 0, 0)	(0, 1, 0, 0)
(0, 0, 1, 0)	(0, 0, 1, 0)
(0, 0, 0, 1)	(0, 0, 0, 1)

Πίνακας Α'.2: Τα πρότυπα εκπαίδευσης $4 - 2 - 4$ Κωδικοποιητή/Αποκωδικοποιητή.

Είσοδος	Έξοδος
(1, 0, 0, 0, 0, 0, 0, 0)	(1, 0, 0, 0, 0, 0, 0, 0)
(0, 1, 0, 0, 0, 0, 0, 0)	(0, 1, 0, 0, 0, 0, 0, 0)
(0, 0, 1, 0, 0, 0, 0, 0)	(0, 0, 1, 0, 0, 0, 0, 0)
(0, 0, 0, 1, 0, 0, 0, 0)	(0, 0, 0, 1, 0, 0, 0, 0)
(0, 0, 0, 0, 1, 0, 0, 0)	(0, 0, 0, 0, 1, 0, 0, 0)
(0, 0, 0, 0, 0, 1, 0, 0)	(0, 0, 0, 0, 0, 1, 0, 0)
(0, 0, 0, 0, 0, 0, 1, 0)	(0, 0, 0, 0, 0, 0, 1, 0)
(0, 0, 0, 0, 0, 0, 0, 1)	(0, 0, 0, 0, 0, 0, 0, 1)

Πίνακας Α'.3: Τα πρότυπα εκπαίδευσης $8 - 3 - 8$ Κωδικοποιητή/Αποκωδικοποιητή.

Α'.6 Αναγνώριση Αριθμών

Το πρόβλημα της αναγνώρισης των αριθμών αποτελείται από 10 πίνακες διαστάσεως 8×8 . Οι πίνακες αυτοί έχουν κωδικοποιηθεί με δυαδικές τιμές και αποτελούν την είσοδο του δικτύου. Οι αριθμοί 0 και 1 και οι δυαδικές τους κωδικοποιήσεις φαίνονται στο Σχήμα Α'.4. Οι έξοδοι έχουν κωδικοποιηθεί με διανύσματα 10 δυαδικών στοιχείων. Τα 10 στοιχεία του εκάστοτε διανύσματος εξόδου έχουν την δυαδική τιμή 0 και μόνο ένα έχει την δυαδική τιμή 1 και βρίσκεται στην αντίστοιχη θέση που έχει ο αριθμός.

Σχήμα Α'.4: Η κωδικοποίηση των αριθμών 0 και 1 και οι αντίστοιχες δυαδικές κωδικοποιήσεις.

Α'.7 Ταξινόμηση των Φυτών της Οικογένειας Iris

Το πρόβλημα της ταξινόμησης των φυτών της οικογένειας Iris [24] αποτελείται από 150 δείγματα, το καθένα από τα οποία έχει τέσσερα χαρακτηριστικά. Το σύνολο περιέχει τρεις κατηγορίες, δύο από τις οποίες είναι γραμμικά μη διαχωρίσιμες. Κάθε κατηγορία αποτελείται από 50 περιπτώσεις και παραπέμπουν στα φυτά της οικογένειας Iris, Setosa, Versicolor και Virginica. Στις Εικόνες Α'.5, Α'.6 και Α'.7 απεικονίζονται τα τρία φυτά της οικογένειας Iris που χρησιμοποιούνται σε αυτό το πρόβλημα.

Σχήμα Α'.5: Iris Setosa.

Σχήμα Α'.6: Iris Versicolor.

Σχήμα Α'.7: Iris Virginica.

