

Νέες Μέθοδοι Εκπαίδευσης Τεχνητών Νευρωνικών Δικτύων, Βελτιστοποίησης και Εφαρμογές

Βασίλειος Π. Πλαγιανάκος

Διδακτορική Διατριβή

Πανεπιστήμιο Πατρών
Σχολή Θετικών Επιστημών
Τμήμα Μαθηματικών
Πάτρα

Επιβλέπων: Καθηγητής Μιχαήλ Ν. Βραχάτης

(Οκτώβριος 2002)

ΑΥΤΗ ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ ΣΤΟΙΧΕΙΟΘΕΤΗΘΗΚΕ ΜΕ ΤΟ ΠΡΟΓΡΑΜΜΑ $\LaTeX$ (ΔΙΑΝΟΜΕΣ MiKTeX ΚΑΙ TeX). Η ΣΥΓΓΡΑΦΗ ΕΓΙΝΕ ΜΕ ΤΗ ΒΟΗΘΕΙΑ ΤΩΝ ΠΡΟΓΡΑΜΜΑΤΩΝ WinEdt (ΣΤΟ ΛΕΙΤΟΥΡΓΙΚΟ ΣΥΣΤΗΜΑ $\text{Microsoft Windows NT}$) ΚΑΙ vi (ΣΤΟ ΛΕΙΤΟΥΡΓΙΚΟ ΣΥΣΤΗΜΑ Red Hat Linux). Η ΤΕΛΙΚΗ ΗΛΕΚΤΡΟΝΙΚΗ ΜΟΡΦΗ ($\text{Portable Document Format - PDF}$) ΔΗΜΙΟΥΡΓΗΘΗΚΕ ΜΕ ΤΟ ΠΡΟΓΡΑΜΜΑ $\text{PDF}\LaTeX$. ΓΙΑ ΤΗΝ ΑΝΑΠΤΥΞΗ ΚΑΙ ΤΟΝ ΕΛΕΓΧΟ ΤΩΝ ΠΡΟΓΡΑΜΜΑΤΩΝ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΕ ΚΥΡΙΩΣ ΤΟ ΠΡΟΓΡΑΜΜΑ Matlab . ΟΙ ΓΡΑΦΙΚΕΣ ΠΑΡΑΣΤΑΣΕΙΣ ΕΓΙΝΑΝ ΜΕ ΤΗ ΒΟΗΘΕΙΑ ΤΩΝ ΠΡΟΓΡΑΜΜΑΤΩΝ Matlab , Mathematica ΚΑΙ Origin , ΚΑΙ Η ΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΣΧΗΜΑΤΩΝ ΜΕ ΤΑ ΠΡΟΓΡΑΜΜΑΤΑ Adobe Photoshop , Gimp ΚΑΙ xv .

Στο εξώφυλλο απεικονίζονται ένας απλουστευμένος ανθρώπινος νευρώνας, το μοντέλο του τεχνητού νευρώνα και ένα πολυστρωματικό τεχνητό νευρωνικό δίκτυο με ένα κρυφό επίπεδο νευρώνων.

Στο δάσκαλό μου, στους γονείς μου
και στην Ελένη, για τη στήριξη και
την υπομονή τους.

Περίληψη

Κανείς δεν συνειδητοποιεί πόσο όμορφο
είναι να ταξιδεύει, μέχρι που γυρίζει
οπίτι του και κοιμάται στο παλιό,
γνώριμό του μαξιλάρι.

—Lin Yutang (1895-1976)

Η παρούσα διατριβή ασχολείται με την μελέτη και την εκπαίδευση Τεχνητών Νευρωνικών Δικτύων (ΤΝΔ) με μεθόδους Βελτιστοποίησης και τις εφαρμογές αυτών. Η παρουσίαση των επιμέρους θεμάτων και αποτελεσμάτων της διατριβής αυτής οργανώνεται ως εξής:

Στο Κεφάλαιο 1 παρέχουμε τους βασικούς ορισμούς και περιγράφουμε τη δομή και τη λειτουργία των ΤΝΔ. Στη συνέχεια, παρουσιάζουμε μια σύντομη ιστορική αναδρομή, αναφέρουμε μερικά από τα πλεονεκτήματα της χρήσης των ΤΝΔ και συνοψίζουμε τους κύριους τομείς όπου τα ΤΝΔ εφαρμόζονται. Τέλος, περιγράφουμε τις βασικές κατηγορίες μεθόδων εκπαίδευσης.

Το Κεφάλαιο 2 αφιερώνεται στη μαθηματική θεμελίωση της εκπαίδευσης ΤΝΔ. Περιγράφουμε τη γνωστή μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος (Backpropagation) και δίνουμε αποδείξεις σύγκλισης για μια κλάση μεθόδων εκπαίδευσης που χρησιμοποιούν μονοδιάστατες ελαχιστοποιήσεις. Στο τέλος του κεφαλαίου παρουσιάζουμε κάποια θεωρητικά αποτελέσματα σχετικά με την ικανότητα των ΤΝΔ να προσεγγίζουν άγνωστες συναρτήσεις.

Στο Κεφάλαιο 3 προτείνουμε μια νέα κλάση μεθόδων εκπαίδευσης ΤΝΔ και αποδεικνύουμε ότι αυτές έχουν την ιδιότητα της ευρείας σύγκλισης, δηλαδή συγκλίνουν σε ένα ελάχιστο της αντικειμενικής συνάρτησης σχεδόν από οποιαδήποτε αρχική συνθήκη. Τα αποτελέσματά μας δείχνουν ότι η προτεινόμενη τεχνική μπορεί να βελτιώσει οποιαδήποτε μέθοδο της κλάσης της οπισθοδρομικής διάδοσης του σφάλματος.

Στο επόμενο κεφάλαιο παρουσιάζουμε την γνωστή μέθοδο QuickProp και μελετάμε τις ιδιότητες σύγκλισής της. Με βάση το θεωρητικό αποτέλεσμα που προκύπτει, κατασκευάζουμε μια νέα τροποποίηση της μεθόδου QuickProp, που έχει την ιδιότητα της ευρείας σύγκλισης και βελτιώνει σημαντικά την κλασική QuickProp μέθοδο.

Στα επόμενα δύο κεφάλαια μελετάμε την εκπαίδευση ΤΝΔ με μεθόδους ολικής Βελτιστοποίησης. Πιο συγκεκριμένα, στο Κεφάλαιο 5 προτείνουμε και μελετάμε διεξοδικά μια νέα κλάση μεθόδων που είναι ικανές να εκπαιδεύσουν ΤΝΔ με περιορισμένα ακέραια βάρη. Στη συνέχεια, επεκτείνουμε τις μεθόδους αυτές έτσι ώστε να υλοποιούνται σε παράλληλους υπολογιστές και να εκπαιδεύουν ΤΝΔ με χρήση συναρτήσεων κατωφλίων.

Το Κεφάλαιο 6 πραγματεύεται την εφαρμογή γνωστών μεθόδων όπως οι Γενετικοί Αλγόριθμοι, η μέθοδος της προσομοιωμένης απόπτωσης (Simulated Annealing) και η μέθοδος βελτιστοποίησης με σμήνος σωματιδίων (Particle Swarm Optimization) στην εκπαίδευση ΤΝΔ. Επίσης, παρουσιάζουμε νέους μετασχηματισμούς της αντικειμενικής συνάρτησης με σκοπό της σταδιακή εξάλειψη των τοπικών ελαχίστων της.

Στο Κεφάλαιο 7 κάνουμε μια σύντομη ανασκόπηση της στοχαστικής μεθόδου της πιο απότομης κλίσης (stochastic gradient descent) για την εκπαίδευση ΤΝΔ ανά πρότυπο εισόδου και προτείνουμε μια νέα τέτοια μέθοδο. Η νέα μέθοδος συγκρίνεται με άλλες γνωστές μεθόδους και τα πειράματά μας δείχνουν ότι υπερτερεί.

Η παρουσίαση του ερευνητικού έργου για αυτή τη διατριβή ολοκληρώνεται με το Κεφάλαιο 8, όπου προτείνουμε και μελετάμε εκτενώς μη μονότονες μεθόδους εκπαίδευσης ΤΝΔ. Η τεχνική που προτείνουμε μπορεί να εφαρμοστεί σε κάθε μέθοδο της κλάσης της οπισθοδρομικής διάδοσης του σφάλματος, με αποτέλεσμα η τροποποιημένη μέθοδος να έχει την ικανότητα, πολλές φορές, να αποφεύγει τοπικά ελάχιστα της αντικειμενικής συνάρτησης.

Η παρουσίαση της διατριβής ολοκληρώνεται με το Κεφάλαιο 9 και δύο Παραρτήματα. Το Κεφάλαιο 9 περιέχει τα γενικά συμπεράσματα της διατριβής. Στο Παράρτημα Α παρουσιάζουμε συνοπτικά μερικά από τα προβλήματα εκπαίδευσης που εξετάσαμε στα προηγούμενα κεφάλαια και τέλος στο Παράρτημα Β δίνουμε την απόδειξη της μεθόδου της οπισθοδρομικής διάδοσης του σφάλματος.

Synopsis

This thesis investigates Optimization methods for Artificial Neural Network training and their applications. In the first two chapters we discuss the basic neural network definitions, well known network architectures and training methods, as well as the theoretical background that supports the development of new efficient and effective training algorithms.

In Chapter 3 a new generalized theoretical result is presented that underpins the development of first-order globally convergent batch training algorithms which employ local learning rates. This result allows us to equip the algorithms of this class with a strategy for adapting the direction of search to a descent one. In this way, a decrease of the batch-error measure at each training iteration is ensured, and convergence of the sequence of weight iterates to a local minimizer of the batch error function is obtained from remote initial weights. The effectiveness of the theoretical result is illustrated in application examples by comparing two well known training algorithms with local learning rates to their globally convergent modifications.

In Chapter 4, a mathematical framework for the convergence analysis of the well known Quickprop method is described. Furthermore, we propose a modification of this method that exhibits improved convergence speed and stability, while at the same time, alleviates the use of heuristic learning parameters. Simulations are conducted to compare and evaluate the performance of the new modified Quickprop algorithm with various popular training algorithms. The results of the experiments indicate that the increased convergence rate achieved by the proposed algorithm, by no means affects its generalization capability and stability.

In Chapter 5, evolutionary neural network training algorithms are presented. These algorithms are applied to train neural networks with weight values confined to a narrow band of integers. We also train the network using threshold activation functions. Furthermore, parallel evolutionary algorithms for integer weight neural network training are introduced. These algorithms have been designed keeping in mind that the resulting integer weights require less bits to be stored and the digital arithmetic operations between them are easier to be implemented in hardware. Another advantage of the proposed evolutionary strategies is that they are capable of continuing the training process “on-chip”, if needed. Our intention is to present results of evolutionary algorithms on this difficult task. Based on the application of the proposed class of methods on classical neural network problems, our experience is that these methods are effective and reliable.

In Chapter 6, we investigate the use of Global Optimization methods for neural network training. To this end, we present strategies for the development of globally convergent modifications of local search methods and we investigate further the use of global search methods for neural network training. The proposed methods tend to lead to desirable weight configurations and allow the network to learn the entire training set, and, in that sense, they improve the efficiency of the training process. Simulation experiments on some notorious for their local minima learning tasks are presented and an extensive comparison of several training algorithms is provided.

In Chapter 7 a method for adapting a global learning rate, i.e. a common learning rate for all the network weights, for on-line training is presented. The proposed technique belongs to the class of stochastic gradient descent methods and takes into consideration

previously computed pieces of information regarding the learning rate adaptation procedure. The proposed algorithm has been implemented and tested in various problems with large data sets and networks. Additionally, we propose a hybrid Evolutionary algorithm for on-line training, capable of training and retraining a neural network when the task is nonstationary.

Finally, in Chapter 8 we present deterministic nonmonotone learning strategies, i.e. deterministic training algorithms in which error function values are allowed to increase at some iterations. To this end, we argue that the current error function value must satisfy a nonmonotone criterion with respect to the maximum error function value of the M previous epochs, and we propose a subprocedure to dynamically compute M . The nonmonotone strategy can be incorporated in any batch training algorithm and provides fast, stable and reliable learning. Extensive experimental results in different classes of problems show that this approach improves the convergence speed and success percentage of first-order training algorithms, and alleviates the need for fine-tuning problem-depended heuristic parameters.

Keywords: Artificial Neural Networks, Training Algorithms, Batch Training, Online Training, Global Convergence, Global Optimization, Integer Weight Neural Networks, Networks with Threshold Activations, Nonmonotone Training, Parallel Implementations.

Ευχαριστίες

Η παρούσα διατριβή δεν θα μπορούσε εκπονηθεί χωρίς τη βοήθεια και την συμπαράσταση πολλών ανθρώπων. Αισθάνομαι πρωτίτως την ανάγκη να ευχαριστήσω θερμά τον Δάσκαλό μου, καθηγητή κ. Μ.Ν. Βραχάτη στον οποίο οφείλεται κατά ένα πολύ μεγάλο βαθμό η υλοποίηση της παρούσας διατριβής. Η ουσιαστική καθοδήγησή του στο ξεπέρασμα των ποικίλων δυσκολιών που συνάντησα κατά τη διάρκεια της έρευνας, οι πολύτιμες συμβουλές και υποδείξεις του, και η ηθική του συμπαράσταση με βοήθησαν τα μέγιστα. Ευχαριστώ επίσης και τα άλλα δύο μέλη της Τριμελούς Συμβουλευτικής Επιτροπής μου, τους καθηγητές κ.κ. Χ.Ε. Μπότσαρη και Αν. Μπούντη, των οποίων η βοήθεια ήταν επίσης καθοριστική.

Κατά τη διάρκεια της εκπόνησης της διατριβής μου είχα τη χαρά και την τιμή να συνεργαστώ με τον λέκτορα του Πανεπιστημίου Brunel της Μεγάλης Βρετανίας κ. Γ.Δ. Μαγουλά, τον οποίο και ευχαριστώ ιδιαίτερος. Τέλος, αισθάνομαι την υποχρέωση να ευχαριστήσω και τους υπόλοιπους συνεργάτες μου, μεταπτυχιακούς φοιτητές του Τμήματος Μαθηματικών του Πανεπιστημίου Πατρών, κ.κ. Ν.Κ. Νούση, Κ.Ε. Παρσόπουλο, Δ.Γ. Σωτηρόπουλο και Ε. Τζανάκη.

Ευχαριστίες επίσης απευθύνονται στον Διευθυντή και στο προσωπικό του Εργαστηρίου Ηλεκτρονικών Υπολογιστών και Πληροφορικής του Τμήματος Μαθηματικών για την συμπαράστασή τους και την διάθεση των απαραίτητων υπολογιστικών πόρων.

Η παρούσα διατριβή στηρίχθηκε οικονομικά από το πρόγραμμα ΥΠΕΡ' 97, της Γενικής Γραμματείας Έρευνας και Τεχνολογίας του Υπουργείου Ανάπτυξης. Στα πλαίσια αυτού του προγράμματος είχα την χαρά να συνεργαστώ με τον καθηγητή του Τμήματος Χημείας κ. Ν.Α. Κατσάνο και την ειδική επιστήμονα του Εθνικού Αρχαιολογικού Μουσείου Αθηνών κ. Ε. Μάγκου, τους οποίους ευχαριστώ θερμά για την βοήθειά τους και τη συμβολή τους στην επιτυχή αποπεράτωση του προγράμματος.

Βασίλειος Π. Πλαγιανάκος

Πάτρα, 2002.

Περιεχόμενα

Περίληψη	v
Ευχαριστίες	ix
I Εισαγωγή και Βασικές Έννοιες	1
1 Εισαγωγή	3
1.1 Από τα Βιολογικά στα Τεχνητά Νευρωνικά Δίκτυα	4
1.2 Ιστορική Αναδρομή	7
1.3 Εφαρμογές των Τεχνητών Νευρωνικών Δικτύων	9
1.4 Ιδιότητες των Τεχνητών Νευρωνικών Δικτύων	9
1.5 Εκπαίδευση των Τεχνητών Νευρωνικών Δικτύων	10
1.5.1 Η μορφολογία του χώρου των βαρών	11
1.5.2 Η αρχικοποίηση των βαρών	11
1.5.3 Κατηγορίες μεθόδων εκπαίδευσης	14
1.5.4 Παράλληλη εκπαίδευση Τεχνητών Νευρωνικών Δικτύων	14
2 Θεωρητικό Υπόβαθρο των Μεθόδων Εκπαίδευσης ΤΝΔ	17
2.1 Εισαγωγή	17
2.2 Η Επιλογή του Ρυθμού Εκπαίδευσης	19
2.3 Αλγόριθμοι Εκπαίδευσης με Ευρεία Σύγκλιση	19
2.4 Βελτιστοποίηση μη Γραμμικών Συναρτήσεων ανά Κατεύθυνση	20
2.4.1 Μελέτη σύγκλισης της σύνθετης μη γραμμικής μεθόδου Jacobi	21
2.4.2 Μελέτη σύγκλισης της σύνθετης μη γραμμικής μεθόδου SOR	22
2.4.3 Μελέτη σύγκλισης μιας τροποποίησης της μεθόδου του Powell	23
2.5 Πρακτική Θεώρηση της Σύγκλισης των Αλγορίθμων Εκπαίδευσης	25
2.6 Τα ΤΝΔ σαν Καθολικοί Προσεγγιστές	25
2.6.1 Θεωρήματα των Kolmogorov και Sprecher	26
II Μαθηματική Θεμελίωση Μεθόδων Εκπαίδευσης ΤΝΔ	27
3 Μαθηματική Θεμελίωση μιας Νέας Κλάσης Αλγορίθμων Ευρείας Σύγκλισης	29
3.1 Εισαγωγή	29
3.2 Μέθοδοι με Διαφορετικό Ρυθμό Εκπαίδευσης για κάθε Βάρος	30
3.3 Ευρεία Σύγκλιση Αλγορίθμων με Τοπικό Ρυθμό Εκπαίδευσης	31
3.4 Αποτελέσματα των Προσομοιώσεων	35
3.4.1 Αναγνώριση αριθμών	36
3.4.2 Προσέγγιση μιας συνεχούς συνάρτησης	38
3.4.3 Αναγνώριση ανωμαλιών σε κολοנוσκοπήσεις	39
3.5 Συμπεράσματα – Συνεισφορά	39

4 Μαθηματική Θεμελίωση της Μεθόδου Quickprop και μια Νέα Τροποποίησή της	41
4.1 Εισαγωγή	41
4.2 Μέθοδοι Χορδής	42
4.3 Η Μέθοδος Quickprop	43
4.4 Αλγόριθμοι Ευρείας Σύγκλισης με Προσαρμοστικό Ρυθμό Εκπαίδευσης . . .	44
4.5 Η Τροποποιημένη Μέθοδος Quickprop	45
4.6 Πειραματικά Αποτελέσματα	47
4.6.1 Αποκλειστικό-ΕΙΤΕ	47
4.6.2 Ταξινόμηση υψής	48
4.6.3 Αναγνώριση αριθμών	49
4.7 Συμπεράσματα - Συνεισφορά	49
 III Μέθοδοι Ολικής Βελτιστοποίησης για την Εκπαίδευση ΤΝΔ	 51
5 Εκπαίδευση Τεχνητών Νευρωνικών Δικτύων με Ακέραια Βάρη	53
5.1 Εισαγωγή	53
5.2 Εκπαίδευση με Διαφοροεξελικτικούς Αλγόριθμους	53
5.2.1 Ο τελεστής μετάλλαξης	54
5.2.2 Ο τελεστής ανασυνδυασμού	55
5.2.3 Αποτελέσματα εκπαίδευσης με ακέραια βάρη	55
5.3 Εκπαίδευση με Περιορισμένα Ακέραια Βάρη	58
5.3.1 Αποτελέσματα εκπαίδευσης με περιορισμένα ακέραια βάρη	58
5.4 Εκπαίδευση με Χρήση Συναρτήσεων Ενεργοποίησης με Κατώφλια	60
5.4.1 Αποτελέσματα εκπαίδευσης με συναρτήσεις ενεργοποίησης με κατώφλια	61
5.5 Εκπαίδευση με Παράλληλους ΔΕΑ	62
5.5.1 Η γήρανση του πληθυσμού	63
5.5.2 Αποτελέσματα εκπαίδευσης με παράλληλους ΔΕΑ	63
5.6 Μελέτη της Γενίκευσης	65
5.7 Συμπεράσματα - Συνεισφορά	67
 6 Εκπαίδευση με Μεθόδους Αποφυγής Τοπικών Ελαχίστων	 71
6.1 Εισαγωγή	71
6.2 Μέθοδοι Ολικής Βελτιστοποίησης για την Εκπαίδευση ΤΝΔ	73
6.2.1 Η μέθοδος της προσομοιωμένης ανόπτησης	73
6.2.2 Γενετικοί αλγόριθμοι	74
6.2.3 Η μέθοδος βελτιστοποίησης με σμήνος σωματιδίων	75
6.3 Μετασχηματισμοί της Συνάρτησης Σφάλματος	77
6.3.1 Η τεχνική της παρεκκλίνουσας τροχιάς	77
6.3.2 Η τεχνική του «εφελκυσμού» της αντικειμενικής συνάρτησης	79
6.4 Αποτελέσματα	81
6.5 Συμπεράσματα - Συνεισφορά	82
 IV Μέθοδοι μη Μονότονης Εκπαίδευσης ΤΝΔ	 85
7 Εκπαίδευση ανά Πρότυπο Εισόδου	87
7.1 Εισαγωγή	87
7.2 Αλγόριθμοι Εκπαίδευσης ανά Πρότυπο Εισόδου	89
7.3 Προσομοιώσεις και Αποτελέσματα	90
7.3.1 Αποκλειστικό-ΕΙΤΕ	91
7.3.2 Αναγνώριση αριθμών	92