Α'.8 Οπτική Αναγνώριση Χειρόγραφων Αριθμών

Το πρόβλημα της οπτικής αναγνώρισης χειρόγραφων αριθμών είναι ένα πραγματικό πρόβλημα ταξινόμησης. Σε αυτό το πείραμα, δημιουργήθηκε μια ψηφιακή βάση δεδομένων συλλέγοντας δείγματα από 43 ανεξάρτητους συγγραφείς. Τα δείγματα που συλλέχθηκαν από 30 συγγραφείς χρησιμοποιήθηκαν για την εκπαίδευση ενώ τα υπόλοιπα για τον έλεγχο της γενίκευσης. Το σύνολο εκπαίδευσης αποτελείται από 3823 πρότυπα, ενώ το σύνολο ελέγχου της γενίκευσης από 1797 πρότυπα. Κάθε ψηφίο έχει κωδικοποιηθεί σε δυαδική μορφή και έχουν τοποθετηθεί σε πίνακες διαστάσεως 32×32 . Για να ελαττώσουμε την διάσταση του διανύσματος εισόδου οι πίνακες διαστάσεως 32×32 διαιρέθηκαν σε μη επικαλυπτόμενα τμήματα διαστάσεως 4×4 και μετρήθηκε ο αριθμός των μονάδων σε κάθε τμήμα. Έτσι δημιουργήθηκε ένας πίνακας διαστάσεως 8×8 . Ωστόσο, στη συνέχεια τα διανύσματα εισόδου μετατράπηκαν έτσι ώστε οι συνιστώσες τους να έχουν τιμές που να βρίσκονται μέσα στο διάστημα $[0, 1]$.

Α'.8.1 Αναγνώριση Φωνηέντων

Ο στόχος αυτού του πραγματικού προβλήματος ταξινόμησης είναι η αναγνώριση ενός φωνηέντος από έναν τυχαίο ομιλητή. Πιο συγκεκριμένα δημιουργήθηκε μια βάση δεδομένων συλλέγοντας δείγματα από 15 διαφορετικούς ομιλητές (8 άντρες και 7 γυναίκες) που έλεγαν τα 11 φωνήεντα 6 φορές (990 δείγματα). Τα δεδομένα χωρίστηκαν σε ένα σύνολο εκπαίδευσης και σε ένα σύνολο ελέγχου της γενίκευσης. Το σύνολο εκπαίδευσης αποτελείται από τέσσερις άντρες και τέσσερις γυναίκες ομιλητές ($11 \times 6 \times 8 = 528$ δείγματα). Το σύνολο ελέγχου της γενίκευσης αποτελείται από τέσσερις άντρες και τρεις γυναίκες ομιλητές ($11 \times 6 \times 7 = 462$ δείγματα). Κάθε δείγμα έχει δέκα χαρακτηριστικά με συνεχείς τιμές που προέρχονται από τα φασματικά δεδομένα (για περισσότερες πληροφορίες για το πρόβλημα αυτό βλέπε τις εργασίες [61] και [53]).

Δημοσιεύσεις - Αναφορές

Δημοσιεύσεις

- [1]. D. G. Sotiropoulos, A. E. Kostopoulos, and T. N. Grapsa. A spectral version of Perry's conjugate gradient method for neural network training. In D. T. Tsahalis, editor, *4th GRACM Congress on Computational Mechanics*, volume 1, pages 172–179, 2002.
- [2]. D. G. Sotiropoulos, A. E. Kostopoulos, and T. N. Grapsa. Training neural networks using two-point stepsize gradient methods. In T. E. Simos and Ch. Tsitouras, editors, *Proceedings of International Conference of Numerical Analysis and Applied Mathematics 2004 (ICNAAM-2004)*, pages 356–359, 2004.
- [3]. A. E. Kostopoulos, D. G. Sotiropoulos, and T. N. Grapsa. A new efficient variable learning rate for Perry's Spectral conjugate gradient training method. In D. T. Tsahalis, editor, *Proceedings of 1st International Conference "From Scientific Computing to Computational Engineering"*, volume 1, pages 66–73, 2004.
- [4]. A. E. Kostopoulos, D. Glotsos, P. Spyridonos, G. Nikiforidis, D. G. Sotiropoulos, and T. N. Grapsa. Comparative evaluation of feedforward and probabilistic neural networks for the automatic classification of brain tumours. In D. T. Tsahalis, editor, *Proceedings of 1st International Conference "From Scientific Computing to Computational Engineering"*, volume 1, pages 309–315, 2004.
- [5]. A. E. Kostopoulos, A. P. Happonen, and U. Ruotsalainen. Three-dimensional alignment of objects in dynamic PET scans using filtered sinusoidal trajectories of sinogram. *Nuclear Instruments and Methods in Physics Research Section A*, 569(2):434-439, 2006.

- [6]. T. N. Grapsa, G. E. Antonelou, and A.E. Kostopoulos. Perturbed Newton method for Unconstrained Optimization. In E. Gallopoulos, E. Houstis, I.S. Kotsireas, D. Noutsos and M.N. Vrahatis, editors, *Proceedings of Conference in Numerical Analysis, Recent Approaches to Numerical Analysis, Theory, Methods and Applications (NumAn2007)*, pages 76–79, 2007.
- [7]. A. E. Kostopoulos, and T. N. Grapsa. Self-scaled conjugate gradient training algorithms. *Neurocomputing*, 72:3000–3019, 2009.
- [8]. A.E. Kostopoulos, G.S. Androulakis, and T.N. Grapsa. A new nonmonotone Newton’s modification for unconstrained optimization, Technical report 09-05, Department of Mathematics, University of Patras, Patras, Greece, 2009.

Αναφορές

Στην εργασία [1]

- [1]. X. Mu, and Y. Zhang. A Modified Gradient-based Backpropagation Training Method for Neural Networks. In *The 2009 IEEE International Conference on Granular Computing, GrC 2009, GrC 2009*, Lushan Mountain, Nanchang, China, 17-19 August 2009, GrC, IEEE, 2009.
- [2]. X. Mu, and Y. Zhang. *Advances in Neural Networks - ISNN 2010*, chapter A New Two-Step Gradient-Based Backpropagation Training Method for Neural Networks. Springer Berlin / Heidelberg, Volume 6063, pages 95-101, 2010.
- [3]. Nazri Mohd Nawi, Meghana R. Ransing, and Rajesh S. Ransing. An Improved Learning Algorithm Based on The Broyden-Fletcher-Goldfarb-Shanno (BFGS) Method For Back Propagation Neural Networks. In *Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA 06)*, pages 152–157, 2006.
- [4]. Nazri Mohd Nawi, Rajesh S. Ransing, and Meghana R. Ransing. A New Method to Improve the Gradient Based Search Direction to Enhance the Computational Efficiency of Back Propagation Based Neural Network Algorithms. In *Proceedings of the Second Asia International Conference on Modelling & Simulation, Asia Modelling Symposium (AMS 2008)*, pages 546-552, 2008.
- [5]. G. Lappas, R. J. Frank, A. A. Albrecht. A computational study on circuit size versus circuit depth. *International Journal on Artificial Intelligence Tools*, 15(2):143-161, 2006.

- [6]. C-C Peng, and G. D. Magoulas. Adaptive Nonmonotone Conjugate Gradient Training Algorithm for Recurrent Neural Networks. In *Proceedings of 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, 29-31 Oct. 2007, volume 2, pages 374-381, 2007.
- [7]. C-C Peng, and G. D. Magoulas. Advanced Adaptive Nonmonotone Conjugate Gradient Training Algorithm for Recurrent Neural Networks. *International Journal on Artificial Intelligence Tools*, 17(5):963-984, 2008.
- [8]. I.E. Livieris, and P. Pintelas. A survey on algorithms for training artificial neural networks, Technical report 08-01, Department of Mathematics, University of Patras, Patras, Greece, 2008.
- [9]. I.E. Livieris, and P. Pintelas. Evaluation of neural network training algorithms using biomedical data, Technical report 08-02, Department of Mathematics, University of Patras, Patras, Greece, 2009.
- [10]. I.E. Livieris, and P. Pintelas. Performance Evaluation of Descent CG Methods for Neural Network Training. In *Proceedings of 9th Hellinic European Research on Computer and its Application Conference (HERCMA '09)*, Athens, Greece, September, 2009.

Στην εργασία [3]

- [1]. M. Terzyiska, Y. Todorov and M. Petrov. NMPC with Adaptive Learning Rate Scheduling of an Internal Fuzzy-Neural Model. In *Proceedings of the International Conference ISAC 2006: Modern Trends in Control, International Study in Automatic Control, Košice*, Slovak Republic, 15–30 June 2006, pages 289-298, 2006.

Στην εργασία [4]

- [1]. E. I. Papageorgiou, P. P. Spyridonos, D. T. Glotsos, C. D. Stylios, P. Ravazoula, G. N. Nikiforidis, P. P. Groumpos. Brain tumor characterization using the soft computing technique of fuzzy cognitive maps. *Applied Soft Computing*, 8(1):820-828, 2008.

Στην εργασία [5]

- [1]. A. P. Happonen, and M. O. Koskinen. Assessment of automated scintigram alignment in the stackgram domain for cardiac stress-rest SPECT studies. *Computerized Medical Imaging and Graphics*, 31:595-606, 2007.