7.3.3 Αναγνώριση των κεφαλαίων γραμμάτων	92
7.4 Υβριδικές Μέθοδοι για την Επανεκπαίδευση ΤΝΔ	92
7.4.1 Το πρόβλημα ταξινόμησης υφής	93
7.4.2 Το πρόβλημα αναγνώρισης ανωμαλιών σε κολονοσκοπήσεις	93
7.5 Συμπεράσματα – Συνεισφορά	96
8 Μη Μονότονοι Αλγόριθμοι Εκπαίδευσης	97
8.1 Μη Μονότονες Στρατηγικές Εκπαίδευσης	97
8.1.1 Ο μη μονότονος ορίζοντας εκπαίδευσης	99
8.1.2 Ανάπτυξη μη μονότονων αλγορίθμων εκπαίδευσης	100
8.1.3 Μοντέλο αλγόριθμου με μεταβλητό ρυθμό εκμάθησης με χρήση της μη μονότονης στρατηγικής	101
8.2 Πειραματικά Αποτελέσματα	102
8.2.1 Συγκριτική μελέτη	102
8.2.2 Μελέτη της επίδρασης του μη μονότονου ορίζοντα εκπαίδευσης	107
8.2.3 Μελέτη της επίδρασης της μη μονότονης στρατηγικής	110
8.2.4 Αποτελέσματα γενίκευσης	113
8.3 Συμπεράσματα – Συνεισφορά	114
V Συμπεράσματα – Παραρτήματα – Βιβλιογραφία – Ευρετήριο	117
9 Συμπεράσματα Διατριβής	119
A Προβλήματα Εκπαίδευσης Νευρωνικών Δικτύων	123
A.1 Αποκλειστικό-ΕΠΕ (XOR)	123
A.2 Ισοτιμία των 3-bit	123
A.3 4-2-4 Κωδικοποιητής/Αποκωδικοποιητής	124
A.4 Το πρόβλημα γενίκευσης MONK	125
A.5 Προσέγγιση μιας συνεχούς συνάρτησης	125
A.6 Αναγνώριση των κεφαλαίων γραμμάτων	125
A.7 Αναγνώριση αριθμών	125
A.8 Ταξινόμηση υφής	126
B Απόδειξη της Μεθόδου Οπισθοδρομικής Διάδοσης του Σφάλματος	129
Βιβλιογραφία	131
Κατάλογος Δημοσιεύσεων Υποψηφίου	143
Ευρετήριο	149

Κατάλογος Σχημάτων

1.1	Απλουστευμένο μοντέλο ενός βιολογικού νευρώνα	5
1.2	Η υπερβολική εφαπτομένη για διάφορες τιμές της παραμέτρου λ	6
1.3	Μοντέλο ενός βιολογικού νευρώνα	6
1.4	Μοντέλο ενός τεχνητού νευρώνα	7
1.5	Ένα πολυστρωματικό πλήρως διασυνδεδεμένο πρόσθιας τροφοδότησης ΤΝΔ .	8
1.6	Παράδειγμα γραφικής παράστασης του χώρου των βαρών ενός ΤΝΔ με ένα μόνο νευρώνα όταν η τιμή της παραμέτρου της σιγμοειδούς είναι $\lambda = 1$ (αριστερά) και η ίδια γραφική παράσταση για $\lambda = 0.1$	11
1.7	Δύο σύνολα προτύπων εκπαίδευσης Α και Β. Τα σύνολα είναι γραμμικώς διαχωρίσιμα (αριστερά). Τα σύνολα δεν είναι γραμμικώς διαχωρίσιμα (δεξιά) . .	12
1.8	Παράδειγμα γραφικής παράστασης του χώρου των βαρών ενός ΤΝΔ με ένα μόνο νευρώνα, όταν το σύνολο εκπαίδευσης δεν είναι γραμμικώς διαχωρίσιμο	12
1.9	Παράδειγμα γραφικής παράστασης του χώρου των βαρών ενός ΤΝΔ με ένα μόνο νευρώνα, όταν το σύνολο εκπαίδευσης αποτελείται από 1, 2, 3 και 4 πρότυπα (δεξιόστροφα ξεκινώντας από επάνω αριστερά)	13
1.10	Επιτάχυνση του χρόνου εκπαίδευσης ανάλογα με το πλήθος των επεξεργασιών	16
3.1	Απεικόνιση της μεθόδου Quickprop για την εκπαίδευση ενός απλού ΤΝΔ με δύο βάρη (με x σημειώνεται το ελάχιστο). Η τροποποιημένη μέθοδος συγκλίνει στο επιθυμητό ελάχιστο (αριστερά), ενώ η κλασική μέθοδος συγκλίνει σε ένα ανεπιθύμητο ακρότατο (δεξιά)	36
3.2	Τυπική γραφική παράσταση της μείωσης των τιμών της συνάρτησης σφάλματος για το πρόβλημα της ισοτιμίας των 3-bit, ξεκινώντας από το ίδιο αρχικό σημείο. Με συνεχή γραμμή βλέπουμε την μέθοδο Quickprop και με διακεκομμένη γραμμή την ευρέως συγκλίνουσα τροποποίησή της	37
3.3	Τυπική γραφική παράσταση της μείωσης των τιμών της συνάρτησης σφάλματος για το πρόβλημα της αναγνώρισης αριθμών, ξεκινώντας από το ίδιο αρχικό σημείο. Με συνεχή γραμμή βλέπουμε την μέθοδο Silva-Almeida και με διακεκομμένη γραμμή την ευρέως συγκλίνουσα τροποποίησή της	39
3.4	Οι εικόνες για το πρόβλημα αναγνώρισης ανωμαλιών σε κολονοσκοπήσεις (αριστερά). Το ποσοστό επιτυχίας σε σχέση με διάστημα αρχικοποίησης των βαρών $(-a, a)$, (δεξιά).	40
5.1	Η επίδραση της παραμέτρου λ στην μορφή μιας σιγμοειδούς συνάρτησης ενεργοποίησης	61
6.1	Περιγραφή ενός απλού Γενετικού Αλγόριθμου	75
6.2	Γραφική παράσταση της συνάρτησης Six Hump Camel Back	78
6.3	Εφαρμόζοντας την τεχνική της παρεκκλίνουσας τροχιάς στη συνάρτηση Six Hump Camel Back, για $\lambda = 1.5$ και $\lambda = 10$	78
6.4	Εφαρμόζοντας την μέθοδο της παρεκκλίνουσας τροχιάς στην εκπαίδευση ΤΝΔ (με x σημειώνουμε το ελάχιστο της συνάρτησης E)	79
6.5	Γραφική παράσταση της συνάρτησης Levy No. 5	81

6.6	Γραφική παράσταση της συνάρτησης Levy No. 5 μετά από το πρώτο στάδιο (αριστερά) και μετά από το δεύτερο στάδιο (δεξιά) του προτεινόμενου μετασχηματισμού	81
7.1	Διαδοχικές εικόνες από ενδοσκόπιο	95
8.1	Το αποκλειστικό-ΕΙΤΕ: (α) η μη μονότονη συμπεριφορά της μεθόδου NMBBP και (β) η συμπεριφορά του προσαρμοστικού ρυθμού εκπαίδευσης	101
8.2	Το πρόβλημα της ισοτιμίας 3-bit: μέσος χρόνος (CPU time) για τη σύγκλιση κάθε αλγόριθμου	103
8.3	Το πρόβλημα προσέγγισης μιας συνεχούς συνάρτησης: μέσος χρόνος (CPU time) για τη σύγκλιση κάθε αλγόριθμου	103
8.4	Το πρόβλημα αναγνώρισης των κεφαλαίων γραμμάτων: μέσος χρόνος (CPU time) για τη σύγκλιση κάθε αλγόριθμου	104
8.5	Το πρόβλημα αναγνώρισης των αριθμών: μέσος χρόνος (CPU time) για τη σύγκλιση κάθε αλγόριθμου	104
8.6	Το πρόβλημα αναγνώρισης των κεφαλαίων γραμμάτων: Μέσος αριθμός υπολογισμών της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της, για τη μέθοδο NMBBP για διάφορες τιμές του M	108
8.7	Το πρόβλημα ταξινόμησης υφής: Μέσος αριθμός υπολογισμών της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της, για τη μέθοδο NMBBP για διάφορες τιμές του M	108
8.8	Το πρόβλημα αναγνώρισης των κεφαλαίων γραμμάτων: Μέσος αριθμός υπολογισμών της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της, για τη μέθοδο NMBPVS για διάφορες τιμές του M	109
8.9	Το πρόβλημα ταξινόμησης υφής: Μέσος αριθμός υπολογισμών της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της, για τη μέθοδο NMBPVS για διάφορες τιμές του M	109
8.10	Το πρόβλημα ταξινόμησης υφής: Ποσοστό επιτυχίας για τη μέθοδο NMBBP για διάφορες τιμές του M	110
8.11	Το αποκλειστικό-ΕΙΤΕ: η συμπεριφορά της μεθόδου NMBPM με προσαρμοζόμενο M^k	111
8.12	Το πρόβλημα αναγνώρισης των κεφαλαίων γραμμάτων: συμπεριφορά σύγκλισης της μεθόδου NMBPVS	112
8.13	Το πρόβλημα αναγνώρισης των αριθμών: ο αριθμός των εκπαιδευμένων ΤΝΔ (από σύνολο 100), που ταξινομούν σωστά τους αριθμούς 0, 1, ..., 9	115
A.1	Γραφική παράσταση της συνάρτησης $f(x) = \sin(x) \cos(2x)$ και τα 20 πρότυπα εκπαίδευσης	126
A.2	Η κωδικοποίηση για το γράμμα Άλφα	127
A.3	Η κωδικοποίηση για τον αριθμό 6	127
A.4	Οι 12 εικόνες υφής που χρησιμοποιήθηκαν για την εξαγωγή των προτύπων εκπαίδευσης	128

Κατάλογος Πινάκων

3.1	Αποτελέσματα από το πρόβλημα αναγνώρισης αριθμών ($E \leq 10^{-1}$)	38
3.2	Αποτελέσματα από το πρόβλημα αναγνώρισης αριθμών ($E \leq 10^{-2}$)	38
3.3	Αποτελέσματα από το πρόβλημα προσέγγισης μιας συνεχούς συνάρτησης . .	40
4.1	Αποτελέσματα από το πρόβλημα του αποκλειστικό-ΕΙΤΕ	48
4.2	Αποτελέσματα από το πρόβλημα ταξινόμησης υφής	48
4.3	Αποτελέσματα από το πρόβλημα αναγνώρισης των αριθμών	49
5.1	Αποτελέσματα από το πρόβλημα του αποκλειστικού ΕΙΤΕ	57
5.2	Αποτελέσματα από το πρόβλημα της ισοτιμίας 3-bit	57
5.3	Αποτελέσματα από το πρόβλημα του 4-2-4 Κωδικοποιητή/Αποκωδικοποιητή	58
5.4	Αποτελέσματα εκπαίδευσης με περιορισμένα βάρη (Αποκλειστικό-ΕΙΤΕ) . . .	59
5.5	Αποτελέσματα εκπαίδευσης με περιορισμένα βάρη (Ισοτιμία 3-bit)	60
5.6	Αποτελέσματα εκπαίδευσης με κατώφλια (Αποκλειστικό-ΕΙΤΕ)	62
5.7	Αποτελέσματα εκπαίδευσης με κατώφλια (Ισοτιμία 3-bit)	62
5.8	Αποτελέσματα ΠΔΕΑ με σιγμοειδείς συναρτήσεις ενεργοποίησης (Αποκλειστικό-ΕΙΤΕ)	64
5.9	Αποτελέσματα ΠΔΕΑ με κατώφλια (Αποκλειστικό-ΕΙΤΕ)	64
5.10	Αποτελέσματα ΠΔΕΑ με σιγμοειδείς συναρτήσεις ενεργοποίησης (Ισοτιμία 3-bit)	65
5.11	Αποτελέσματα ΠΔΕΑ με κατώφλια (Ισοτιμία 3-bit)	65
5.12	Αποτελέσματα ΠΔΕΑ με σιγμοειδείς συναρτήσεις ενεργοποίησης (4-2-4 Κωδικοποιητή/Αποκωδικοποιητή)	65
5.13	Αποτελέσματα ΠΔΕΑ με κατώφλια (4-2-4 Κωδικοποιητή/Αποκωδικοποιητή)	66
5.14	Σύγκριση της γενίκευση στα προβλήματα MONK	66
5.15	Η τοπολογία των δικτύων για τα προβλήματα MONK	67
5.16	Ακέραια βάρη και πολώσεις για το πρόβλημα MONK-1	68
5.17	Ακέραια βάρη και πολώσεις για το πρόβλημα MONK-2	69
5.18	Ακέραια βάρη και πολώσεις για το πρόβλημα MONK-3	70
6.1	Συγκριτικά αποτελέσματα	83
7.1	Αποτελέσματα από το πρόβλημα του αποκλειστικού ΕΙΤΕ	91
7.2	Αποτελέσματα από το πρόβλημα αναγνώρισης αριθμών	92
7.3	Αποτελέσματα από το πρόβλημα αναγνώρισης κεφαλαίων γραμμάτων	93
7.4	Αποτελέσματα από το πρόβλημα αναγνώρισης ανωμαλιών σε κολονοσκοπήσεις	96
8.1	Αποτελέσματα από το πρόβλημα της ισοτιμίας 3-bit	105
8.2	Αποτελέσματα από το πρόβλημα προσέγγισης μιας συνεχούς συνάρτησης . .	106
8.3	Αποτελέσματα από το πρόβλημα αναγνώρισης γραμμάτων	106
8.4	Αποτελέσματα από το πρόβλημα αναγνώρισης αριθμών	107
8.5	Αποτελέσματα από το πρόβλημα αναγνώρισης αριθμών	110
8.6	Αποτελέσματα από το πρόβλημα του αποκλειστικού ΕΙΤΕ	111
8.7	Αποτελέσματα από το πρόβλημα ταξινόμησης υφής	111
8.8	Αποτελέσματα από το πρόβλημα της προσέγγισης μιας συνεχούς συνάρτησης	112
8.9	Αποτελέσματα από το πρόβλημα της προσέγγισης μιας συνεχούς συνάρτησης	113

8.10	Αποτελέσματα από το πρόβλημα αναγνώρισης των γραμμάτων	113
8.11	Αποτελέσματα από το πρόβλημα γενίκευσης MONK	114
A.1	Τα πρότυπα εκπαίδευσης του αποκλειστικού-ΕΙΤΕ	124
A.2	Τα πρότυπα εκπαίδευσης της ισοτιμίας 3-bit	124
A.3	Τα πρότυπα εκπαίδευσης του 4-2-4 Κωδικοποιητή/Αποκωδικοποιητή	124
A.4	Ο δυαδικός πίνακας κωδικοποίησης για το γράμμα Άλφα	126
A.5	Ο δυαδικός πίνακας κωδικοποίησης για τον αριθμό 6	127

Μέρος I

Εισαγωγή και Βασικές Έννοιες

Εισαγωγή

Όταν κάποιος ανακαλύψει την αλήθεια για κάτι με μεγάλο κόπο,
τότε, επιθεωρώντας πιο προσεκτικά την ανακάλυψή του,
συχνά διαπιστώνει ότι αυτό που τον κούρασε πολύ για να βρεθεί
θα μπορούσε να παρατηρηθεί με την μεγαλύτερη ευκολία.

—Galileo Galilei (1564–1642)

Με τον όρο *Τεχνητό Νευρωνικό Δίκτυο* (ΤΝΔ) αποκαλούμε ένα μαθηματικό μοντέλο αποτελούμενο από ένα μεγάλο αριθμό ανεξάρτητων υπολογιστικών στοιχείων, που ονομάζονται νευρώνες (neurons), τα οποία διασυνδέονται μεταξύ τους και είναι οργανωμένα σε στρώματα (layers). Με άλλα λόγια, ένα ΤΝΔ είναι ένας μαζικά παράλληλος κατανεμημένος επεξεργαστής, που έχει την έμφυτη ιδιότητα να αποθηκεύει εμπειρική γνώση και να την έχει διαθέσιμη για χρήση στο μέλλον. Τα ΤΝΔ προσομοιάζουν τον ανθρώπινο εγκέφαλο σε δύο σημεία: (α) η γνώση του ΤΝΔ αποκτάται μέσω μιας διαδικασίας μάθησης, και (β) οι σύνδεσμοι μεταξύ των νευρώνων (που ονομάζονται *συντελεστές βάρους* ή απλά *βάρη*) χρησιμοποιούνται για να αποθηκευτεί αυτή η γνώση. Η διαδικασία για να επιτύχουμε την εκπαίδευση του ΤΝΔ, ονομάζεται *αλγόριθμος εκπαίδευσης* [47].

Στην πραγματικότητα βέβαια, τα ΤΝΔ είναι πολύ απλούστερα από τα βιολογικά. Τα ΤΝΔ παρέχουν ένα εναλλακτικό αλγοριθμικό μοντέλο, το οποίο είναι εμπνευσμένο από τα βιολογικά μοντέλα, σύμφωνα με το οποίο οι υπολογισμοί γίνονται παράλληλα και μαζικά, και η εκπαίδευση αντικαθιστά την ανάπτυξη προγράμματος.

Από τους παραπάνω ορισμούς γίνεται φανερό ότι τα ΤΝΔ είναι μια νέα τεχνική για επεξεργασία πληροφοριών. Μπορούμε να πούμε ότι αποτελούν προσπάθεια προσομοίωσης, με τη βοήθεια υπολογιστών, του ανθρώπινου νευρικού συστήματος και λειτουργούν εντελώς διαφορετικά από τις συνήθεις μεθόδους. Μερικά παραδείγματα θα αποσαφηνίσουν τη διαφορά τους από τις συμβατικές υπολογιστικές μεθόδους.

Ας υποθέσουμε ότι θέλουμε να ερμηνεύσουμε το παρακάτω σχήμα:

Το κα_ό το παλ_κάρ_,
ξ_ρει κ_1 άλ_ο μο_οπ_τι. Σ_ου κου_ού τ_ν π_ρτα,
_σο θέ_εις β_ώντ_. Όσ_ δε φτ_νει η αλε_ού
τ_ κά_ει κρε_αστ_ρια.

Αν προσπαθήσουμε διαβάζοντας σειριακά από τα αριστερά προς τα δεξιά ένα γράμμα τη φορά, όπως θα έκανε ένας υπολογιστής, είναι πολύ δύσκολο αν όχι αδύνατο να καταλάβουμε το νόημα. Αν όμως κοιτάζουμε όλο το σχήμα θα καταλάβουμε εύκολα ότι αποτελείται από τρία τμήματα και με λίγη προσπάθεια θα βρούμε τις τρεις γνωστές παροιμίες. Έτσι φαίνεται ότι μπορέσαμε να κάνουμε κάτι που και οι πιο ισχυροί υπολογιστές δυσκολεύονται ή αποτυγχάνουν αν δεν έχουν τις ειδικές συντακτικές γνώσεις και την εμπειρία που διαθέτουν οι άνθρωποι. Το ανθρώπινο μυαλό, αν και σαφώς πιο αργό από έναν σύγχρονο υπολογιστή, υπερίσχυσε.