Στην εργασία [7]

- [1]. S. Chabaa, A. Zeroual, and J. Antari. Identification and Prediction of Internet Traffic Using Artificial Neural Networks. *Journal of Intelligent Learning Systems and Applications*, 2:147-155, 2010.
- [2]. M. Kabsch-Korbutowicz, and M. Kutylowska. Artificial Neural Networks - A Viable Option for Predicting Changes in Water Turbidity after Treatment by Coagulation/Ultrafiltration. *OCHRONA ŚRODOWISKA*, 32(2):5-20, 2010.
- [3]. Gao Donglian, and Shan Rui. TR-BFGS Algorithm for Multi-layer Feed-Forward Neural Networks . In *Cryptography and Network Security, Data Mining and Knowledge Discovery, E-Commerce & Its Applications and Embedded Systems (CDEE), 2010 First ACIS International, 23-24 October 2010*, pages 143-146, 2010.
- [4]. Y. Zhang, Z. Dong, L. Wu, and S. Wang. A hybrid method for MRI brain image classification. *Expert Systems with Applications*, 38(8):10049-10053 , 2011.
- [5]. J. C. F. Wong, and J. L. Xie. Inverse determination of a heat source from natural convection in a porous cavity. *Computers & Fluids*, 52:1-14, 2011.
- [6]. Jeff C. F. Wong and Bartosz Protas. Application of scaled nonlinear conjugate-gradient algorithms to the inverse natural convection problem. *Optimization Methods and Software*, iFirst:1-27, 2011.

Βιβλιογραφία

- [1] H. Akaike On successive transformation of probability distribution and its application to the analysis of the optimum gradient method. *Ann. Inst. Statist. Math. Tokyo*, 11:1–17, 1959.
- [2] M. Al Baali Improved Hessian approximations for the limited memory BFGS method. *Numerical Algorithms*, 22:99–112, 1999.
- [3] A. D. Anastasiadis, G. D. Magoulas and M. N. Vrahatis. New globally convergent training scheme based on the resilient propagation algorithm. *Neurocomputing*, 64:253–270, 2005.
- [4] A. D. Anastasiadis, G. D. Magoulas and M. N. Vrahatis. Sign-based learning schemes for pattern classification. *Pattern Recognition Letters*, 26:1926–1936, 2005.
- [5] J. Barzilai and J. M. Borwein. Two point step size gradient methods. *IMA J. Numer. Anal.*, 8:141–148, 1988.
- [6] R. Battiti First- and second-order methods for learning: Between steepest descent and newton’s method. *Neural Computation*, 4:141–166, 1992.
- [7] E. M. L. Beale *Numerical Methods for Nonlinear Optimization*, chapter A derivation of conjugate gradients, pages 39–43. Academic Press, London, 1972.
- [8] S. Becker and Y. Le Cun. Improving the convergence of the back-propagation learning with second order methods. In D. S. Touretzky, G. E. Hinton, T. J. Sejnowski, editor, *1988 Connectionist Models Summer School*.
- [9] E. G. Birgin and J. M. Martinez. A spectral conjugate gradient method for unconstrained optimization. *Applied Mathematics and Optimization*, 43:117–128, 1999.
- [10] C. M. Bishop *em Neural networks for pattern recognition*. Oxford University Press Inc, New York, 1995.

- [11] A. Bortoletti, C. Di Fiore, S. Fanelli and P. Zellini. A new class of quasi-Newtonian methods for optimal learning in MLP-networks. *IEEE Trans. Neural Networks*, 14:263–273, 2003.
- [12] R. P. Brent. *Algorithms for minimization without derivatives*. Englewood Cliffs, NJ: Prentice Hall, 1973.
- [13] A. E. Jr. Bryson and Y. C. Ho. *Applied Optimal Control*. Blaisdell, 1969.
- [14] A. Cauchy. Méthode générale pour la résolution des systèmes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25:536–538, 1847.
- [15] L. W. Chan and F. Fallside. An adaptive training algorithm for back-propagation networks. *Computers Speech and Language*, 2:205–218, 1987.
- [16] C. Charalambous. Conjugate gradient algorithm for efficient training of artificial neural networks. *IEEE Proceedings*, 139:301–310, 1992.
- [17] Y. H. Dai, L. Z. Liao and D. Li. On restart procedures for the conjugate gradient method. *Numerical Algorithms*, 35:249–260, 2004.
- [18] W. C. Davidon. Variable metric method for minimization. Technical report, Technical Report ANL-5990, Argonne National Laboratory, 1959.
- [19] C. Decaestecker, I. Salmon, O. Dewitte, I. Camby, P. Van Ham, J. Pasteels, J. Brotchi and R. Kiss. Nearest-neighbor classification for identification of aggressive versus nonaggressive astrocytic tumors by means of image cytometry-generated variables. *Journal of Neurosurgery*, 86:532–537, 1997.
- [20] C. Di Fiore, S. Fanelli, F. Lepore and P. Zellini. Matrix algebras in quasi-Newton methods for unconstrained minimization. *Numerische Mathematik*, 94:4979–5000, 2003.
- [21] C. Di Fiore, S. Fanelli and P. Zellini. An efficient generalization of Battiti-Shanno's quasi-Newton algorithm for optimal learning in MLP-networks. In N. R. Pal et al., editors, *ICONIC 2004 LNCS 3316*, volume 1.
- [22] L. C. W. Dixon, P. G. Ducksbury and P. Singth. A new three term conjugate gradient method. Technical report, Technical Report No. 130, Numerical Optimization Centre, The Hatfield Polytechnic, Hatfield, Hertfordshire, England, 1985.
- [23] S. E. Fahlman. Faster-learning variations on back-propagation: An empirical study. In D. S. Touretzky, G. E. Hinton, T. J. Sejnowski, editors, *1988 Connectionist Models Summer School*.
- [24] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7:179–188, 1936.

- [25] R. Fletcher. *Practical Methods of Optimization*. New York: Wiley, 1987.
- [26] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *Computer Journal*, 7:149–154, 1964.
- [27] J. D. Gibbons. *Nonparametric Statistical Inference*. M. Dekker, 2nd Ed., 1985.
- [28] D. Glotsos, P. Spyridonos, P. Petalas, D. Cavouras, I. Ravazoula, S. Dadioti, E. Lekka and G. Nikiforidis. Supporting the regular diagnostic procedure followed by the experts in astrocytomas malignancy grading by means of an automatic classification methodology using Support Vector Machines. *Analytical and Quantitative Cytology and Histology*, 26:77–83, 2004.
- [29] L. Grippo, F. Lampariello and S. Lucidi. A nonmonotone line search technique for Newton’s method. *SIAM Journal of Numerical Analysis*, 23:707–716, 1986.
- [30] J. Han, J. Sun and W. Sun. Global Convergence of Non-Monotone Descent Methods for Unconstrained Optimization Problems. *Journal of Computational and Applied Mathematics*, 2002.
- [31] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice–Hall Inc, 1999.
- [32] D. O. Hebb. *The Organization of Behavior: A Neurophysiological*. New York: Wiley, 1949.
- [33] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49:409–436, 1952.
- [34] H. C. Hsin, C. C. Li, M. Sun and R. J. Scabassi. An adaptive training algorithm for back-propagation neural networks. *IEEE Transactions on System, Man, and Cybernetics*, 25:521–514, 1995.
- [35] C. Igel and M. Hüsken. Empirical evaluation of the improved RPROP learning algorithms. *Neurocomputing*, 50:105–123, 2003.
- [36] R. A. Jacobs. Increased rates of convergence through learning rate adaption. *Neural Networks*, 1:295–307, 1988.
- [37] M. Jiang, G. Gielen, B. Zhang and Z. Luo. Fast learning algorithms for feedforward neural networks. *Applied Intelligence*, 18:37–54, 2003.
- [38] E. M. Johansson, F. U. Dowla and D. M. Goodman. Backpropagation learning for multi-layer feed-forward neural networks using the conjugate gradient training method. Technical report, Technical Report UCRL-JC-104850, Lawrence Livermore National Laboratory, 1990.