Πριν δώσουμε απάντηση για ποιο λόγο αυτό συμβαίνει ας δούμε ένα ακόμα παράδειγμα. Ο ανθρώπινος εγκέφαλος για να εκτελέσει αναγνώριση προσώπων, δηλαδή να αναγνωρίσει γνωστά μας πρόσωπα σε άγνωστα περιβάλλοντα, χρειάζεται περίπου 100–200 χιλιοστά του δευτερολέπτου, ενώ ένας πολύ ισχυρός σύγχρονος ηλεκτρονικός υπολογιστής θα χρειαζόταν ίσως ολόκληρες ημέρες για να εκτελέσει μια τέτοια λειτουργία [24].

Και πάλι ο ανθρώπινος εγκέφαλος υπερσχύει. Υπάρχουν δύο λόγοι για αυτό: η δυνατότητα *μάθησης* και η *παράλληλη επεξεργασία* των δεδομένων. Η ικανότητα του ανθρώπινου εγκεφάλου να φτιάχνει από πολύ μικρή ηλικία και διαρκώς να διορθώνει κανόνες (αυτό που συνήθως ονομάζουμε εμπειρία), καθώς και η δυνατότητά του για μαζική παράλληλη επεξεργασία δεδομένων, του δίνει πλεονέκτημα απέναντι στους σημερινούς σειριακούς υπολογιστές σε πολλά προβλήματα που απαιτούν λύση σε πραγματικό χρόνο (real time).

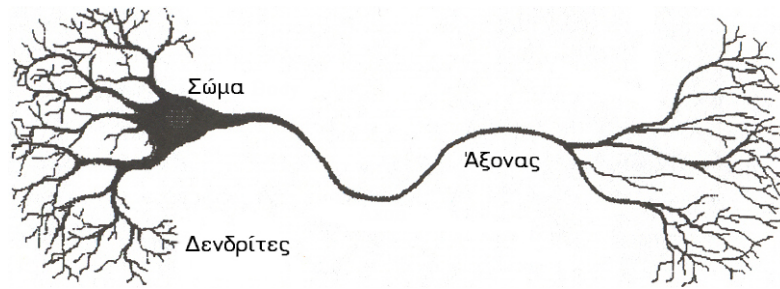
Έτσι η εξομοίωση του παράλληλου ανθρώπινου εγκεφάλου (έστω και με τη χρήση σειριακών υπολογιστών), έχει πολλά να προσφέρει στην επίλυση καθημερινών προβλημάτων, που με τις συνήθεις μεθόδους φαίνονται άλυτα. Γι' αυτό και σήμερα με τη βοήθεια της Βιολογίας, των Γνωστικών Επιστημών και της Νευροεπιστήμης προσπαθούμε να κατανοήσουμε όσο το δυνατόν καλύτερα τον ανθρώπινο εγκέφαλο, που αποτελεί τον τελειότερο παράλληλο επεξεργαστή που ξέρουμε. Ελπίζουμε έτσι να χρησιμοποιήσουμε πολλά από αυτά τα στοιχεία για να φτιάξουμε τον μελλοντικό παράλληλο υπολογιστή, που πιθανά θα βασίζεται στην παράλληλη επεξεργασία των ΤΝΔ.

1.1 Από τα Βιολογικά στα Τεχνητά Νευρωνικά Δίκτυα

Πριν δούμε την λειτουργία και περιγράψουμε ένα ΤΝΔ, ας δούμε πώς λειτουργεί ο ανθρώπινος εγκέφαλος. Τα Βιολογικά Νευρωνικά Δίκτυα που υπάρχουν στον ανθρώπινο εγκέφαλο αποτελούνται από *νευρώνες*. Ο νευρώνας είναι το μικρότερο τμήμα του εγκεφάλου που είναι ικανό να επεξεργαστεί πληροφορίες και η ύπαρξή του διαφοροποιεί τα ζώα από τα φυτά (τα φυτά δεν έχουν νευρώνες). Τυπικά, η επεξεργασία στους νευρώνες γίνεται 5 με 6 τάξεις μεγέθους πιο αργά από ότι στις σύγχρονες ψηφιακές λογικές πύλες. Ο χρόνος για τις ψηφιακές λογικές πύλες μετριέται σε δισεκατομμυριοστά του δευτερολέπτου (nanoseconds), ενώ στους νευρώνες σε χιλιοστά του δευτερολέπτου (milliseconds). Όμως, ο εγκέφαλος αντισταθμίζει τη σχετικά αργή ταχύτητα λειτουργίας των νευρώνων με τον πραγματικά τεράστιο αριθμό τους και τον τεράστιο αριθμό των μεταξύ τους συνδέσεων. Υπολογίζεται ότι υπάρχουν 10 δισεκατομμύρια νευρώνες και 60 τρισεκατομμύρια συνδέσεις στον φλοιό του ανθρώπινου εγκεφάλου [143]. Οι νευρώνες αποτελούνται από 3 βασικά τμήματα, όπως φαίνεται και στο Σχήμα 1.1. Αυτά είναι: (α) το σώμα, (β) ο άξονας, και (γ) οι δενδρίτες.

Πιο αναλυτικά, θα λέγαμε ότι οι *δενδρίτες* λαμβάνουν σήματα από γειτονικούς νευρώνες. Τα σήματα αυτά είναι ηλεκτρικοί παλμοί που διαδίδονται μεταξύ του άξονα του νευρώνα πομπού και των δενδριτών του νευρώνα δέκτη, με τη βοήθεια χημικών διεργασιών. Το σημείο των χημικών διεργασιών, όπου ο άξονας ενός νευρώνα μεταδίδει το σήμα στους δενδρίτες του επόμενου ονομάζεται *σύναψη*. Πρέπει να σημειώσουμε ότι αυτές οι διεργασίες μεταβάλλουν τα εισερχόμενα σήματα, αλλάζοντας συνήθως την συχνότητά τους. Στη συνέχεια το *σώμα* αθροίζει τα εισερχόμενα σήματα και όταν αρκετά σήματα έχουν ληφθεί αποστέλλει το επεξεργασμένο σήμα στους γειτονικούς του νευρώνες. Η μετάδοση του σήματος γίνεται μέσω του *άξονα*. Έτσι, κάθε νευρώνας δέχεται πολλά σήματα σαν είσοδο και μετά την επεξεργασία τους μεταδίδει μόνο ένα σε όλους τους νευρώνες με τους οποίους συνδέεται.

Σειρά έχει τώρα να δούμε πώς από αυτό το απλουστευμένο βιολογικό μοντέλο περνάμε στα Τεχνητά Νευρωνικά Δίκτυα. Οι κόμβοι ή τεχνητοί νευρώνες ή απλά νευρώνες στα ΤΝΔ θεωρούνται συνήθως ως απλουστευμένα πρότυπα των βιολογικών νευρώνων και στις μεταξύ τους συνδέσεις αντιστοιχεί ένας πραγματικός αριθμός, που ονομάζεται *συντελεστής βάρους* ή απλά *βάρος* και χρησιμοποιείται (όπως και οι συνάψεις μεταξύ των ανθρώπινων νευρώνων) για την τροποποίηση των εισόδων του νευρώνα. Συνήθως, κάθε νευρώνας (εκτός από τους



Σχήμα 1.1: Απλουστευμένο μοντέλο ενός βιολογικού νευρώνα

νευρώνες εισόδου) θεωρούμε ότι έχει ακόμα μία σύνδεση που έχει ένα βάρος που ονομάζεται *πόλωση* ή *μεροληψία* (bias) και σταθερή είσοδο με την τιμή 1. Η χρήση της πόλωσης βοηθά το ΤΝΔ να έχει καλύτερη ικανότητα ταξινόμησης. Στα επόμενα σε όλα τα ΤΝΔ που θα μελετήσουμε θα χρησιμοποιούμε πολώσεις και αναφερόμενοι στα βάρη θα εννοούμε και τις πολώσεις.

Η εμπειρία και η γνώση που αποκτά το ΤΝΔ αποθηκεύεται στα βάρη του. Η προσαρμογή των τιμών των βαρών έτσι ώστε το ΤΝΔ να έχει την επιθυμητή απόκριση ονομάζεται *εκπαίδευση* και γίνεται με τη βοήθεια διαφόρων *αλγορίθμων εκπαίδευσης*.

Κάθε νευρώνας βρίσκεται σε μια εσωτερική κατάσταση που ονομάζεται επίπεδο ενεργοποίησης, που αποτελεί τη έξοδο του νευρώνα και εξαρτάται από τις εισόδους που λαμβάνονται. Τονίζουμε ότι κάθε νευρώνας στέλνει μόνο ένα σήμα κάθε φορά στους γειτονικούς του νευρώνες. Τα περισσότερα ΤΝΔ για να υπολογίσουν την ενεργοποίησή τους υπολογίζουν το γινόμενο κάθε εισόδου επί το αντίστοιχο βάρος και αθροίζουν όλα αυτά τα γινόμενα. Τελικά η ενεργοποίηση είναι η εικόνα του αποτελέσματος μέσω μιας συνεχούς συνάρτησης (*συνάρτηση ενεργοποίησης*). Ο ρόλος της συνάρτησης ενεργοποίησης συνήθως είναι να περιορίσει την ενεργοποίηση μέσα σε κάποιο επιθυμητό διάστημα. Οι πιο γνωστές συναρτήσεις ενεργοποίησης είναι οι ακόλουθες:

- Η δυαδική συνάρτηση με κατώφλι θ :

$$f_1(x) = \begin{cases} 1, & x \geq \theta \\ 0, & x < \theta \end{cases}$$

- Η διπολική συνάρτηση με κατώφλι θ :

$$f_2(x) = \begin{cases} 1, & x \geq \theta \\ -1, & x < \theta \end{cases}$$

- Η ταυτοτική συνάρτηση:

$$f_3(x) = x, \quad \forall x \in \mathbb{R}$$

- Η λογιστική συνάρτηση:

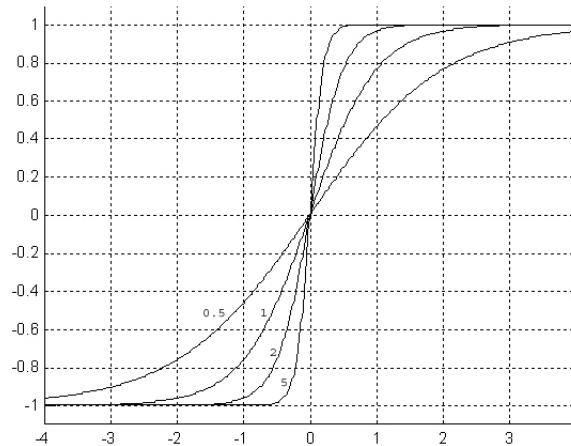
$$f_4(x) = \frac{1}{1 + e^{-\lambda_1 x}}$$

- Η υπερβολική εφαπτομένη:

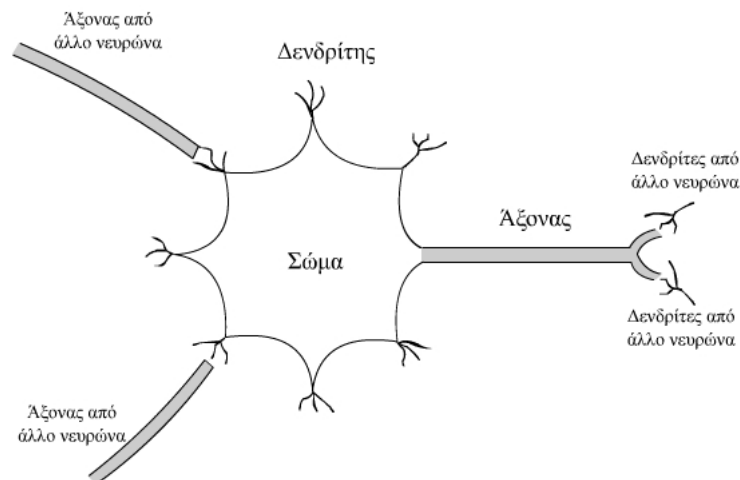
$$f_5(x) = \tanh(\lambda_2 x)$$

Οι συναρτήσεις ενεργοποίησης f_1 και f_2 περιορίζουν την ενεργοποίηση του νευρώνα στις τιμές $\{0, 1\}$ και $\{-1, 1\}$, αντίστοιχα. Οι f_4 και f_5 περιορίζουν την έξοδο του νευρώνα στο διάστημα $(0, 1)$ και $(-1, 1)$, αντίστοιχα. Τέλος η f_3 επιτρέπει αυθαίρετα μεγάλες ή μικρές ενεργοποιήσεις. Οι παράμετροι λ_1 και λ_2 στις συναρτήσεις f_4 και f_5 αντίστοιχα ρυθμίζουν την μορφή της σιγμοειδούς. Όσο μεγαλύτερες τιμές παίρνουν οι παράμετροι αυτές τόσο πιο απότομη γίνεται η σιγμοειδής και προσεγγίζει τον κατακόρυφο άξονα. Συνήθως στην πράξη επιλέγουμε τις τιμές $\lambda_1 = 1$ και $\lambda_2 = 1/2$. Στο Σχήμα 1.2 απεικονίζεται

η συνάρτηση ενεργοποίησης f_5 για διάφορες τιμές της παραμέτρου λ_2 . Η επιλογή της κατάλληλης συνάρτησης ενεργοποίησης γίνεται ανάλογα με το είδος των προτύπων (δυναμικά, διπολικά, πραγματικοί αριθμοί κτλ.) και είναι κρίσιμη για την εκπαίδευση. Στα Σχήματα 1.3 και 1.4 απεικονίζονται ένας βιολογικός και ένας τεχνητός νευρώνας και δίνεται ο τύπος υπολογισμού της ενεργοποίησης του τεχνητού νευρώνα.



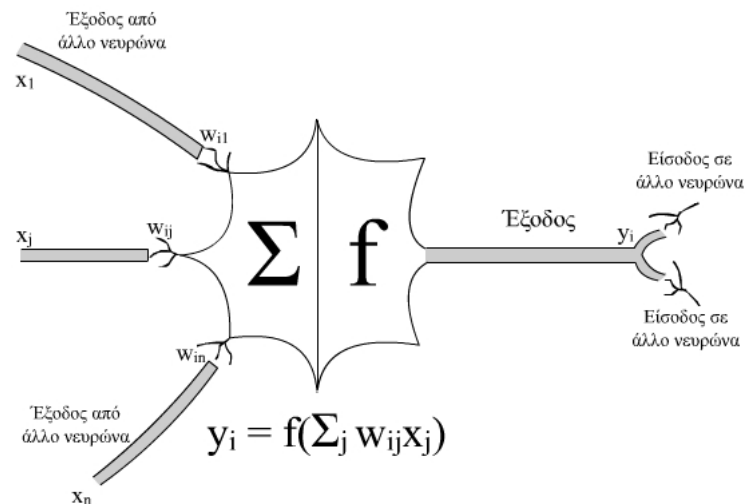
Σχήμα 1.2: Η υπερβολική εφαπτομένη για διάφορες τιμές της παραμέτρου λ



Σχήμα 1.3: Μοντέλο ενός βιολογικού νευρώνα

Ο τρόπος με τον οποίο είναι διασυνδεδεμένοι οι νευρώνες ενός δικτύου ονομάζεται *τοπολογία* ή *αρχιτεκτονική* του ΤΝΔ. Συνήθως, είναι βολικό να βλέπουμε τους νευρώνες σαν να είναι τοποθετημένοι σε *στρώματα* ή *επίπεδα* (layers). Όλοι οι νευρώνες ενός στρώματος, σχεδόν πάντα, έχουν την ίδια συνάρτηση ενεργοποίησης και συμπεριφέρονται με τον ίδιο τρόπο. Κάθε ΤΝΔ έχει τουλάχιστον δύο στρώματα: το στρώμα εισόδου από όπου εισέρχονται τα πρότυπα εκπαίδευσης και το στρώμα εξόδου από το οποίο παίρνουμε την έξοδο του δικτύου.

Αν ένα ΤΝΔ δεν αποτελείται μόνο από το στρώμα εισόδου και το στρώμα εξόδου, τότε τα υπόλοιπα στρώματά του ονομάζονται *κρυφά στρώματα* (hidden layers) και το ΤΝΔ λέγεται *πολυστρωματικό* (multilayer). Τα πολυστρωματικά ΤΝΔ είναι δυσκολότερο να εκπαιδευτούν



Σχήμα 1.4: Μοντέλο ενός τεχνητού νευρώνα

και απαιτούν εμπειρία και προσοχή στο σχεδιασμό τους, αλλά μπορούν να επιλύσουν τα περισσότερα από τα πραγματικά προβλήματα.

Ένας άλλος τρόπος κατηγοριοποίησης των ΤΝΔ είναι ανάλογα με τη φορά και τον τρόπο διάδοσης των πληροφοριών μεταξύ των νευρώνων. Αν το σήμα διαδίδεται έτσι ώστε να μην υπάρχει νευρώνας που η έξοδός του είναι είσοδος κάποιου νευρώνα του ιδίου ή προηγούμενου στρώματος, τότε θα λέμε ότι το ΤΝΔ είναι πρόσθιας τροφοδότησης (feedforward). Επίσης, αν κάθε νευρώνας ενός στρώματος, διασυνδέεται με όλους του νευρώνες του επόμενου στρώματος, τότε λέμε ότι το ΤΝΔ είναι *πλήρως διασυνδεδεμένο* (fully interconnected).

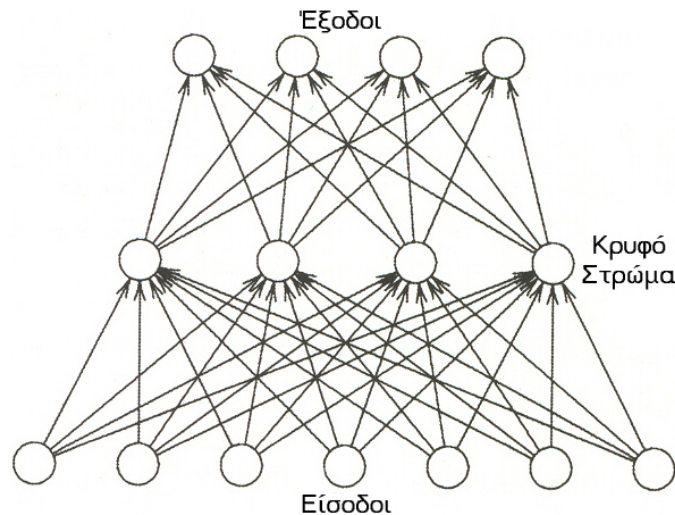
Αντίθετα, αν η έξοδος ενός νευρώνα κινείται προς νευρώνες του ιδίου ή προηγούμενων στρώματος, τότε το ΤΝΔ λέγεται ότι έχει *ανάδραση* (feedback). Τα ΤΝΔ που οι συνδέσεις μεταξύ των νευρώνων σχηματίζουν κύκλους ονομάζονται *δυναμικά ΤΝΔ* και η εκπαίδευσή τους αντιστοιχεί ουσιαστικά στην εύρεση ενός σημείου ισορροπίας του συστήματος και γίνεται με μεθόδους μελέτης Δυναμικών Συστημάτων και της Θεωρίας του Χάους [2]. Η μελέτη των δυναμικών συστημάτων μπορεί να χρησιμοποιηθεί και στην πρόβλεψη χρονοσειρών από ΤΝΔ. Στην περίπτωση αυτή η αρχιτεκτονική του ΤΝΔ μπορεί να καθοριστεί ανάλογα με την χρονοσειρά της οποίας θέλουμε να προβλέψουμε μελλοντικές τιμές [118].

Στο υπόλοιπο της παρούσας Διατριβής θα ασχοληθούμε μόνο με πλήρως διασυνδεδεμένα πρόσθιας τροφοδότησης ΤΝΔ, όπως αυτό που φαίνεται στο Σχήμα 1.5. Συμπερασματικά, ένα ΤΝΔ χαρακτηρίζεται από: (α) την αρχιτεκτονική του, (β) τις συναρτήσεις ενεργοποίησης που χρησιμοποιεί, και (γ) τον αλγόριθμο εκπαίδευσης.

1.2 Ιστορική Αναδρομή

Σε αυτή την ενότητα θα προσπαθήσουμε να κάνουμε μια σύντομη ιστορική αναδρομή και να αναφέρουμε τις σημαντικότερες στιγμές της μελέτης των ΤΝΔ και των αλγορίθμων εκπαίδευσής τους [5, 45].

Η ιστορία των Τεχνητών Νευρωνικών Δικτύων, κατά πολλούς, ξεκινά το 1873 όταν ο ψυχολόγος Alexander Bain πρότεινε την μελέτη του ανθρώπινου εγκεφάλου σαν ένα δίκτυο που μεταδίδει σήματα (signal-transmitting network). Η συνέχεια γίνεται στο τέλος του 19ου και στις αρχές του 20ου αιώνα. Η αρχή έγινε από επιστήμονες όπως οι Hermann von Helmholtz, Ernst Mach και Ivan Pavlov, που προέρχονταν από διαφορετικούς επιστημονικούς κλάδους, όπως η Φυσική, η Ψυχολογία, η Ιατρική, κτλ. Οι εργασίες τους αφορούν γενικά θεωρίες



Σχήμα 1.5: Ένα πολυστρωματικό πλήρως διασυνδεδεμένο πρόσθιας τροφοδότησης ΤΝΔ

μάθησης, την εξαρτημένη θεωρία, τη θεωρία των συνειρμών, γνωστικές θεωρίες, θεωρίες όρασης και φυσιολογίας κτλ. και δεν περιλαμβάνουν κάποιο συγκεκριμένο μαθηματικό μοντέλο περιγραφής της λειτουργίας των δικτύων.

Η σύγχρονη αντίληψη για τα ΤΝΔ ξεκινά γύρω στο 1940 με την εργασία των Warren McCulloch και Walter Pitts [86], που έδειξαν ότι τα ΤΝΔ μπορούν γενικά να υπολογίσουν κάθε αριθμητική ή λογική συνάρτηση. Η εργασία τους πολύ συχνά θεωρείται η απαρχή της μελέτης των ΤΝΔ και των εφαρμογών τους.

Οι Warren McCulloch και Walter Pitts ακολουθήθηκαν από τον Donald Hebb [48] που παρατήρησε ότι η κλασική εξαρτημένη θεωρία (όπως προτάθηκε από τον Pavlov) ισχύει λόγω των ιδιοτήτων των νευρώνων και πρότεινε ένα μηχανισμό μάθησης των βιολογικών νευρώνων (Hebb rule). Οι πρώτες πρακτικές εφαρμογές ήρθαν την δεκαετία του 1950 με την ανακάλυψη του δικτύου perceptron και του αντίστοιχου αλγόριθμου εκπαίδευσης από τον Frank Rosenblatt [131]. Αν και μόνο μια μικρή κλάση προβλημάτων μπορούσε να επιλυθεί από το perceptron το ενδιαφέρον για τα ΤΝΔ αυξήθηκε. Την ίδια εποχή οι Bernard Widrow και Ted Hoff [171] παρουσίασαν ένα αλγόριθμο εκπαίδευσης γραμμικών ΤΝΔ με τις ίδιες περίπου δυνατότητες και δομή με το perceptron. Ο αλγόριθμός τους χρησιμοποιείται ακόμα και σήμερα.

Δυστυχώς, αυτά τα ΤΝΔ είχαν κάποιους σημαντικούς περιορισμούς, όπως έδειξαν οι Martin Minsky και Seymour Papert [91] το 1969. Τα προβλήματα αυτά μπορούσαν να επιλυθούν από κάποια νέα πιο πολύπλοκα (πολυστρωματικά) ΤΝΔ, αλλά εκείνη την εποχή δεν υπήρχαν αλγόριθμοι για την εκπαίδευσή τους. Έτσι πολλοί επηρεασμένοι από τους Minsky και Papert πίστεψαν ότι αυτό είναι το τέλος της έρευνας για τα ΤΝΔ, και έτσι για πολλά χρόνια η έρευνα ατόνησε.

Η επόμενη σημαντική εξέλιξη στον τομέα των ΤΝΔ που μελετάμε ήρθε την δεκαετία του 1980, με την ανακάλυψη της μεθόδου της οπισθοδρομικής διάδοσης του σφάλματος (Back Propagation – BP) από πολλούς ερευνητές ανεξάρτητα. Η εργασία όμως που είχε την μεγαλύτερη επιρροή ήταν αυτή των David Rumelhart και James McClelland [133, 134]. Η μέθοδος της οπισθοδρομικής διάδοσης του σφάλματος είναι ικανή να εκπαιδεύει ΤΝΔ που δεν έχουν τις αδυναμίες του παρελθόντος και άνοιξε νέους δρόμους στην μελέτη και την έρευνα των ΤΝΔ.

Τα τελευταία χρόνια έχουν ανακαλυφθεί αρκετοί αλγόριθμοι εκπαίδευσης και αρχιτεκτονικές ΤΝΔ, και έχει γραφτεί πληθώρα από σχετικά ερευνητικά άρθρα. Η τρέχουσα έρευνα

κινείται προς την κατεύθυνση επίλυσης σύγχρονων και δύσκολων πρακτικών προβλημάτων, αλλά και προς την μαθηματική θεμελίωση των νέων αποτελεσμάτων. Επίσης, την τελευταία δεκαετία έχουν προταθεί διάφορες νέες εμπορικές εφαρμογές των ΤΝΔ, μερικές από τις οποίες θα δούμε στη συνέχεια.

1.3 Εφαρμογές των Τεχνητών Νευρωνικών Δικτύων

Στον ακόλουθο κατάλογο θα παρουσιάσουμε μερικές από τις βασικές εφαρμογές των ΤΝΔ σε διάφορους τομείς της Επιστήμης και της Τεχνολογίας. Οι περισσότερες από αυτές τις εφαρμογές έχουν ήδη υλοποιηθεί και πολλές από αυτές αποτελούν εμπορικά προϊόντα.

Αεροπλοΐα. Δημιουργία αυτόματων πιλότων και προγραμμάτων προσομοίωσης πτήσης, συστήματα ελέγχου πτήσης, ανίχνευση ελαττωμάτων σε τμήματα των αεροπλάνων.

Βιολογία. Βοήθεια στην κατανόηση του εγκεφάλου και άλλων συστημάτων, δημιουργία μοντέλων αμφιβληστροειδούς χιτώνα και κοχλίας.

Γεωλογία. Ανάλυση πιθανότητας ύπαρξης πετρελαίου σε γεωλογικούς σχηματισμούς, ανάλυση πετρωμάτων σε ορυχεία, ανάλυση της μόλυνσης του περιβάλλοντος.

Επιχειρήσεις. Αξιολόγηση υποψηφίων για κάποια θέση, βελτιστοποίηση του συστήματος κράτησης θέσεων σε μεταφορικά μέσα, αναγνώριση γραφικού χαρακτήρα.

Ιατρική. Ανάλυση ομιλίας για την κατασκευή ακουστικών βοηθημάτων, διάγνωση βασισμένη στα συμπτώματα, έλεγχος χειρουργείου, εξαγωγή συμπερασμάτων από ακτινογραφίες, ανάλυση καρδιογραφήμάτων και εγκεφαλογραφημάτων, εντοπισμός καρκίνου σε κολονοσκοπήσεις και μαστογραφίες.

Κατασκευές. Αυτόματος έλεγχος, έλεγχος γραμμής παραγωγής, έλεγχος ποιότητας, επιλογή τμημάτων κατά το στάδιο της συναρμολόγησης.

Οικονομία. Υπολογισμός κινδύνου σε δάνεια και υποθήκες, αναγνώριση πλαστογραφιών, μετάφραση χειρόγραφων εντύπων, εκτίμηση τιμών μετοχών και συναλλάγματος.

Περιβάλλον. Πρόγνωση του καιρού, ανάλυση τάσεων και καιρικών συνθηκών.

Άμυνα. Χειρισμός μη επανδρωμένων οχημάτων και αεροπλάνων, αναγνώριση σημάτων από radar, δημιουργία «έξυπνων» όπλων, αναγνώριση και σκόπευση στόχων, βελτιστοποίηση αξιοποίησης αποθεμάτων, κρυπτογραφία.

Υπολογιστές. Αναγνώριση ομιλίας, εντοπισμός φωνηέντων φθόγγων, μετατροπή κειμένου σε ομιλία, δρομολόγηση πληροφοριών σε δίκτυα υπολογιστών.

1.4 Ιδιότητες των Τεχνητών Νευρωνικών Δικτύων

Το επιστημονικό ενδιαφέρον για τα ΤΝΔ προκύπτει κυρίως από τη δυνατότητά τους να επιλύουν δύσκολα και ενδιαφέροντα υπολογιστικά προβλήματα του πραγματικού κόσμου. Η χρήση των ΤΝΔ προσφέρει τις ακόλουθες πολύ χρήσιμες ιδιότητες και δυνατότητες [47].

Μη γραμμικότητα. Οι νευρώνες, γενικά, είναι μη γραμμικοί, αφού βασίζονται σε μη γραμμικές συναρτήσεις ενεργοποίησης. Κατά συνέπεια, το ΤΝΔ αφού αποτελείται από την σύνθεση πολλών νευρώνων, είναι μη γραμμικό.

Συσχέτιση Εισόδου-Εξόδου. Κατά την εκπαίδευση παρουσιάζουμε στο ΤΝΔ πρότυπα εισόδου ή εκπαίδευσης (που ουσιαστικά κωδικοποιούν το δοθέν πρόβλημα) και τις αντίστοιχες επιθυμητές εξόδους. Σκοπός είναι το ΤΝΔ να φτάσει σε μια τέτοια κατάσταση όπου για κάθε πρότυπο εκπαίδευσης, η εξοδος του να ταυτίζεται με την επιθυμητή εξοδο. Έτσι δημιουργείται μια συσχέτιση μεταξύ των δεδομένων εισόδου και εξόδου, χωρίς όμως τη χρήση κάποιου προκαθορισμένου στατιστικού ή άλλου μοντέλου.

Προσαρμογή. Τα ΤΝΔ έχουν την ικανότητα να μεταβάλλουν τα βάρη τους ανάλογα με το περιβάλλον τους, δηλαδή ανάλογα με τα πρότυπα εισόδου. Έτσι ένα ΤΝΔ είναι δυνατό να συνεχίσει να εκπαιδεύεται για να αντιμετωπίσει μια μικρή αλλαγή των προτύπων ή ακόμα και μη στατικά προβλήματα.

Απόκριση βασισμένη σε ενδείξεις. Τα εκπαιδευμένα ΤΝΔ μπορούν όχι μόνο να ταξινομούν και να τοποθετούν τα πρότυπα εισόδου σε κλάσεις, αλλά επιπρόσθετα δίνουν και τον βαθμό εμπιστοσύνης αυτής της απόφασης. Έτσι μπορούν να ταξινομήσουν και νέα, άγνωστα κατά τη διάρκεια της εκπαίδευσης, πρότυπα. Αυτή η ικανότητα των ΤΝΔ ονομάζεται *γενίκευση*.

Συναφείς πληροφορίες. Η γνώση αντιπροσωπεύεται από τη δομή και την κατάσταση του ΤΝΔ. Κάθε νευρώνας πιθανά επηρεάζει και επηρεάζεται από όλους τους υπόλοιπους νευρώνες. Συνεπώς, συναφείς πληροφορίες αντιμετωπίζονται με φυσικό τρόπο από το ΤΝΔ.

Ανεκτικότητα σε σφάλματα. Τα ΤΝΔ που έχουν υλοποιηθεί σε υλικό (hardware) έχουν την ιδιότητα της ανεκτικότητας σε σφάλματα, γιατί η απόδοση του συστήματος μειώνεται ομαλά σε περίπτωση λάθους. Για παράδειγμα, αν καταστραφεί ένας νευρώνας, το ΤΝΔ δεν θα αχρηστευθεί, αλλά θα συνεχίσει να λειτουργεί με κάπως χειρότερη απόδοση.

Δυνατότητα VLSI υλοποίησης. Η μαζικά παράλληλη φύση των ΤΝΔ τα καθιστά ιδανικά για υλοποίηση σε υλικό με χρήση της τεχνολογίας ολοκλήρωσης πολύ μεγάλης κλίμακας (Very Large Scale Integration – VLSI). Αποτέλεσμα αυτής της υλοποίησης είναι η εξαιρετικά γρήγορη απόκριση του συστήματος και η δυνατότητα της χρησιμοποίησής του σαν μέρος ενός μεγαλύτερου και πολύπλοκου συστήματος (embedded system).

Ομοιομορφία ανάλυσης και σχεδιασμού. Όλα τα μοντέλα ΤΝΔ μοιράζονται κάποιες βασικές αρχές, όπως την έννοια του νευρώνα, των συνδέσμων και της εκπαίδευσης. Αποτέλεσμα αυτού είναι η ευκολότερη διασπορά ιδεών μεταξύ των ερευνητών.

Βιολογική αναλογία. Η κατασκευή των ΤΝΔ είναι εμπνευσμένη από τον ανθρώπινο εγκέφαλο. Έτσι οι Νευροβιολόγοι συχνά μελετούν τα ΤΝΔ για να καταλάβουν καλύτερα την λειτουργία του ανθρώπινου εγκεφάλου και τα αποτελέσματα αυτής της έρευνας βοηθούν την περαιτέρω ανάπτυξη των ΤΝΔ. Αυτός ο κύκλος τροφοδοτεί και τις δύο Επιστήμες και δίνει στα ΤΝΔ ιδιαίτερη ερευνητική αξία.

1.5 Εκπαίδευση των Τεχνητών Νευρωνικών Δικτύων

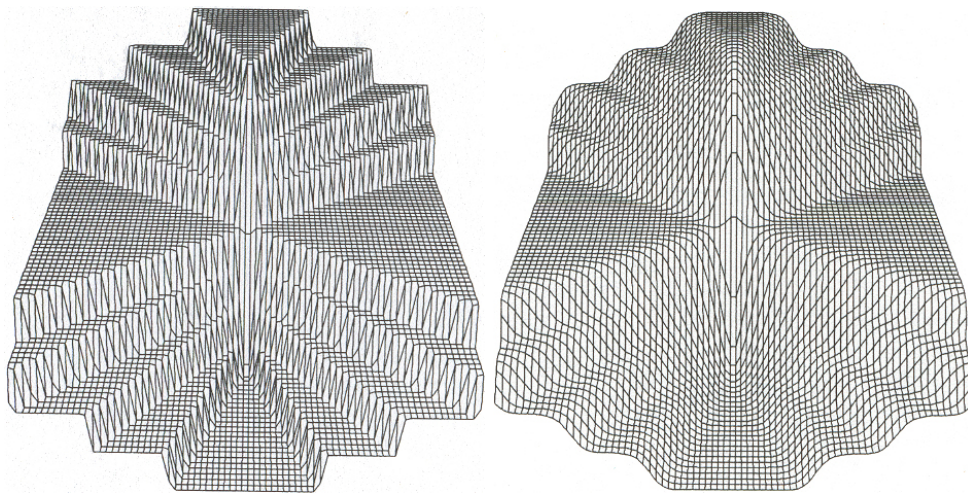
Ο όρος «εκπαίδευση» αναφέρεται στην διαδικασία μεταβολής των βαρών του ΤΝΔ με τέτοιο τρόπο ώστε το δίκτυο να «μαθαίνει» την σχέση μεταξύ των προτύπων εκπαίδευσης και της επιθυμητής εξόδου, με σκοπό την επίλυση κάποιου πρακτικού προβλήματος, όπως η αναγνώριση και ταξινόμηση προτύπων, η προσέγγιση μιας άγνωστης συνάρτησης, η πρόβλεψη των μελλοντικών τιμών μιας χρονοσειράς κτλ. Η προσαρμογή αυτή γίνεται με τη βοήθεια του αλγόριθμου εκπαίδευσης, που συχνά είναι ένας αλγόριθμος βελτιστοποίησης. Το πρόβλημα της αποδοτικής εκπαίδευσης ΤΝΔ είναι δύσκολο και απαιτεί προσεκτική επιλογή της μεθόδου εκπαίδευσης και της αρχιτεκτονικής του ΤΝΔ. Έχει βρεθεί ότι το πρόβλημα της εκμάθησης μιας τυχαίας απεικόνισης από ένα ΤΝΔ είναι στην χειρότερη περίπτωση NP-complete [16]. Εν τούτοις, υπάρχουν αποδοτικοί αλγόριθμοι (πολυωνυμικού χρόνου) που μπορούν να εκπαιδεύσουν ΤΝΔ, η αρχιτεκτονική των οποίων δημιουργείται κατά τη διάρκεια της εκπαίδευσης. Το μειονέκτημα είναι ότι τα δίκτυα αυτά δεν είναι πλήρως διασυνδεδεμένα και τελικά μπορεί να έχουν πολύ μεγάλο αριθμό νευρώνων.

Ο ανεπαρκής αριθμός νευρώνων στο κρυφό στρώμα, η ακατάλληλη αρχικοποίηση των βαρών και η λανθασμένη ρύθμιση των ευρετικών παραμέτρων κάνουν την εκπαίδευση πιο δύσκολη με αποτέλεσμα την σύγκλιση σε τοπικά ελάχιστα με μεγάλη συναρτησιακή τιμή.

Τελικά το ΤΝΔ δεν καταφέρνει να εκπαιδευτεί σε όλα τα πρότυπα εισόδου και η απόδοσή του δεν είναι η αναμενόμενη.

1.5.1 Η μορφολογία του χώρου των βαρών

Η εκπαίδευση ΤΝΔ είναι ένα πολύ δύσκολο πρόβλημα κυρίως λόγω της μορφής του χώρου των βαρών. Διαισθητικά, ο χώρος των βαρών παρουσιάζει πολλαπλά τοπικά ελάχιστα διότι είναι η σύνθεση των μη γραμμικών συναρτήσεων ενεργοποίησης (που έχουν ελάχιστα σε διαφορετικά σημεία), με αποτέλεσμα πολλές φορές η τελική συνάρτηση να μην είναι κυρτή [42]. Ένας άλλος παράγοντας που επηρεάζει την μορφή του χώρου των βαρών όταν χρησιμοποιούνται σιγμοειδείς συναρτήσεις είναι η παράμετρος λ . Για μικρές τιμές αυτής της παραμέτρου συχνά ο χώρος των βαρών φαίνεται να γίνεται πιο ομαλός [56, 128], όπως φαίνεται και στο Σχήμα 1.6.



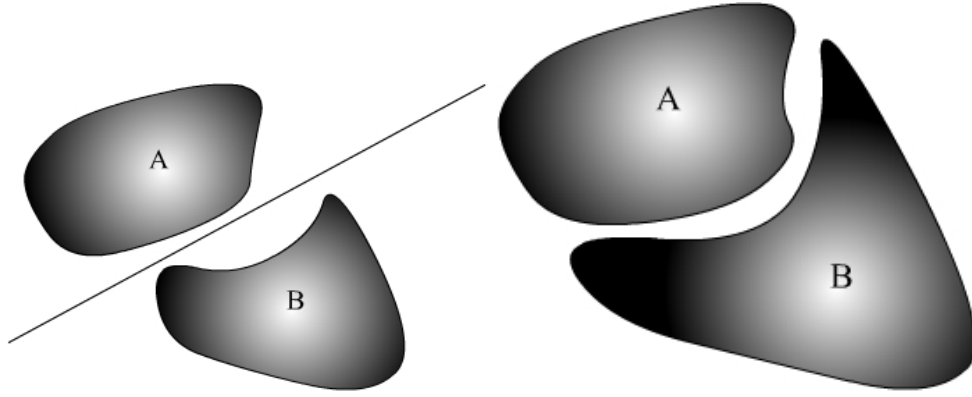
Σχήμα 1.6: Παράδειγμα γραφικής παράστασης του χώρου των βαρών ενός ΤΝΔ με ένα μόνο νευρώνα όταν η τιμή της παραμέτρου της σιγμοειδούς είναι $\lambda = 1$ (αριστερά) και η ίδια γραφική παράσταση για $\lambda = 0.1$

Το πρόβλημα επιδεινώνεται λόγω της πολύ μεγάλης διάστασης του χώρου (τυπικές αρχιτεκτονικές ΤΝΔ αποτελούνται από χιλιάδες βάρη) και από το γεγονός ότι σχεδόν πάντα το σύνολο εκπαίδευσης δεν είναι γραμμικώς διαχωρίσιμο (βλ. Σχήμα 1.7). Στην περίπτωση αυτή δημιουργούνται στενές περιοχές ανεπιθύμητων τοπικών ελαχίστων που μπορούν να παγιδεύσουν τις μεθόδους εκπαίδευσης [56], όπως φαίνεται και στο Σχήμα 1.8.