- [39] A. Kaye and D. Walker. Low-grade astrocytomas: controversies in management. *Journal of Clinical Neuroscience*, 7:475–483, 2000.
- [40] K. M. Khoda and C. Storey. Efficient implementation of a generalized polak–ribière algorithm for nonlinear optimization. *International Journal of Computer Mathematics*, 48:53–61, 1993.
- [41] A. E. Kostopoulos, D. G. Sotiropoulos and T. N. Grapsa. A new efficient variable learning rate for Perry’s Spectral conjugate gradient training method. In D. T. Tsahalis, editor, *Proceedings of 1st International Conference "From Scientific Computing to Computational Engineering"*, volume 1. pages 66–73, 2004
- [42] A. E. Kostopoulos, D. Glotsos, P. Spyridonos, G. Nikiforidis, D. G. Sotiropoulos and T. N. Grapsa. Comparative evaluation of feedforward and probabilistic neural networks for the automatic classification of brain tumours. In D. T. Tsahalis, editor, *Proceedings of 1st International Conference "From Scientific Computing to Computational Engineering"*, volume 1, pages 309–315, 2004.
- [43] A. E. Kostopoulos and T. N. Grapsa. Self-scaled conjugate gradient training algorithms. *Neurocomputing*, 72:3000–3019, 2009.
- [44] Y. LeCun. Une procedure d’apprentissage pour reseau a seuil assymetrique. *Cognitiva*, 85:599–604, 1985.
- [45] Y. LeCun. *Efficient Learning and Second-order Methods, A Tutorial at NIPS 93*. Denver, 1993.
- [46] D. C. Liu and J. Nocedal. On the L-BFGS method for large scale optimization. *Math. Programming*, 45:503–528, 1989.
- [47] D. Luenberger. *Introduction to Linear and Nonlinear Programming*. Addison Wesley, 1984.
- [48] G. D. Magoulas, M. N. Vrahatis and G. S. Androulakis. Effective backpropagation with variable stepsize. *Neural Networks*, 10:69–82, 1997.
- [49] G. D. Magoulas, M. N. Vrahatis and G. S. Androulakis. Improving the convergence of the back-propagation algorithm using learning rate adaptation techniques. *Neural Computation*, 11:1769–1796, 1999.
- [50] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.
- [51] M. L. Minsky and S. A. Papert. *Perceptrons*. Cambridge, MA: MIT Press, 1969.
- [52] M. F. Møller. A scaled conjugate gradient algorithm, for fast supervised learning. *Neural Networks*, 6:525–533, 1993.

- [53] P. M. Murphy and D. W. Aha. UCI repository of machine learning databases. Technical report, University of California, Department of Informatics and Computer Science, Irvine, CA, 1994.
- [54] L. Nazareth. A conjugate direction algorithm without line searches. *Journal of Optimization Theory and Applications*, 23:373–387, 1977.
- [55] D. Nguyen and B. Widrow. Improving the learning speed of 2-layer neural network by choosing initial values of the adaptive weights. *Biological Cybernetics*, 59:257–263, 1990.
- [56] J. Nocedal. Thoery of algorithms for unconstrained optimization. *Acta Numerica*, pages 199–242, 1992.
- [57] J. Nocedal and J. S. Wright. *Numerical Optimization*. Springer-Verlag, 1999.
- [58] D. B. Parker. Learning –logic: Casting the cortex of the human brain in silicon. Technical Report TR–47, Center for Computational Research in Economics and Management Sciense, Cambridge, MA: MIT Press, 1985.
- [59] D. B. Parker. Optimal algorithms for adaptive networks: Second order backpropagation, second order direct propagation, and second order Hebbian learning. In *IEEE 1st International Conference on Neural Networks*, volume 2, pages 593–600, 1987.
- [60] A. Perry. A modified conjugate gradient algorithm. *Operations Research*, 26:26–43, 1978.
- [61] M. Pfister and R. Rojas. Hybrid learning algorithms for neural networks - The adaptive inclusion of second order information. *Zeitschrift für Angewandte Mathematik und Mechanik*, 79:215–218, 1996.
- [62] M. Pfister and R. Rojas. Speeding up back-propagation-acomparison of orthogonal techniques. In *Joint Conference on Neural Networks*, pages 517–523, 1993.
- [63] V. P. Plagianakos, D. G. Sotiropoulos and M. N. Vrahatis. *Recent Advances in Circuits and Systems*, chapter Automatic adaptation of learning rate for back-propagation neural networks, pages 337–341. World Scientific Publishing Co. Pte. Ltd., 1998.
- [64] V. P. Plagianakos, G. D. Magoulas and M. N. Vrahatis. Deterministic nonmonotone strategies for effective training of multilayer perceptrons. *IEEE Trans. Neural Networks*, 13:1268–1284, 2002.
- [65] E. Polak and G. Ribière. Note sur la convergence de méthodes de directions conjuguées. *Revue Française d'Informatique et de Recherche Opérationnelle*, 16:35–43, 1969.