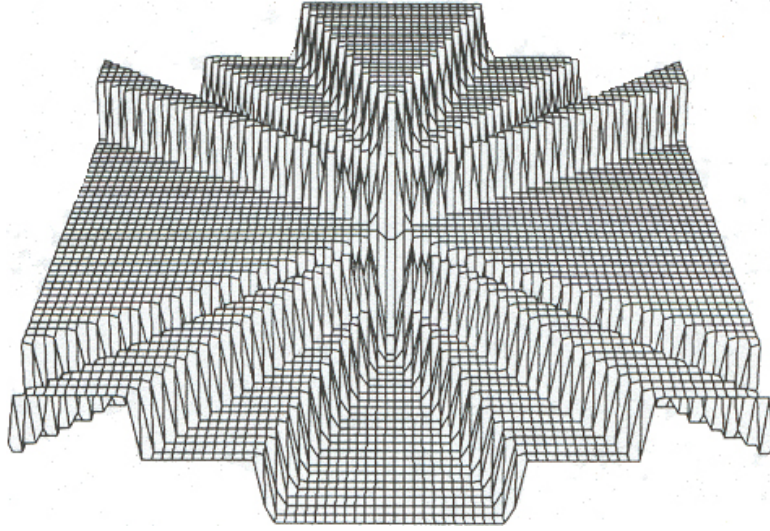
Τέλος, σημαντικό ρόλο παίζει και ο αριθμός των προτύπων, αφού για κάθε νέο πρότυπο η μορφή της συνάρτησης σφάλματος γίνεται πιο απότομη και δημιουργούνται νέες περιοχές με πολλά τοπικά ελάχιστα καθώς επίσης και επίπεδες περιοχές με σχεδόν μηδενική κλίση. Η επίδραση του αριθμού των προτύπων στη μορφή του χώρου των βαρών σε ένα ΤΝΔ με ένα μόνο νευρώνα φαίνεται στο Σχήμα 1.9.

1.5.2 Η αρχικοποίηση των βαρών

Σε αυτή την υποενότητα θα αναφέρουμε κάποιους τρόπους αρχικοποίησης των βαρών. Αν και η κατάλληλη αρχικοποίηση των βαρών αποτελεί ανοικτό πρόβλημα, συνήθως τα βάρη του ΤΝΔ αρχικοποιούνται με μικρούς πραγματικούς αριθμούς. Οι τιμές αυτές μπορεί να είναι από την ομοιόμορφη κατανομή στο διάστημα $(-1, 1)$ ή κάποιο ακόμα πιο περιορισμένο. Εναλλακτικά, μπορούν να χρησιμοποιηθούν κάποιες εμπειρικές τεχνικές για την επιλογή των αρχικών βαρών, έτσι ώστε να βοηθηθεί η διαδικασία της εκπαίδευσης [35, 135, 181].



Σχήμα 1.7: Δύο σύνολα προτύπων εκπαίδευσης A και B. Τα σύνολα είναι γραμμικώς διαχωρίσιμα (αριστερά). Τα σύνολα δεν είναι γραμμικώς διαχωρίσιμα (δεξιά)



Σχήμα 1.8: Παράδειγμα γραφικής παράστασης του χώρου των βαρών ενός ΤΝΔ με ένα μόνο νευρώνα, όταν το σύνολο εκπαίδευσης δεν είναι γραμμικώς διαχωρίσιμο

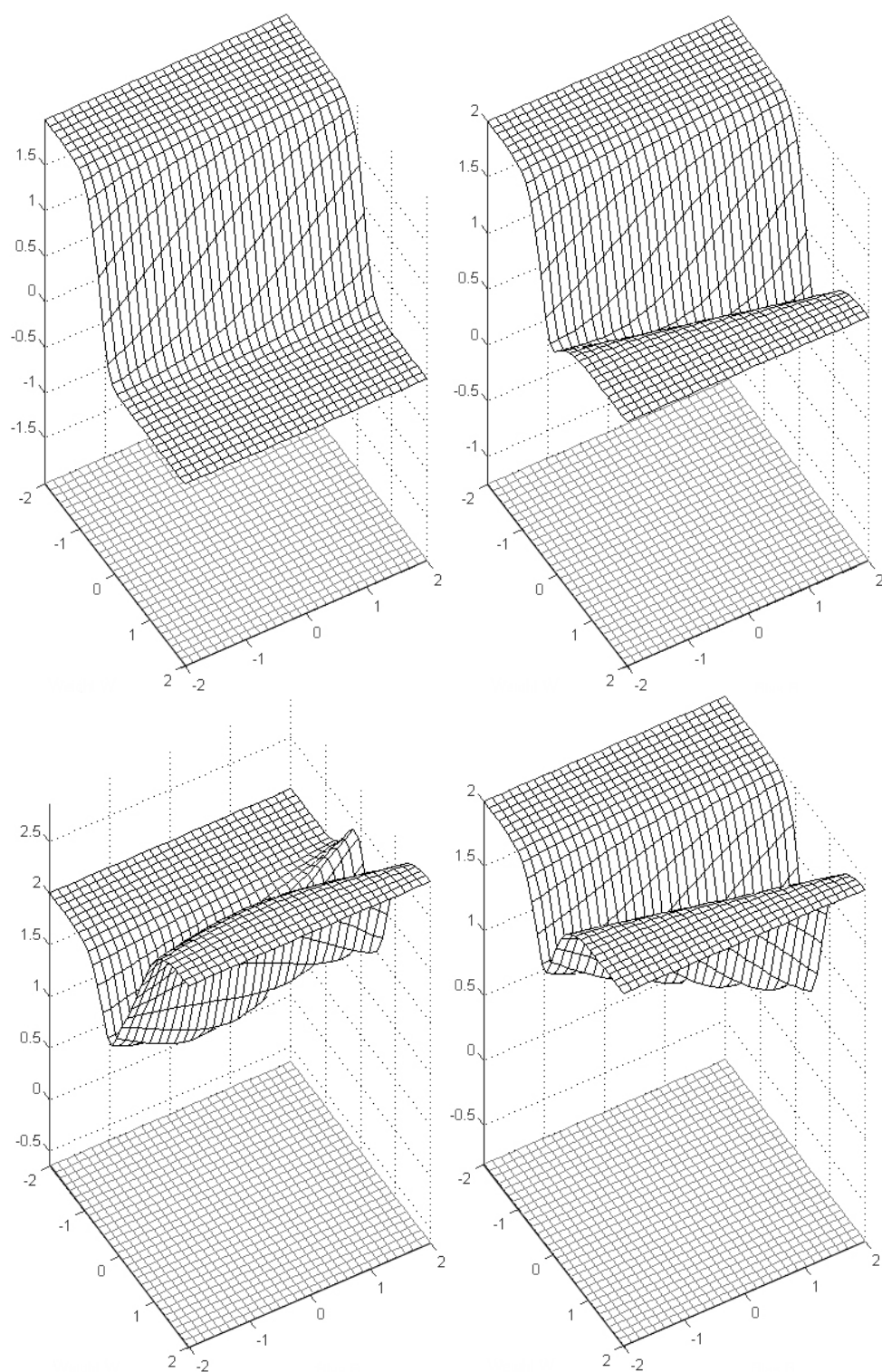
Οι Nguyen και Widrow [72, 94] πρότειναν μια από τις πιο γνωστές τεχνικές αρχικοποίησης των βαρών. Αυτή η τεχνική αποτρέπει τον πρόωρο κορεσμό στους κρυφούς νευρώνες υπολογίζοντας το διάστημα από το οποίο επιλέγονται τα βάρη που συνδέουν την είσοδο με το κρυφό στρώμα. Πρώτα η τεχνική αυτή υπολογίζει την παράμετρο ρ :

$$\rho = 0.7(\mathcal{M}^{1/\mathcal{N}}), \quad (1.1)$$

όπου $\mathcal{N}$ είναι ο αριθμός των νευρώνων εισόδου και $\mathcal{M}$ είναι ο αριθμός των κρυφών νευρώνων. Στη συνέχεια επιλέγονται τα βάρη $w = (w_{11}, \dots, w_{nm}, \dots, w_{\mathcal{N}\mathcal{M}})$ τυχαία από την ομοιόμορφη κατανομή στο διάστημα $(-1, +1)$. Τέλος, τα βάρη των συνδέσεων μεταξύ των νευρώνων εισόδου και των κρυφών νευρώνων, υπολογίζονται από τον ακόλουθο τύπο:

$$w_{nm} = \frac{\rho w_{nm}}{\|w\|}. \quad (1.2)$$

Αυτή η διαδικασία έχει σαν σκοπό να κατανείμει τα αρχικά βάρη έτσι ώστε να είναι πιο πιθανό κάθε πρότυπο εισόδου να προκαλεί αποδοτική εκπαίδευση των κρυφών νευρώνων, με αποτέλεσμα να επιταχύνεται η διαδικασία της εκπαίδευσης [94].



Σχήμα 1.9: Παράδειγμα γραφικής παράστασης του χώρου των βαρών ενός ΤΝΔ με ένα μόνο νευρώνα, όταν το σύνολο εκπαίδευσης αποτελείται από 1, 2, 3 και 4 πρότυπα (δεξιόστροφα ξεκινώντας από επάνω αριστερά)

1.5.3 Κατηγορίες μεθόδων εκπαίδευσης

Οι μέθοδοι εκπαίδευσης ΤΝΔ μπορούν να χωριστούν σε δύο βασικές κατηγορίες: (α) *μεθόδους εκπαίδευσης με επίβλεψη* (supervised learning), και (β) *μεθόδους εκπαίδευσης χωρίς επίβλεψη* (unsupervised learning). Στην πρώτη περίπτωση είναι αναγκαία η παρουσία ενός «δασκάλου», ενώ στη δεύτερη το ΤΝΔ πρέπει να οργανωθεί και να εκπαιδευτεί από μόνο του.

Εκπαίδευση με επίβλεψη ονομάζεται η διαδικασία της προσαρμογής ενός συστήματος έτσι ώστε να έχει συγκεκριμένη απόκριση σε συγκεκριμένες εισόδους. Στην περίπτωση των ΤΝΔ η πραγματική έξοδος του ΤΝΔ συγκρίνεται με την επιθυμητή έξοδο και υπολογίζεται η διαφορά τους. Η διαφορά αυτή αποτελεί το σφάλμα εκπαίδευσης του δικτύου. Στη συνέχεια τα βάρη του ΤΝΔ μεταβάλλονται με τέτοιο τρόπο ώστε στην επόμενη επανάληψη η τιμή του σφάλματος να μειωθεί. Για να είναι δυνατή η εκπαίδευση με επίβλεψη πρέπει πριν αρχίσει η εκπαίδευση να υπάρχει διαθέσιμο ένα σύνολο με πρότυπα εκπαίδευσης και για καθένα από αυτά η επιθυμητή απόκριση του δικτύου. Εκεί βρίσκεται και η συμβολή του «δασκάλου» που πρέπει να χαρακτηρίσει όλα τα πρότυπα εισόδου.

Πολλές φορές ο χαρακτηρισμός των προτύπων εισόδου είναι δύσκολος (όταν αυτά είναι πάρα πολλά) ή αδύνατος (όταν προέρχονται από μια άγνωστη διεργασία). Στις περιπτώσεις αυτές είναι δυνατό να χρησιμοποιηθεί η εκπαίδευση χωρίς επίβλεψη. Πιο συγκεκριμένα, τα βάρη του ΤΝΔ μεταβάλλονται μόνο σε σχέση με τις εισόδους. Με τον τρόπο αυτό, συνήθως, δημιουργούνται ομαδοποιήσεις και το ΤΝΔ μαθαίνει τις συσχετίσεις των προτύπων εισόδου. Έτσι, νέα πρότυπα κατηγοριοποιούνται σύμφωνα με τις υπάρχουσες ομάδες και το δίκτυο μπορεί για παράδειγμα να επιτύχει αναγνώριση προτύπων. Οι πιο γνωστοί κανόνες εκπαίδευσης χωρίς επίβλεψη προτάθηκαν από τους Tuevo Kohonen [63, 64], James Anderson [4] και Stephen Grossberg [44].

Τέλος, ένας άλλος τρόπος εκπαίδευσης είναι η *εκπαίδευση με ενίσχυση* (reinforcement learning) [11]. Σύμφωνα με αυτή την τεχνική ορίζεται ένας στόχος με κάπως αόριστο τρόπο. Έτσι αντί να δίνουμε την επιθυμητή τιμή της εξόδου, δίνουμε μια βαθμολογία της απόδοσης του ΤΝΔ. Για παράδειγμα, ο στόχος μπορεί να είναι το ΤΝΔ να μάθει να παίζει σκάκι. Στην περίπτωση αυτή δεν βαθμολογούμε κάθε μία κίνηση ξεχωριστά, παρά μόνο το τελικό αποτέλεσμα (νίκη, ήττα ή ισοπαλία). Εάν το ΤΝΔ κερδίσει, τότε η τάση του συστήματος να κάνει τις ίδιες κινήσεις (ή μια ακολουθία από κινήσεις) ενισχύεται· αλλιώς η τάση αυτή εξασθενίζεται. Ο τρόπος αυτός εκπαίδευσης είναι υπολογιστικά επίπονος, αλλά ικανός να εκπαιδεύσει ΤΝΔ με βάση ανθρώπινες εμπειρίες που συνήθως συσχετίζονται με το αποτέλεσμα και όχι με κάθε βήμα μιας διαδικασίας.

Από τα παραπάνω καταλαβαίνουμε ότι δεν μπορεί να υπάρξει μία μέθοδος που να αντιμετωπίζει με τον καλύτερο δυνατό τρόπο όλα τα προβλήματα. Στο υπόλοιπο αυτής της Διατριβής θα εστιάσουμε την προσοχή μας σε διάφορες μεθόδους εκπαίδευσης με επίβλεψη, που βασίζονται σε γνωστές αλλά και πρωτότυπες μεθόδους βελτιστοποίησης γενικών μη γραμμικών συναρτήσεων, για την εκπαίδευση πλήρως διασυνδεδεμένων ΤΝΔ πρόσθιας τροφοδότησης. Στο επόμενο κεφάλαιο, θα μελετήσουμε το θεωρητικό υπόβαθρο πάνω στο οποίο στηρίζονται αυτές οι μέθοδοι.

1.5.4 Παράλληλη εκπαίδευση Τεχνητών Νευρωνικών Δικτύων

Η παράλληλη επεξεργασία, δηλαδή η μέθοδος όπου η παράλληλη επίλυση μικρών υπο-προβλημάτων έχει σαν αποτέλεσμα την γρήγορη επίλυση ενός μεγάλου και δύσκολου προβλήματος, έχει αναπτυχθεί πολύ τα τελευταία χρόνια [69]. Τα ΤΝΔ σαν αρχιτεκτονικές που εκτελούν αριθμητικούς υπολογισμούς χρησιμοποιώντας δομή παράλληλης κατανομής επεξεργασίας μπορούν να υλοποιηθούν και να εκτελεστούν από παράλληλες υπολογιστικές μηχανές.

Γενικά, η εκπαίδευση ενός ΤΝΔ με τη βοήθεια κάποιας μεθόδου της κλάσης της οπιθοδρομικής διάδοσης του σφάλματος [133] που θα μελετήσουμε στο επόμενο κεφάλαιο,

αποτελείται από τα ακόλουθα βήματα σε κάθε επανάληψη:

1. Παρουσίαση ολόκληρου του συνόλου των προτύπων εκπαίδευσης και υπολογισμός της εξόδου (ενεργοποίηση) του ΤΝΔ.
2. Υπολογισμός της τιμής της συνάρτησης σφάλματος για όλα τα πρότυπα.
3. Υπολογισμός του διανύσματος των μερικών παραγώγων της συνάρτησης σφάλματος.
4. Προσαρμογή των βαρών με βάση τον αλγόριθμο εκπαίδευσης.

Τα Βήματα 1, 2 και 3 μπορούν να εκτελεστούν παράλληλα, εάν το σύνολο των προτύπων εκπαίδευσης διαμεριστεί κατάλληλα σε πολλούς επεξεργαστές. Αντίθετα το Βήμα 4 είναι καλύτερα να εκτελείται από ένα επεξεργαστή, αθροίζοντας τις τιμές της συνάρτησης σφάλματος που θα λάβει από όλους τους άλλους επεξεργαστές. Όταν το σύνολο των προτύπων εκπαίδευσης είναι μεγάλο, τότε η μεθοδολογία που θα περιγράψουμε μπορεί να αυξήσει σημαντικά την ταχύτητα εκπαίδευσης.

Για την υλοποίηση της παράλληλης εκπαίδευσης ΤΝΔ, μπορούμε να χρησιμοποιήσουμε κάποιο υπολογιστικό σύστημα που αποτελείται από πολλούς επεξεργαστές ή μπορούμε να ακολουθήσουμε το παράδειγμα της κατανεμημένης επεξεργασίας. Γενικά τα υπολογιστικά συστήματα με πολλούς επεξεργαστές είναι πολύ ακριβά και σχετικά δύσκολο να αναβαθμιστούν στο μέλλον. Από την άλλη μεριά, μπορούμε να επιτύχουμε την αξιοποίηση απλών προσωπικών υπολογιστών, που συνδέονται μεταξύ τους μέσω ενός δικτύου, για να επιλύσουμε προβλήματα παράλληλης κατανεμημένης επεξεργασίας [25]. Η υπολογιστική ισχύς που μπορούμε να επιτύχουμε με τον τρόπο αυτό πολλές φορές μπορεί να συγκριθεί ή και να ξεπεράσει την ισχύ ενός σύγχρονου παράλληλου υπολογιστικού συστήματος.

Η διαδικασία που θα περιγράψουμε βασίζεται στην υλοποίηση μιας Παράλληλης Εικονικής Μηχανής (ΠΕΜ) (Parallel Virtual Machine – PVM) [39], που αποτελείται από 15 προσωπικούς υπολογιστές και χρησιμοποιείται για την παράλληλη εκπαίδευση ΤΝΔ με πολύ μεγάλα σύνολα προτύπων εκπαίδευσης. Στην εργασία [104] δίνουμε τις λεπτομέρειες της υλοποίησης της ΠΕΜ και μια αναλυτική εκτίμηση του συνολικού κόστους της, που δεν ξεπερνά τα 11,000 ευρώ για ένα δίκτυο με 15 υπολογιστές, χρησιμοποιώντας το υπολογιστικό παράδειγμα Beowulf [1, 150]. Αξίζει να σημειωθεί ότι αφού υλοποιηθεί η ΠΕΜ μπορεί να χρησιμοποιηθεί και για την επίλυση άλλων υπολογιστικά δαπανηρών προβλημάτων (βλ. την εργασία [115] όπου χρησιμοποιήσαμε την ΠΕΜ για την μαζική εύρεση ριζών ειδικών συναρτήσεων).

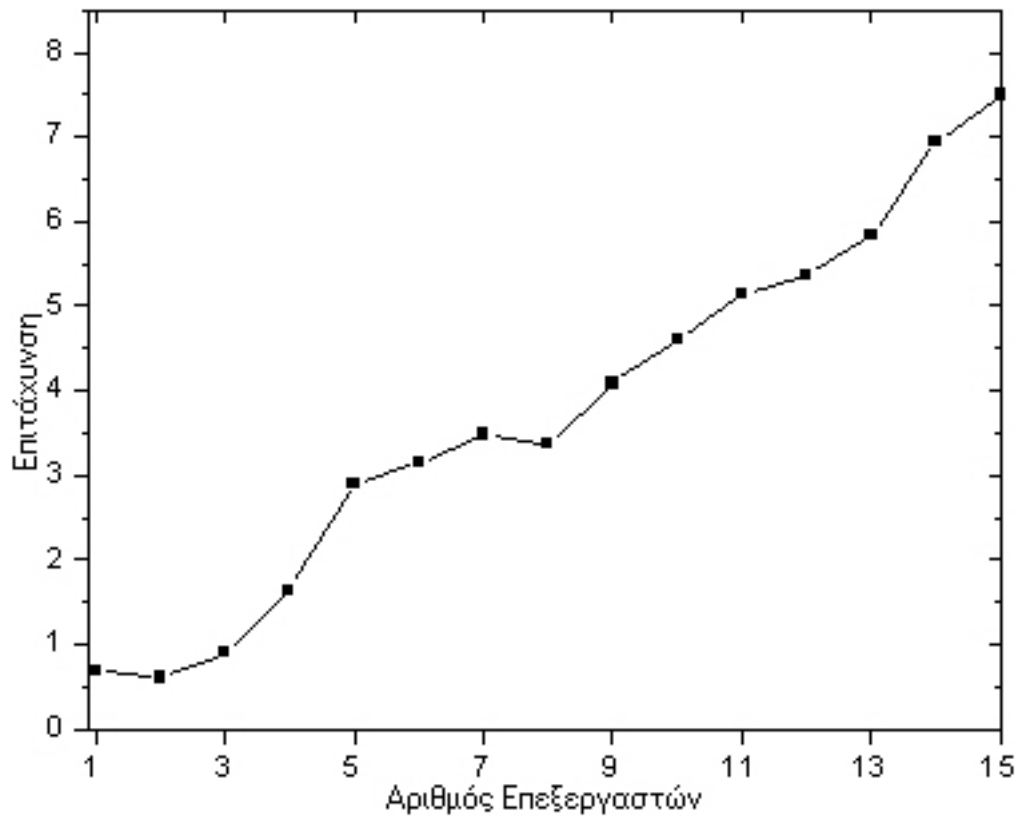
Η ΠΕΜ είναι ένα σύνολο από προγραμματιστικά εργαλεία και βιβλιοθήκες, που μπορούν να χρησιμοποιηθούν για την προσομοίωση ενός παράλληλου επεξεργαστή με τόσους επεξεργαστές όσοι είναι οι υπολογιστές που θα διασυνδέσουμε για να σχηματίσουμε την ΠΕΜ. Αυτό επιτυγχάνεται με τις συναρτήσεις επικοινωνίας και ελέγχου που παρέχει η ΠΕΜ. Οι υπολογιστές αυτοί μπορεί να είναι διαφορετικής αρχιτεκτονικής και δυνατοτήτων και να έχουν διαφορετικό λειτουργικό σύστημα [39]. Το βασικό γνώρισμα της ΠΕΜ είναι ότι τελικά το σύνολο των υπολογιστών που την αποτελεί, συμπεριφέρεται σαν ένας παράλληλος υπολογιστής.

Για να εκπαιδεύσουμε ΤΝΔ παράλληλα χρησιμοποιήσαμε το master-slave μοντέλο παράλληλης επεξεργασίας [39]. Σύμφωνα με αυτό ένα επεξεργαστής διευθύνει και επιβλέπει την εκπαίδευση (master) και πολλοί επεξεργαστές (slaves) επεξεργάζονται το πρόβλημα και επικοινωνούν με τον master για να λάβουν το πρόβλημα ή να στείλουν την λύση.

Η διαδικασία που προτείνουμε είναι η διαμέριση του συνόλου των προτύπων εκπαίδευσης από τον master επεξεργαστή και η αποστολή των μερών στους slave επεξεργαστές. Αυτό έχει σαν αποτέλεσμα κάθε slave επεξεργαστής να υπολογίζει το μέρος της τιμής και του διανύσματος των μερικών παραγώγων της συνάρτησης σφάλματος που αντιστοιχεί στα πρότυπα που έχει. Στη συνέχεια αποστέλλει τις τιμές αυτές στον master επεξεργαστή, όπου αθροίζονται για να βρεθούν τελικά οι τιμή και το διάνυσμα των μερικών παραγώγων για όλο το

σύνολο προτύπων εκπαίδευσης. Τέλος, ακολουθώντας κάποια μέθοδο εκπαίδευσης γίνεται η προσαρμογή των βαρών και τα νέα βάρη αποστέλλονται στους slave επεξεργαστές για την επόμενη επανάληψη (βλ. και την εργασία [104] για μια αναλυτική παρουσίαση της μεθόδου σε ψευδοκώδικα).

Στο Σχήμα 1.10 παρουσιάζουμε την επιτάχυνση (speedup) της παράλληλης διαδικασίας εκπαίδευσης σε σχέση με τη σειριακή εκπαίδευση ανάλογα με το πλήθος των επεξεργαστών που χρησιμοποιούμε, στο πρόβλημα της αναγνώρισης ανωμαλιών σε κολοנוσκοπήσεις (για λεπτομέρειες σχετικά με το πρόβλημα βλ. την εργασία [104] και τις Υποενότητες 3.4.3 και 7.4.2).



Σχήμα 1.10: Επιτάχυνση του χρόνου εκπαίδευσης ανάλογα με το πλήθος των επεξεργαστών

Θεωρητικό Υπόβαθρο των Μεθόδων Εκπαίδευσης ΤΝΔ

Τίποτα δεν είναι τόσο
πρακτικό όσο η Θεωρία.

—J. Robert Oppenheimer (1904–1967)

Ο όρος Τεχνητό Νευρωνικό Δίκτυο (ΤΝΔ), όπως είδαμε και στο Κεφάλαιο 1, αναφέρεται σε μια αρχιτεκτονική που εκτελεί αριθμητικούς υπολογισμούς χρησιμοποιώντας δομή μαζικού παραλληλισμού (massively parallel structure) ή παράλληλης κατανομής εργασίας (parallel distributed processing). Το επιστημονικό ενδιαφέρον για τα διάφορα μοντέλα ΤΝΔ προκύπτει κυρίως από τη δυνατότητά τους να επιλύουν δύσκολα και ενδιαφέροντα υπολογιστικά προβλήματα του πραγματικού κόσμου. Οι κόμβοι, ή τεχνητοί νευρώνες, στα ΤΝΔ θεωρούνται συνήθως ως απλουστευμένα πρότυπα των βιολογικών νευρώνων, δηλαδή τα πραγματικά κύτταρα του ανθρώπινου εγκεφάλου, και τα βάρη σύνδεσης μεταξύ των κόμβων μοιάζουν με τις συνάψεις μεταξύ των ανθρώπινων νευρώνων. Θετικά ή αρνητικά βάρη αντιστοιχούν σε συνάψεις (συνδέσεις) που μεταδίδουν ή αναστέλλουν ερεθίσματα από άλλους νευρώνες. Τα ΤΝΔ παρέχουν ένα εναλλακτικό αλγοριθμικό μοντέλο, το οποίο έχει εμπνευστεί από βιολογικά μοντέλα, σύμφωνα με το οποίο οι υπολογισμοί γίνονται παράλληλα και μαζικά, και η εκπαίδευση αντικαθιστά την ανάπτυξη προγράμματος.

2.1 Εισαγωγή

Η εκπαίδευση ΤΝΔ αποτελεί ένα μέσο δυναμικής αναπαράστασης κωδικοποιημένης πληροφορίας στους νευρώνες ενός ΤΝΔ. Η προσέγγιση ότι η εκπαίδευση ΤΝΔ με επίβλεψη αντιστοιχεί στην ελαχιστοποίηση μιας αντικειμενικής συνάρτησης (συνάρτηση σφάλματος), οδηγεί στην ανάπτυξη αλγορίθμων εκπαίδευσης που βασίζονται στην αριθμητική ελαχιστοποίηση χωρίς περιορισμούς. Έτσι το ζητούμενο είναι η ελαχιστοποίηση της συνάρτησης σφάλματος E , δηλαδή η εύρεση ενός διανύσματος $w^* = (w_1^*, w_2^*, \dots, w_n^*) \in \mathbb{R}^n$, τέτοιου ώστε:

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

όπου η συνάρτηση E ορίζεται συνήθως ως το άθροισμα, για όλα τα πρότυπα εισόδου, των τετραγώνων των διαφορών της πραγματικής εξόδου του ΤΝΔ και της επιθυμητής:

$$E(w) = \frac{1}{2} \sum_{p=1}^P \sum_{j=1}^{N_L} (y_{j,p}^L - t_{j,p})^2, \quad (2.1)$$

όπου P είναι ο συνολικός αριθμός προτύπων, $y_{j,p}^L$ η έξοδος του j νευρώνα που ανήκει στο L στρώμα, N_L ο αριθμός των νευρώνων του στρώματος εξόδου, και $t_{j,p}$ η επιθυμητή έξοδος του j νευρώνα εξόδου στο πρότυπο p .

Η αναγωγή του προβλήματος της εκπαίδευσης ΤΝΔ σε ένα πρόβλημα μη γραμμικών ελαχιστών τετραγώνων έχει πολλά πλεονεκτήματα, όπως η γνώση της τιμής του ολικού ελαχίστου και η υπάρχουσα εμπειρία από τις μεθόδους που έχουν ήδη κατασκευαστεί για την επίλυση αυτού του προβλήματος. Επίσης, είναι προφανές ότι οι μέθοδοι που αναπτύσσονται για την εκπαίδευση ΤΝΔ μπορούν να χρησιμοποιηθούν και για την επίλυση άλλων επιστημονικών προβλημάτων, όπως η μοντελοποίηση άγνωστων συστημάτων (βλ. για παράδειγμα τις εργασίες [59, 60, 132] όπου μελετάμε την επίδραση αέριων ρύπων σε στερεές επιφάνειες).

Γενικά, η ελαχιστοποίηση της συνάρτησης σφάλματος $E(w)$ πραγματοποιείται με την σταδιακή μεταβολή των βαρών από ένα αλγόριθμο εκπαίδευσης. Το διάνυσμα μεταβολής των βαρών δείχνει την κατεύθυνση που πρέπει να ακολουθηθεί στο χώρο των βαρών, έτσι ώστε να ελαττωθεί η τιμή της E . Τα βάρη του ΤΝΔ μεταβάλλονται σύμφωνα με το ακόλουθο επαναληπτικό σχήμα:

$$w^{k+1} = w^k + \Delta w^k, \quad k = 0, 1, \dots,$$

όπου w^{k+1} το νέο διάνυσμα των βαρών, w^k είναι το τρέχον διάνυσμα των βαρών και Δw^k είναι το διάνυσμα μεταβολής των βαρών.

Διαφορετικές τιμές για την διόρθωση Δw^k δημιουργούν διαφορετικούς αλγόριθμους εκπαίδευσης, οι οποίοι είναι συνήθως πρώτης τάξης και βασίζονται σε πληροφορίες σχετικά με το διάνυσμα των μερικών παραγώγων της συνάρτησης σφάλματος (κλίση) για τον υπολογισμό του Δw^k . Υπάρχουν βέβαια και αλγόριθμοι εκπαίδευσης δεύτερης τάξης που χρησιμοποιούν και πληροφορίες σχετικά με τον πίνακα των δευτέρων παραγώγων της συνάρτησης σφάλματος. Στην εργασία [14] ο Battiti παρουσιάζει μια επισκόπηση τεχνικών ελαχιστοποίησης με χρήση παραγώγων πρώτης και δεύτερης τάξης που εφαρμόζονται για εκπαίδευση των ΤΝΔ με επίβλεψη.

Μια μεγάλη ποικιλία από μεθόδους της Αριθμητικής Ανάλυσης έχουν χρησιμοποιηθεί για την επιτάχυνση της εκπαίδευσης, χρησιμοποιώντας πληροφορίες από τις παραγώγους δεύτερης τάξης [14, 85, 92, 99, 146, 169]. Όμως, οι μέθοδοι δεύτερης τάξης είναι πολλές φορές υπολογιστικά επίπονες, όταν το ΤΝΔ έχει μερικές χιλιάδες βάρη [15]. Επιπρόσθετα, δεν είναι βέβαιο ότι το επιπλέον υπολογιστικό κόστος θα επιταχύνει την διαδικασία εκπαίδευσης, ειδικά σε μη κυρτές συναρτήσεις μακριά από το ελάχιστο [28, 95], όπως είναι η συναρτήσεις που δημιουργούνται κατά την εκπαίδευση ΤΝΔ [14]. Για τους παραπάνω λόγους η βελτίωση και η δημιουργία νέων αποδοτικών μεθόδων εκπαίδευσης ΤΝΔ που βασίζονται σε πληροφορίες πρώτης τάξης παρουσιάζει μεγαλύτερο ενδιαφέρον.

Η πιο γνωστή κλάση μεθόδων εκπαίδευσης ΤΝΔ είναι η μέθοδος της πιο απότομης καθόδου (steepest descent), που χρησιμοποιεί σαν διόρθωση, Δw^k , τον όρο $-\eta \nabla E(w^k)$, όπου η είναι μια ευρετική σταθερή παράμετρος (ρυθμός εκπαίδευσης) που συνήθως παίρνει τιμές στο διάστημα $(0, 1)$ (η βέλτιστη τιμή του βήματος η εξαρτάται από την μορφή της πολυδιάστατης συνάρτησης σφάλματος) και $\nabla E(w^k)$ είναι το διάνυσμα των μερικών παραγώγων της συνάρτησης σφάλματος, που υπολογίζεται με την εφαρμογή του κανόνα της αλυσίδας στα στρώματα του ΤΝΔ. Είναι προφανές ότι η συνάρτηση σφάλματος πρέπει να ικανοποιεί τις υποθέσεις για την ύπαρξη παραγώγων πρώτης τάξης. Ασφαλώς αυτό περιορίζει τη μορφή της $E(w)$ και καθιστά απαραίτητο οι νευρώνες του ΤΝΔ να χρησιμοποιούν παραγωγίσιμες συναρτήσεις ενεργοποίησης που επιτρέπουν τον ορισμό της παραγώγου για κάθε νευρώνα. Η μέθοδος για τον υπολογισμό της κλίσης της συνάρτησης σφάλματος ονομάζεται *οπισθοδρομική διάδοση του σφάλματος* (Backpropagation – BP) [133].

Η μέθοδος της οπισθοδρομικής διάδοσης του σφάλματος χαρακτηρίζεται από καλή απόδοση όταν οι αρχικές τιμές του διανύσματος βαρών είναι σχετικά μακριά από το ελάχιστο, κάτι που ισχύει στις περισσότερες περιπτώσεις εκπαίδευσης των ΤΝΔ. Ωστόσο, η σύγκλιση της στην περιοχή του ελαχίστου χαρακτηρίζεται από βραδύτητα. Σημαντικοί περιορισμοί

για τη χρήση της μεθόδου στα ΤΝΔ είναι η αδυναμία της για εγγύηση σύγκλισης σε κάποιο τοπικό ελάχιστο καθώς και η χρήση σταθερού μήκους βήματος που πολλές φορές εμποδίζει τη σύγκλιση και δεν εγγυάται τη μείωση της συνάρτησης σφάλματος σε κάθε επανάληψη του αλγορίθμου εκπαίδευσης. Στο Παράρτημα Β παραθέτουμε την απόδειξη της μεθόδου της οπισθοδρομικής διάδοσης του σφάλματος. Στα επόμενα κεφάλαια θα προτείνουμε και θα μελετήσουμε νέες τροποποιήσεις αυτής της μεθόδου, που σύμφωνα με τα πειραματικά αποτελέσματά μας επιδεικνύουν μεγαλύτερη ταχύτητα σύγκλισης και αυξημένο ποσοστό επιτυχίας.

Τέλος, στην περίπτωση εκπαίδευσης ανά πρότυπο εισόδου p (βλ. και το Κεφάλαιο 7 για μια εκτενή μελέτη τέτοιων μεθόδων) χρησιμοποιείται μια στιγμιαία προσέγγιση της κλίσης της συνάρτησης σφάλματος, που δεν είναι άλλη από τη στήλη του πίνακα των μερικών παραγώγων που αντιστοιχεί στο πρότυπο p . Έχει βρεθεί πειραματικά ότι σε πολλά προβλήματα εκπαίδευσης ΤΝΔ, ο παραπάνω πίνακας των μερικών παραγώγων έχει μεγάλο συντελεστή αστάθειας, γεγονός που οδηγεί σε ελλιπείς πληροφορίες σχετικά με τις κατευθύνσεις ανίχνευσης και έχει ως αποτέλεσμα εξαιρετικά βραδύ χρόνο εκπαίδευσης [137].

2.2 Η Επιλογή του Ρυθμού Εκπαίδευσης

Κάνοντας μια μικρή ιστορική αναδρομή στο πρόβλημα της επιλογής του ρυθμού εκπαίδευσης πρέπει να αναφερθεί ότι πρώτος ο Goldstein στην εργασία [41] πρότεινε μια σχέση που βασίζεται στην Εσσιανή της συνάρτησης σφάλματος και μπορεί να χρησιμοποιηθεί για τον καθορισμό του ρυθμού εκπαίδευσης. Ο ίδιος επίσης απέδειξε τη σύγκλιση της μεθόδου με την προϋπόθεση ότι η συνάρτηση σφάλματος είναι δύο φορές συνεχώς παραγωγίσιμη και έδωσε μια εκτίμηση του ρυθμού σύγκλισής της για την περίπτωση που είναι γνωστό ένα φράγμα της στάθμης της Εσσιανής. Ακολουθώντας μια διαφορετική προσέγγιση, ο Armijo πρότεινε την πρώτη μέθοδο της πιο απότομης καθόδου που επέτρεπε μεταβλητό βήμα σε κάθε επανάληψη, η^k , και απέδειξε τη σύγκλιση της υπό λιγότερο αυστηρές προϋποθέσεις [7]. Μια βελτίωση της μεθόδου αυτής για την εκπαίδευση νευρωνικών δικτύων προτάθηκε στην εργασία [82].

Κατάλληλα επιλεγμένες τιμές των βημάτων εκπαίδευσης βοηθούν να αποφευχθεί η σύγκλιση σε μέγιστα ή σαγματικά σημεία του χώρου των βαρών. Παρόλα αυτά είναι γνωστό ότι αυτή η αντιμετώπιση δεν είναι αποδοτική. Για παράδειγμα, προβλήματα υπάρχουν όταν ο χώρος των βαρών χαρακτηρίζεται από μακρές και βαθιές «χαράδρες», με απότομες «οτροφές» και ο «πυθμένος» είναι ελαφρά κεκλιμένος [57, 133]. Επίσης, αυτή η αντιμετώπιση του προβλήματος εισάγει δυσκολίες στην προσπάθεια δημιουργίας μεθόδων ευρείας σύγκλισης [66, 71]. Βέβαια, υπάρχουν θεωρητικά αποτελέσματα που εγγυώνται την σύγκλιση της μεθόδου όταν ένας σταθερός ρυθμός εκπαίδευσης χρησιμοποιείται. Στην περίπτωση αυτή, ο ρυθμός εκπαίδευσης πρέπει να είναι ανάλογος του αντιστρόφου της σταθεράς Lipschitz της συνάρτησης σφάλματος, που στην πράξη δεν είναι εύκολο να υπολογιστεί [7, 82, 162].

2.3 Αλγόριθμοι Εκπαίδευσης με Ευρεία Σύγκλιση

Όπως αναφέρουν οι Dennis και Schnabel [30, σελ. 5], ο όρος «αλγόριθμος ευρείας σύγκλισης»¹ χρησιμοποιείται για να δηλώσει ότι μια μέθοδος είναι κατασκευασμένη έτσι ώστε να συγκλίνει σε ένα τοπικό ελάχιστο μιας μη γραμμικής συνάρτησης, από σχεδόν οποιοδήποτε αρχικό σημείο. Επιπρόσθετα, ο Nocedal [95, σελ. 200] τονίζει ότι ένας αλγόριθμος λέγεται ότι έχει την ιδιότητα της ευρείας σύγκλισης όταν συγκλίνει σε ένα τοπικό ελάχιστο από «απομακρυσμένα» αρχικά σημεία. Η ιδιότητα της ευρείας σύγκλισης διευκολύνει ιδιαίτερα τη

¹Αντί για τον όρο *ευρεία σύγκλιση* μπορούν να χρησιμοποιηθούν οι ισοδύναμοι όροι *καθολική* ή *ολική σύγκλιση* για την απόδοση του Αγγλικού όρου *global convergence*.