- [66] M. J. D. Powell. Restart procedures for the conjugate gradient method. *Mathematical Programming*, 12:241–254, 1977.
- [67] R. A. Prayson, D. P. Agamanolis, M. L. Cohen and M. L. Estes. Interobserver reproducibility among neuropathologists and surgical pathologists in fibrillary astrocytoma grading. *Neurological sciences*, 175:475–483, 2000.
- [68] S. Ramón y Cajál. *Histologie du Systéms Nerveux de l’homme et des vertébrés*. Paris: Maloine, 1911.
- [69] M. Raydan. The Barzilai and Borwein Gradient Method for the Large Scale Unconstrained Minimization Problem. *SIAM Journal of Optimization*, 13:321–326, 1993.
- [70] M. Riedmiller and H. Braun. A direct adaptive method for faster back-propagation learning: The rprop algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591, 1993.
- [71] F. Rosenblatt. The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958.
- [72] F. Rosenblatt. On the convergence of reinforcement procedures in simple perceptrons. Technical Report VG–1196–G–4, Cornel Aeronautical Laboratory, 1960.
- [73] D. E. Rumelhart, G. E. Hinton and R. J. Williams. Learning representations of backpropagation errors. *Nature (London)*, 323:533–536, 1986.
- [74] D. E. Rumelhart and J. L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*. Cambridge, MA: MIT Press, 1986.
- [75] L. E. Scales. *Introduction to non-linear optimization*. New York: Springer-Verlag, 1985.
- [76] D. F. Shanno. Globally convergent conjugate gradient methods. *Mathematical Programming*, 33:61–67, 1985.
- [77] D. F. Shanno and K. H. Phua. Minimization of unconstrained multivariate functions. *ACM Transactions on Mathematical Software*, 2:87–94, 1976.
- [78] D. F. Shanno and P. C. Kettler. Optimal conditioning of quasi-Newton methods. *Mathematical Computation*, 24:657–664, 1970.
- [79] F. Silva and Almeida. *Advanced Neural Computers*, chapter Speeding up Back-propagation, pages 151–158. Elsevier Science Publishers B.V. (North Holland, Amsterdam), 1990.

- [80] D. G. Sotiropoulos, A. E. Kostopoulos and T. N. Grapsa. A spectral version of Perry's conjugate gradient method for neural network training. In D. T. Tsahalis, editor, *4th GRACM Congress on Computational Mechanics*, volume 1, pages 172–179, 2002.
- [81] D. G. Sotiropoulos, A. E. Kostopoulos and T. N. Grapsa. Training neural networks using two-point stepsize gradient methods. In T. E. Simos and Ch. Tsitouras, editor, *Proceedings of International Conference of Numerical Analysis and Applied Mathematics 2004 (ICNAAM-2004)*, pages 356–359, 2004.
- [82] P. Spyridonos, V. Zolota, D. Cavouras, G. Zenebissis, D. Glotsos and G. Nikiforidis. A computer-based image analysis system for classification of astrocytomas according the WHO grading system. In *Proceedings of the IASTED International Conference on Signal Processing Pattern Recognition and Applications*, pages 371–374, 2002.
- [83] S. Theodoridis and K. Koutroubas. *Pattern Recognition*, chapter Feature Generation II, pages 233–270. Academic Press, 1999.
- [84] T. P. Vogl, J. K. Mangis, J. K. Rigler, W. T. Zink and D. L. Alkon. Accelerating the convergence of the back-propagation method. *Biological Cybernetics*, 59:257–263, 1988.
- [85] M. N. Vrahatis, G. D. Magoulas, V. P. Plagianakos. Globally convergent modification of the Quickprop method. *Neural Processing Letters*, 12:159–170, 2000.
- [86] P. J. Werbos. *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. PhD thesis, Harvard University, Cambridge, 1974.
- [87] B. Widrow. *Self-Organizing Systems*, chapter Generalization and information storage in networks of adeline "neurons", pages 435–461. Washington, DC: Spartan Books, 1962.
- [88] B. Widrow and M. E. Hoff, Jr. Adaptive switching circuits. *IRE WESCON Convention record*, pages 96–104, 1960.
- [89] F. Wilcoxon. Individual comparison by ranking methods. *Biometrics Bull.*, 1:80–83, 1945.
- [90] J. Z. Zhang, N. Y. Deng and L. H. Chen. New quasi-Newton equation and related methods for unconstrained optimization. *Journal of Optimization Theory and Applications*, 102:147–167, 1999.