διαδικασία εκπαίδευσης ΤΝΔ καθώς τις περισσότερες φορές η εκπαίδευση ενός προβλήματος ξεκινά για το δίκτυο χρησιμοποιώντας τυχαία αρχικά βάρη, ως επί το πλείστον μακριά από ένα ελάχιστο, ενώ ο χρήστης καλείται να ρυθμίσει ευρετικά διάφορες παραμέτρους κρίσιμες για τη σύγκλιση του αλγόριθμου και την επιτυχία της εκπαίδευσης. Τονίζουμε ότι οι αλγόριθμοι εκπαίδευσης που βασίζονται στη μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος δε συγκλίνουν πάντα σε ένα τοπικό ελάχιστο όταν η αρχική τιμή του διανύσματος βαρών βρίσκεται μακριά από τη γειτονιά του τοπικού ελαχίστου.

Η ευρεία σύγκλιση μπορεί να επιτευχθεί εφαρμόζοντας μεθόδους μη ακριβούς ευθύγραμμης (μονοδιάστατης) ανίχνευσης. Αυτές οι μέθοδοι είναι γνωστές για την ευκολία της υλοποίησής τους σε λογισμικό και για τη μικρή υπολογιστική πολυπλοκότητά τους. Η ενσωμάτωσή τους σε οποιονδήποτε αλγόριθμο που επαναληπτικά προσαρμόζει τα βάρη ακολουθώντας κατευθύνσεις μείωσης της συνάρτησης σφάλματος εξασφαλίζει στον αλγόριθμο την ιδιότητα της ευρείας σύγκλισης [73, 76, 162], όπως θα δούμε και στο Κεφάλαιο 3.

Οι ιδιότητες σύγκλισης των μεθόδων ευθύγραμμης (μονοδιάστατης) ανίχνευσης μπορούν να μελετηθούν εξετάζοντας την κατεύθυνση ανίχνευσης, όπως αυτή ορίζεται από τη γωνία που σχηματίζεται μεταξύ της κατεύθυνσης της πιο απότομης καθόδου και της κατεύθυνσης ανίχνευσης της εκάστοτε μεθόδου, d^k , δηλαδή:

$$\cos \theta_k = \frac{-\nabla E(w^k)^\top d^k}{\|\nabla E(w^k)\| \|d^k\|}, \quad (2.2)$$

και λαμβάνοντας υπόψη το μέγεθος του βήματος. Το μέγεθος του βήματος καθορίζεται από μια επανάληψη της μεθόδου ευθύγραμμης ανίχνευσης. Μια στρατηγική που προτείνεται δέχεται ως βήμα η^k ένα θετικό αριθμό που ικανοποιεί τις παρακάτω συνθήκες:

$$E(w^k + \eta^k d^k) - E(w^k) \leq \sigma_1 \eta^k \nabla E(w^k)^\top d^k, \quad (2.3)$$

$$\nabla E(w^k + \eta^k d^k)^\top d^k \geq \sigma_2 \nabla E(w^k)^\top d^k, \quad (2.4)$$

όπου $0 < \sigma_1 < \sigma_2 < 1$. Οι δύο ανισότητες είναι γνωστές ως συνθήκες του Wolfe [174, 175]. Η πρώτη ανισότητα εξασφαλίζει ότι η συνάρτηση σφάλματος μειώνεται επαρκώς σε κάθε επανάληψη του αλγόριθμου εκπαίδευσης, ενώ η δεύτερη εμποδίζει το ρυθμό εκπαίδευσης να γίνει πολύ μικρός. Οι δύο αυτές ανισότητες αρκούν για να διασφαλιστεί η σύγκλιση σε ένα ελάχιστο αρκεί η γωνία μεταξύ της κατεύθυνσης ανίχνευσης και του διανύσματος των κλίσεων να είναι μικρότερη από 90° [174, 175].

Το θεώρημα των Wolfe και Zoutendijk [174, 175, 180] μπορεί να χρησιμοποιηθεί για να εξασφαλιστεί ευρεία σύγκλιση και είναι ανεξάρτητο από τη μέθοδο που χρησιμοποιείται για το καθορισμό των κατευθύνσεων μείωσης ή των μηκών των βημάτων. Σύμφωνα με αυτό το θεωρητικό αποτέλεσμα ένας αλγόριθμος εκπαίδευσης που βασίζεται στη μέθοδο της πιο απότομης καθόδου έχει την ιδιότητα της ευρείας σύγκλισης αν χρησιμοποιεί ευθύγραμμη (μονοδιάστατη) ανίχνευση που ικανοποιεί τις συνθήκες του Wolfe για τον καθορισμό του ρυθμού εκπαίδευσης. Στο Κεφάλαιο 3 παρουσιάζουμε αυτό το θεώρημα και με τη βοήθειά του αποδεικνύουμε ευρεία σύγκλιση σε μια νέα κλάση μεθόδων εκπαίδευσης ΤΝΔ.

2.4 Βελτιστοποίηση μη Γραμμικών Συναρτήσεων ανά Κατεύθυνση

Είναι χρήσιμο να δούμε και κάποιες μεθόδους που ελαχιστοποιούν την συνάρτηση σφάλματος εκτελώντας μονοδιάστατες ελαχιστοποιήσεις. Είναι γνωστό ότι το σημείο ελαχίστου x^* μιας συνεχώς παραγωγίσιμης συνάρτησης f πρέπει να ικανοποιεί τις απαραίτητες συνθήκες:

$$\nabla f(x^*) = \theta^n = (0, 0, \dots, 0). \quad (2.5)$$

Η Εξίσωση (2.5) είναι ένα σύστημα n μη γραμμικών εξισώσεων που η λύση τους δίνει το x^* . Έτσι, για να ελαχιστοποιήσουμε την συνάρτηση f μπορούμε να βρούμε μια λύση του παραπάνω συστήματος, με την προϋπόθεση η λύση αυτή να αντιστοιχεί σε σημείο ελαχίστου. Αυτό είναι ισοδύναμο με την επίλυση του ακόλουθου συστήματος εξισώσεων:

$$\begin{aligned}\partial_1 f(x_1, x_2, \dots, x_n) &= 0, \\ \partial_2 f(x_1, x_2, \dots, x_n) &= 0, \\ &\vdots \\ \partial_n f(x_1, x_2, \dots, x_n) &= 0,\end{aligned}\tag{2.6}$$

όπου $\partial_i f(x_1, \dots, x_i, \dots, x_n)$ δηλώνουν τις μερικές παραγώγους της f ως προς την i συνιστώσα.

Στη συνέχεια θα μελετήσουμε τις μη γραμμική μέθοδο *Jacobi* και τη μη γραμμική μέθοδο *SOR* για την επίλυση του Συστήματος (2.6). Επίσης θα μελετήσουμε την μέθοδο του *Powell* καθώς και μια τροποποίησή της [168].

2.4.1 Μελέτη σύγκλισης της σύνθετης μη γραμμικής μεθόδου *Jacobi*

Η κλάση των μη γραμμικών μεθόδων *Jacobi* χρησιμοποιείται συχνά για την επίλυση του Συστήματος (2.6). Το κύριο χαρακτηριστικό τους είναι ότι είναι αλγόριθμοι που μπορούν να υλοποιηθούν αποδοτικά σε παράλληλους υπολογιστές [96]. Ξεκινώντας από ένα τυχαίο αρχικό διάνυσμα $x^0 \in \mathcal{D}$, στην k επανάληψη εκτελούμε μονοδιάστατη ελαχιστοποίηση της συνάρτησης:

$$f(x_1^k, \dots, x_{i-1}^k, x_i, x_{i+1}^k, \dots, x_n^k),\tag{2.7}$$

κατά μήκος της i κατεύθυνσης και παίρνουμε το ελάχιστο $\hat{x}_i$. Προφανώς για το $\hat{x}_i$ ισχύει:

$$\partial_i f(x_1^k, \dots, x_{i-1}^k, \hat{x}_i, x_{i+1}^k, \dots, x_n^k) = 0.\tag{2.8}$$

Αυτό αντιστοιχεί σε μονοδιάστατη ελαχιστοποίηση γιατί όλες οι συνιστώσες του διανύσματος x^k , εκτός από την i συνιστώσα, παραμένουν σταθερές. Η i συνιστώσα υπολογίζεται σύμφωνα με την εξίσωση:

$$x_i^{k+1} = x_i^k + \tau_k(\hat{x}_i - x_i^k),\tag{2.9}$$

για κάποια παράμετρο χαλάρωσης τ_k . Συνεπώς, η αντικειμενική συνάρτηση (2.7) ελαχιστοποιείται μονοδιάστατα προς κάθε κατεύθυνση i .

Ανάλογα την χρησιμοποιούμενη μέθοδο μονοδιάστατης ελαχιστοποίησης, μπορούν να κατασκευαστούν διάφορες σύνθετες μη γραμμικές μέθοδοι *Jacobi*. Αξίζει να σημειώσουμε ότι ο αριθμός των επαναλήψεων της μονοδιάστατης ελαχιστοποίησης εξαρτάται από την ζητούμενη ακρίβεια. Έτσι μεγάλος υπολογιστικός κόπος απαιτείται για να βρεθούν με ακρίβεια τα ελάχιστα ανά κάθε κατεύθυνση. Αν και ο υπολογιστικός κόπος αυξάνεται με την διάσταση του προβλήματος, δεν είναι βέβαιο ότι η εύρεση με μεγάλη ακρίβεια του ελαχίστου ανά κατεύθυνση θα επιταχύνει τελικά την ελαχιστοποίηση της f , όταν το αρχικό διάνυσμα είναι μακριά από κάποιο ελάχιστό της. Έτσι στην πράξη πολλές φορές για να βρεθεί το $\hat{x}_i$ εκτελείται μόνο μία επανάληψη της μεθόδου μονοδιάστατης ελαχιστοποίησης [96, 160].

Στη συνέχεια μελετάμε τη σύγκλιση της σύνθετης μη γραμμικής μεθόδου *Jacobi*. Ο σκοπός μας είναι να δείξουμε ότι υπάρχει γειτονιά ενός ελαχίστου της αντικειμενικής συνάρτησης, στην οποία μπορεί να αποδειχθεί η σύγκλιση. Η ανάλυση της σύγκλισης γίνεται κάτω από κατάλληλες προϋποθέσεις και παρέχει χρήσιμες πληροφορίες για αυτή την κλάση μεθόδων.

Θεώρημα 2.1 [162] Έστω $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ δύο φορές συνεχώς παραγωγίσιμη στην ανοικτή περιοχή $\mathcal{D}_0 \subset \mathcal{D}$ του σημείου $x^* \in \mathcal{D}$ για το οποίο $\nabla f(x^*) = \theta^n$ και η Εοσιανή, $H(x^*)$ είναι

θετικά ορισμένη και έχει την ιδιότητα A^π . Τότε υπάρχει ανοικτή σφαίρα $\mathcal{S} = \mathcal{S}(x^*, r)$ στο $\mathcal{S}_0$ (όπου $\mathcal{S}(x^*, r)$ συμβολίζει την ανοικτή σφαίρα με κέντρο x^* και ακτίνα r), τέτοια ώστε η ακολουθία των διανυσμάτων $\{x^k\}_{k=0}^\infty$ που δημιουργείται από την μη γραμμική μέθοδο Jacobi συγκλίνει στο x^* που ελαχιστοποιεί την f .

Απόδειξη. Είναι προφανές ότι οι αναγκαίες και ικανές συνθήκες για να είναι το x^* σημείο ελαχίστου της f ικανοποιούνται αφού $\nabla f(x^*) = \Theta^n$ και η Εοσιανή είναι θετικά ορισμένη στο x^* . Η εύρεση του σημείου αυτού είναι ισοδύναμη με την επαναληπτική επίλυση του Συστήματος (2.6) με την εφαρμογή της μη γραμμικής μεθόδου Jacobi και μιας οποιασδήποτε μεθόδου μονοδιάστατης ελαχιστοποίησης.

Έστω η διαμέριση της $H(x^*)$ σε έναν διαγώνιο πίνακα D , έναν αυστηρά κάτω τριγωνικό L και έναν αυστηρά άνω τριγωνικό πίνακα $L^\top$:

$$H(x^*) = D(x^*) - L(x^*) - L^\top(x^*). \quad (2.10)$$

Αφού, $H(x^*)$ είναι συμμετρική και θετικά ορισμένη, τότε $D(x^*)$ είναι θετικά ορισμένος [157]. Επίσης, αφού η $H(x^*)$ έχει την ιδιότητα A^π , οι ιδιοτιμές του

$$\Phi(x^*) = D(x^*)^{-1} [L(x^*) + L^\top(x^*)],$$

είναι πραγματικές και $\rho(\Phi(x^*)) < 1$ [8] (όπου $\rho(A)$ είναι η φασματική ακτίνα του πίνακα A). Επομένως, υπάρχει ανοικτή σφαίρα $\mathcal{S} = \mathcal{S}(x^*, r)$ στο $\mathcal{S}_0$, τέτοια ώστε για κάθε αρχικό διάνυσμα $x^0 \in \mathcal{S}$, υπάρχει ακολουθία $\{x^k\}_{k=0}^\infty \subset \mathcal{S}$ που ικανοποιεί την μη γραμμική μέθοδο Jacobi και $\lim_{k \rightarrow \infty} x^k = x^*$ [96]. Το Θεώρημα αποδείχθηκε. $\square$

Παρατήρηση 2.1 Ο Young πρότεινε μια κλάση πινάκων, που μπορούν να γραφτούν σε block-τριδιαγώνια μορφή [177]. Τα στοιχεία του πίνακα $A = [a_{ij}]$ χωρίζονται σε δύο σύνολα. Γενικά, κάθε διαμέριση ενός n -διάστατου διανύσματος $x = (x^{(1)}, \dots, x^{(m)})$, σε τμήματα $x^{(p)}$ διάστασης n_p , $p = 1, \dots, m$ (με $\sum_{p=1}^m n_p = n$) ορίζεται μοναδικά από την διαμέριση $\pi = \{\pi_p\}_{p=1}^m$ του συνόλου των πρώτων n ακεραίων, όπου π_p περιέχει τους ακεραίους $s_p + 1, \dots, s_p + n_p$, $s_p = \sum_{j=1}^{p-1} n_j$. Η ίδια διαμέριση π επιφέρει την διαμέριση κάθε $n \times n$ πίνακα A σε blocks A_{ij} διάστασης $n_i \times n_j$. Να σημειωθεί ότι οι πίνακες A_{ii} είναι τετραγωνικοί.

Έτσι, ο πίνακας A έχει την ιδιότητα A^π [8, 177] εάν ο πίνακας $PAP^\top$, μπορεί να γραφτεί στην block-τριδιαγώνια μορφή:

$$PAP^\top = \begin{bmatrix} D_1 & L_1^\top & & & \mathcal{O} \\ L_1 & D_2 & L_2^\top & & \\ & \ddots & \ddots & \ddots & \\ & & L_{r-2} & D_{r-1} & L_{r-1}^\top \\ \mathcal{O} & & & L_{r-1} & D_r \end{bmatrix},$$

όπου οι πίνακες D_i , $i = 1, \dots, r$ είναι μη ιδιάζοντες. Μια αλγοριθμική διαδικασία για το μετασχηματισμό ενός συμμετρικού πίνακα στην παραπάνω τριδιαγώνια μορφή, βρίσκεται στο βιβλίο του Stewart [151, σελ. 335].

2.4.2 Μελέτη σύγκλισης της σύνθετης μη γραμμικής μεθόδου SOR

Ξεκινώντας από ένα τυχαίο αρχικό διάνυσμα $x^0 \in \mathcal{D}$, η μη γραμμική μέθοδος SOR, στην k επανάληψη, ελαχιστοποιεί μονοδιάστατα την συνάρτηση:

$$f(x_1^{k+1}, \dots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \dots, x_n^k), \quad (2.11)$$

κατά μήκος της i κατεύθυνσης και βρίσκει το ελάχιστο $\hat{x}_i$. Και σε αυτή την περίπτωση η i συνιστώσα υπολογίζεται σύμφωνα με την Εξίσωση (2.9). Η βασική διαφορά από τη μη γραμμική μέθοδο Jacobi είναι ότι ο υπολογισμός του x_i στην k επανάληψη, χρησιμοποιεί τις τιμές όλων των προηγούμενων υπολογισμένων μεταβλητών στην k επανάληψη. Παρακάτω δίνουμε ένα θεωρητικό αποτέλεσμα σύγκλισης της μη γραμμικής μεθόδου SOR.

Θεώρημα 2.2 [162] Έστω $f: \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ δύο φορές συνεχώς παραγωγίσιμη σε μια ανοικτή περιοχή $\mathcal{S}_0 \subset \mathcal{D}$ ενός τοπικού ελαχίστου $x^* \in \mathcal{D}$. Τότε υπάρχει ανοικτή σφαίρα $\mathcal{S} = \mathcal{S}(x^*, r)$ στο $\mathcal{S}_0$ τέτοια ώστε η ακολουθία $\{x^k\}$ που δημιουργείται από τη μη γραμμική μέθοδο SOR συγκλίνει στο x^* .

Απόδειξη. Αφού το x^* είναι σημείο τοπικού ελαχίστου της αντικειμενικής συνάρτησης f έχουμε ότι $\nabla f(w^*) = 0$ και η Ερσιανή $H(x^*)$ είναι θετικά ορισμένη. Η εύρεση ενός τέτοιου σημείου ισοδυναμεί με την αντίστοιχη λύση $x^* \in \mathcal{D}$ του Συστήματος (2.6) με την εφαρμογή της μη γραμμικής μεθόδου SOR. Έστω ότι:

$$\Phi_\tau(x^*) = [D(x^*) - \tau L(x^*)]^{-1} \left[(1 - \tau)D(w^*) + \tau L^\top(x^*) \right],$$

για $\tau \in (0, 2)$ όπου D και L ορίζονται όπως στην Εξίσωση (2.10). Τώρα, αφού η $H(x^*)$ είναι συμμετρική και θετικά ορισμένη, τότε $D(x^*)$ είναι μη ιδιάζων [157]. Από το Θεώρημα του Ostrowski [157], έχουμε ότι η φασματική ακτίνα $\rho(\Phi_\tau(x^*)) < 1$ για κάθε $\tau \in (0, 2)$ και συνεπώς, από το Θεώρημα της μεθόδου SOR [96], υπάρχει ανοικτή σφαίρα $\mathcal{S} = \mathcal{S}(x^*, r)$ στο $\mathcal{S}_0$, έτσι ώστε για κάθε $x^0 \in \mathcal{S}$, υπάρχει μοναδική ακολουθία $\{x^k\} \subset \mathcal{S}$ που ικανοποιεί τη μη γραμμική μέθοδο SOR και $\lim_{k \rightarrow \infty} x^k = x^*$. Το Θεώρημα αποδείχθηκε. $\square$

2.4.3 Μελέτη σύγκλισης μιας τροποποίησης της μεθόδου του Powell

Εδώ περιγράφουμε σύντομα τη μέθοδο ελαχιστοποίησης του Powell και προτείνουμε μια τροποποίησή της. Η μέθοδος του Powell [126] βασίζεται στη χρήση συζυγών κατευθύνσεων και η κύρια ιδέα είναι ότι το ελάχιστο μιας θετικά ορισμένης τετραγωνικής μορφής μπορεί να βρεθεί εκτελώντας το πολύ n διαδοχικές ακριβείς μονοδιάστατες ελαχιστοποιήσεις κατά μήκος συζυγών κατευθύνσεων, όπου n είναι ο αριθμός των μεταβλητών της αντικειμενικής συνάρτησης. Η τεχνική αυτή μπορεί να εφαρμοστεί και σε συναρτήσεις που δεν είναι τετραγωνικές, προσθέτοντας μία ακόμα ανίχνευση προς μία νέα σύνθετη κατεύθυνση μετά από τις πρώτες n ακριβείς μονοδιάστατες ελαχιστοποιήσεις.

Μια επανάληψη της μεθόδου του Powell αποτελείται από τα ακόλουθα βήματα, όπου x^0 είναι το αρχικό σημείο, και u_i , $i = 1, 2, \dots, n$ είναι το αρχικό σύνολο των κατευθύνσεων ανίχνευσης, που είναι το σύνολο των διανυσμάτων της βάσης του χώρου e_i , $i = 1, 2, \dots, n$:

1. Για $i = 1, 2, \dots, n$ υπολόγισε το λ_i έτσι ώστε να ελαχιστοποιείται η $f(x^{i-1} + \lambda_i u_i)$, και όρισε $x^i = x^{i-1} + \lambda_i u_i$.
2. Για $i = 1, 2, \dots, n - 1$, αντικατέστησε την κατεύθυνση u_i από την u_{i+1} .
3. Αντικατέστησε την κατεύθυνση u_n από την $(x^n - x^0)$.
4. Υπολόγισε το λ έτσι ώστε να ελαχιστοποιείται η $f(x^n + \lambda u_n)$, και θέσε $x^0 = x^n + \lambda u_n$.

Για γενικές (όχι τετραγωνικές) συναρτήσεις απαιτούνται περισσότερες από n επαναλήψεις και η μέθοδος τερματίζεται όταν ικανοποιηθεί κάποιο κριτήριο τερματισμού. Η μέθοδος του Powell στην πράξη διαλέγει τις νέες κατευθύνσεις με τρόπο που πολλές φορές αυτές είναι γραμμικά εξαρτημένες. Υπάρχουν πολλές διαδικασίες για την αντιμετώπιση αυτού του προβλήματος, αλλά η πιο απλή είναι μετά από n ή $(n+1)$ επαναλήψεις να χρησιμοποιούνται πάλι οι διευθύνσεις της βάσης του χώρου. Στη συνέχεια παρουσιάζουμε μια τροποποίηση

της μεθόδου του Powell, που για την ελαχιστοποίηση χρησιμοποιεί μόνο τα σχετικά μεγέθη των συναρτησιακών τιμών της $f(x + \lambda u)$.

Θέλουμε λοιπόν να ελαχιστοποιήσουμε τη συνάρτηση $f(x^0 + \lambda u)$ κατά μήκος της κατεύθυνσης u . Ο τρόπος που επιλέγουμε είναι να χρησιμοποιήσουμε μια μονοδιάστατη μέθοδο εύρεσης ριζών, για τον υπολογισμό της τιμής του $\lambda \neq 0$ έτσι ώστε:

$$f(x^0 + \lambda u) - f(x^0) = 0. \quad (2.12)$$

Εάν $\hat{\lambda}$ είναι η λύση της παραπάνω εξίσωσης, τότε το σημείο $\hat{x}^0 = x^0 + \hat{\lambda}u$ έχει την ίδια συναρτησιακή τιμή με το σημείο x^0 . Στη συνέχεια επιλέγουμε ένα σημείο που ανήκει στο ευθύγραμμο τμήμα με άκρα x^0 και $\hat{x}^0$, το οποίο να έχει μικρότερη συναρτησιακή τιμή από αυτά τα άκρα. Έτσι μπορούμε να επιλέξουμε αυτό το σημείο, σύμφωνα με την ακόλουθη σχέση:

$$x^1 = x^0 + \gamma (\hat{x}^0 - x^0), \quad \gamma \in (0, 1).$$

Για την επίλυση της μονοδιάστατης εξίσωσης (2.12), χρησιμοποιούμε μια τροποποίηση της μεθόδου της διχοτόμησης (bisection) [161, 164]. Η μέθοδος αυτή δεν απαιτεί τις ακριβείς συναρτησιακές τιμές, αλλά μόνο τα μεγέθη των συναρτησιακών τιμών σε δύο σημεία, δηλαδή απλά συγκρίνει τις συναρτησιακές τιμές. Άλλα πλεονεκτήματα της μεθόδου είναι ότι μπορεί να υλοποιηθεί παράλληλα, συγκλίνει πάντα εντός του δοθέντος διαστήματος και είναι βέλτιστη, δηλαδή έχει ασυμπτωτικά τον καλύτερο ρυθμό σύγκλισης [144].

Τα ακόλουθα θεωρήματα δείχνουν ότι κάθε μέθοδος που ελαχιστοποιεί κατά μήκος ενός συνόλου γραμμικώς ανεξάρτητων συζυγών κατευθύνσεων έχει την ιδιότητα του τετραγωνικού τερματισμού (quadratically termination property).

Θεώρημα 2.3 [126, 179] *Εάν μια τετραγωνική συνάρτηση $f(x)$, n μεταβλητών ελαχιστοποιείται διαδοχικά κατά μήκος ενός συνόλου n γραμμικώς ανεξάρτητων συζυγών κατευθύνσεων, το ολικό ελάχιστο της f θα βρεθεί σε n ή λιγότερες επαναλήψεις, ανεξάρτητα από το αρχικό σημείο και την διαδοχή των κατευθύνσεων.*

Ορισμός 2.1 *Μια μέθοδος που ελαχιστοποιεί την f σύμφωνα με τις απαιτήσεις του Θεωρήματος 2.3, έχει την ιδιότητα του τετραγωνικού τερματισμού (quadratic termination).*

Θεώρημα 2.4 [126, 179] *Οι κατευθύνσεις που χρησιμοποιούνται από τη μέθοδο του Powell είναι συζυγείς.*

Θεώρημα 2.5 *Η τροποποιημένη μέθοδος που προτείνουμε εντοπίζει το ελάχιστο μιας τετραγωνικής συνάρτησης $f(x)$, n μεταβλητών, σε n ή λιγότερες επαναλήψεις, χρησιμοποιώντας μόνο τις σχετικό μέγεθος των συναρτησιακών τιμών της f , ανεξάρτητα από το αρχικό σημείο και την σειρά με την οποία λαμβάνουμε τις κατευθύνσεις της μονοδιάστατης ελαχιστοποίησης.*

Απόδειξη. Αφού η τροποποιημένη μέθοδος χρησιμοποιεί τις κατευθύνσεις u_i της μεθόδου Powell, από το Θεώρημα 2.4 αυτές είναι συζυγείς. Οι αρχικές κατευθύνσεις είναι η βάση του χώρου, άρα γραμμικώς ανεξάρτητες. Επίσης, η γραμμική εξάρτηση των κατευθύνσεων αποφεύγεται με την επαναχρησιμοποίηση των διανυσμάτων της βάσης του χώρου μετά από n ή $(n + 1)$ επαναλήψεις. Συνεπώς οι προϋποθέσεις του Θεωρήματος 2.3 ικανοποιούνται. Το θεώρημα αποδείχθηκε. $\square$

Η ιδιότητα του τετραγωνικού τερματισμού είναι πολύ σημαντική, γιατί δείχνει ότι η σύγκλιση της μεθόδου είναι ταχύτερη ακόμα και σε γενικές (όχι τετραγωνικές) συναρτήσεις. Στην πράξη όλες οι μέθοδοι απαιτούν περισσότερες από n επαναλήψεις, γιατί το ελάχιστο σημείο σε κάθε μονοδιάστατη ελαχιστοποίηση δεν βρίσκεται ακριβώς, με αποτέλεσμα τελικά οι κατευθύνσεις να μην είναι συζυγείς.

2.5 Πρακτική Θεώρηση της Σύγκλισης των Αλγορίθμων Εκπαίδευσης

Τα θεωρητικά αποτελέσματα είναι χρήσιμα για την κατανόηση και τη μελέτη της συμπεριφοράς των αλγορίθμων εκπαίδευσης. Ωστόσο οι αλγόριθμοι, ελαχιστοποιώντας τη μη γραμμική συνάρτηση σφάλματος, έχουν να αντιμετωπίσουν κάποια από τα δυσκολότερα πρακτικά προβλήματα που εμφανίζονται κατά τη βελτιστοποίηση μη γραμμικών συναρτήσεων. Οι κυριότερες δυσκολίες είναι οι ακόλουθες:

- *Το κόστος υπολογισμού των τιμών της συνάρτησης σφάλματος και των παραγώγων της.* Στις εφαρμογές το υπολογιστικό κόστος αποτελεί το βασικό κριτήριο επιλογής του αλγόριθμου εκπαίδευσης, καθώς σε πολλές περιπτώσεις είναι προτιμότερες μερικές ακόμα επαναλήψεις ενός αλγόριθμου που βασίζεται στη μέθοδο της πιο απότομης καθόδου από τη χρήση περίπλοκων αλγορίθμων τοπικής σύγκλισης.
- *Οι μη ακριβείς τιμές της συνάρτησης σφάλματος.* Είναι γνωστό πως οι αριθμητικοί υπολογισμοί υπόκεινται σε σφάλματα ακρίβειας [172]. Οι αριθμητικές πράξεις που απαιτούνται στις προσομοιώσεις των αλγορίθμων εκπαίδευσης επηρεάζουν την ακρίβεια των τιμών της συνάρτησης σφάλματος [176]. Επιπλέον, τα χαρακτηριστικά των μη γραμμικών νευρώνων εμποδίζουν τον ακριβή υπολογισμό των συναρτησιακών τιμών του σφάλματος και οδηγούν σε κορεσμό, τόσο στις προσομοιώσεις όσο και στις υλοποιήσεις των ΤΝΔ [52].
- *Τα πολλαπλά ελάχιστα της συνάρτησης σφάλματος.* Η συνάρτηση σφάλματος δεν είναι κατ' ανάγκη κυρτή και δημιουργείται από την υπέρθεση των μη γραμμικών συναρτήσεων ενεργοποίησης που ελαχιστοποιούνται σε διαφορετικά σημεία. Όταν η τιμή της συνάρτησης σφάλματος σε ένα ελάχιστο είναι μικρότερη από την «επιθυμητή», τίθεται το θέμα της ποιότητας του ελάχιστου. Για παράδειγμα, σε προβλήματα προσέγγισης συναρτήσεων ή αναγνώρισης συστημάτων υπάρχουν πολλά «επιθυμητά» ελάχιστα που προσεγγίζουν, άγνωστο πόσο καλά, το ολικό ελάχιστο. Σε αυτές τις περιπτώσεις το πρόβλημα μπορεί να εξαλειφθεί χρησιμοποιώντας αρκετά μεγάλο αριθμό δεδομένων. Δυστυχώς, υπάρχουν και περιπτώσεις που ο αλγόριθμος εκπαίδευσης συγκλίνει σε «ανεπιθύμητα» ελάχιστα, δηλαδή σε ελάχιστα με συναρτησιακές τιμές μεγαλύτερες από την επιθυμητή. Αυτό συμβαίνει για διάφορους λόγους, π.χ. όταν το πλήθος των κρυφών νευρώνων δεν επαρκεί για τη συγκεκριμένη εφαρμογή ή όταν ο αλγόριθμος αρχικοποιείται με ακατάλληλα αρχικά βάρη και εμποδίζει το ΤΝΔ από το να μάθει πλήρως όλα τα πρότυπα.

2.6 Τα ΤΝΔ σαν Καθολικοί Προσεγγιστές

Κλείνουμε αυτό το κεφάλαιο με μια αναφορά σε θεωρήματα που αποδεικνύουν ότι τα ΤΝΔ είναι *καθολικοί προσεγγιστές* (universal approximators). Μια από τις συχνές χρήσεις των ΤΝΔ είναι να προσεγγίζουν άγνωστες συναρτήσεις. Η απάντηση στο ερώτημα πόσο καλοί προσεγγιστές είναι τα πολυστρωματικά ΤΝΔ, δίνεται από τα ακόλουθα θεωρήματα. Τα Θεώρημα 2.6 και 2.7, που δόθηκαν από τους Kolmogorov και Sprecher αντίστοιχα, αποδεικνύουν ότι τα ΤΝΔ πρόσθιας τροφοδότησης με ένα κρυφό επίπεδο είναι καθολικοί προσεγγιστές, δηλαδή μπορούν να προσεγγίσουν ακριβώς *κάθε* συνεχή συνάρτηση. Τέλος το Θεώρημα 2.8, διατυπώνει τα παραπάνω θεωρητικά αποτελέσματα στην ορολογία των ΤΝΔ.

2.6.1 Θεωρήματα των Kolmogorov και Sprecher

Θεώρημα 2.6 [65] Κάθε συνεχής συνάρτηση $f(x_1, \dots, x_n)$ ορισμένη στο $I^n, n \geq 2$, όπου I είναι το κλειστό μοναδιαίο διάστημα, $I = [0, 1]$, μπορεί να γραφτεί στη μορφή:

$$f(x) = \sum_{j=1}^{2n-1} \xi_j \left(\sum_{i=1}^n \phi_{ij}(x_i) \right),$$

όπου ξ_j και ϕ_{ij} είναι συνεχείς συναρτήσεις μιας μεταβλητής και οι ϕ_{ij} είναι μονότονες συναρτήσεις που δεν εξαρτώνται από την f .

Θεώρημα 2.7 [149] Για κάθε ακέραιο $n \geq 2$ υπάρχει μια πραγματική μονοτόνως αύξουσα συνάρτηση $\phi(x), \phi : [0, 1] \rightarrow [0, 1]$, που εξαρτάται από το n και έχει την ακόλουθη ιδιότητα: για κάθε δεδομένο $\delta > 0$, υπάρχει ένας ρητός $\epsilon, 0 < \epsilon < \delta$, τέτοιος ώστε κάθε πραγματική συνεχής συνάρτηση n μεταβλητών, $f(x)$, που ορίζεται στο I^n , να μπορεί να γραφτεί:

$$f(x) = \sum_{j=1}^{2n-1} \xi \left(\sum_{i=1}^n \lambda^i \phi(x_i + \epsilon(j-1)) + j-1 \right),$$

όπου ξ είναι συνεχής πραγματική συνάρτηση και λ είναι μια σταθερά που δεν εξαρτάται από την f .

Θεώρημα 2.8 [49] Έστω μια συνεχής συνάρτηση $f : I^n \rightarrow \mathbb{R}^m$, όπου $I = [0, 1]$, τότε η f μπορεί να προσεγγιστεί ακριβώς από ένα ΤΝΔ πρόσδιας τροφοδότησης που έχει n εισόδους, m εξόδους και $(2n+1)$ κρυφούς νευρώνες.

Η συνάρτηση ενεργοποίησης του j κρυφού νευρώνα πρέπει να είναι έχει την ακόλουθη μορφή:

$$z_j = \sum_{i=1}^n \lambda^i \phi(x_i + \epsilon j) + j,$$

όπου η πραγματική σταθερά λ και η πραγματική συνεχής μονοτόνως αύξουσα συνάρτηση ϕ δεν εξαρτώνται από την f (αν και εξαρτώνται από το n) και η σταθερά ϵ ικανοποιεί τις προϋποθέσεις του Θεωρήματος 2.7. Ο k -οστος νευρώνας εξόδου έχει συνάρτηση ενεργοποίησης:

$$y_k = \sum_{j=1}^{2n-1} g_k z_j,$$

όπου g_k είναι πραγματικές συνεχείς συναρτήσεις που εξαρτώνται από την f και το ϵ .

Οι Hornik, Stinchcombe και White [53] έδειξαν ότι μπορούμε να χρησιμοποιήσουμε οποιαδήποτε αύξουσα συνάρτηση ενεργοποίησης $h(x)$, όπου $0 \leq h(x) \leq 1$, για κάθε x και επίσης:

$$\lim_{x \rightarrow -\infty} h(x) = 0, \quad \lim_{x \rightarrow \infty} h(x) = 1.$$

Πρέπει να σημειωθεί ότι για να ισχύει αυτό απαιτείται μεγαλύτερος αριθμός κρυφών νευρώνων. Τέλος, κάτω από κάποιες συνθήκες είναι δυνατό να προσεγγιστεί και η παράγωγος της άγνωστης συνάρτησης f [54].

Μέρος II

Μαθηματική Θεμελίωση Μεθόδων Εκπαίδευσης ΤΝΔ

Μαθηματική Θεμελίωση μιας Νέας Κλάσης Αλγορίθμων Ευρείας Σύγκλισης

Το μόνο καλό είναι η γνώση
και το μόνο κακό η άγνοια.

—Σωκράτης (469-399 π.Χ.)

Σε αυτό το κεφάλαιο παρουσιάζεται ένα νέο γενικό θεωρητικό αποτέλεσμα που υποστηρίζει την ανάπτυξη αιτιοκρατικών αλγορίθμων πρώτης τάξης, ευρείας σύγκλισης για την εκπαίδευση ΤΝΔ ανά ομάδα προτύπων εισόδου [76, 80]. Οι συγκεκριμένοι αλγόριθμοι χρησιμοποιούν διαφορετικό ρυθμό εκπαίδευσης για κάθε κατεύθυνση, που ονομάζεται τοπικός ρυθμός εκπαίδευσης (local learning rate) [73, 76, 80]. Το θεωρητικό αυτό αποτέλεσμα μας επιτρέπει να εξοπλίσουμε οποιονδήποτε αλγόριθμο αυτής της κλάσης με μια στρατηγική προσαρμογής της κατεύθυνσης ανίχνευσης, έτσι ώστε αυτή να είναι πάντα κατεύθυνση μείωσης. Με τον τρόπο αυτό η τιμή της συνάρτησης σφάλματος μειώνεται σε κάθε επανάληψη με βεβαιότητα (μονότονη μείωση) και επιτυγχάνεται η ευρεία σύγκλιση της μεθόδου. Η αποτελεσματικότητα του θεωρητικού αυτού αποτελέσματος παρουσιάζεται συγκρίνοντας δύο πολύ γνωστές μεθόδους εκπαίδευσης ΤΝΔ με τις προτεινόμενες ευρείας σύγκλισης τροποποιήσεις τους.

3.1 Εισαγωγή

Το πρόβλημα της προσαρμογής του ρυθμού εκπαίδευσης στους αλγόριθμους οπισθοδρομικής διάδοσης του σφάλματος (αλγόριθμοι πρώτης τάξης) έχει ερευνηθεί εξαντλητικά και έχουν προταθεί διάφορες στρατηγικές προσαρμογής για να βελτιώσουν τη διαδικασία της εκπαίδευσης, όπως οι ακόλουθες:

- Στην αρχή χρησιμοποιείται ένας σχετικά μικρός ρυθμός εκπαίδευσης η^0 , ο οποίος αυξάνεται στην επόμενη επανάληψη, $k + 1$, εάν διαδοχικές επαναλήψεις μειώνουν την τιμή της συνάρτησης σφάλματος, ή ο ρυθμός εκπαίδευσης μειώνεται δραστικά αν υπάρξει σημαντική αύξηση της τιμής της συνάρτησης σφάλματος [13, 159].
- Στην αρχή χρησιμοποιείται ένας σχετικά μικρός ρυθμός εκπαίδευσης η^0 , ο οποίος αυξάνεται στην $k + 1$ επανάληψη, εάν η κατεύθυνση της πιο απότομης καθόδου της συνάρτησης σφάλματος παραμένει σχεδόν σταθερή για διαδοχικές επαναλήψεις, ή ο ρυθμός εκπαίδευσης μειώνεται δραστικά αν υπάρξει σημαντική αλλαγή στην κατεύθυνση της πιο απότομης καθόδου της συνάρτησης σφάλματος [23].
- Χρησιμοποιείται ένας διαφορετικός τοπικός ρυθμός εκπαίδευσης για κάθε βάρος $w_i^k \in \mathbb{R}^n$ ($i = 1, 2, \dots, n$), δηλαδή $\eta_1^k, \eta_2^k, \dots, \eta_n^k$, που αυξάνεται αν διαδοχικές διορθώσεις των βαρών είναι προς την ίδια κατεύθυνση· διαφορετικά μειώνεται [57, 103, 129, 145].

Σ' αυτό το κεφάλαιο θα εστιάσουμε στην κλάση των αλγορίθμων εκπαίδευσης πρώτης τάξης που χρησιμοποιούν διαφορετικούς ρυθμούς εκπαίδευσης για κάθε βάρος. Από την σκοπιά της βελτιστοποίησης συναρτήσεων, στην πραγματικότητα αυτές οι μέθοδοι χρησιμοποιούν ένα διαφορετικό βήμα για κάθε μια κατεύθυνση του χώρου της αντικειμενικής συνάρτησης. Αυτή η προσέγγιση βοηθά την μέθοδο να κινηθεί γρηγορότερα και πιο αποδοτικά σε περιοχές όπου η κλίση προς κάποια κατεύθυνση είναι μικρή, ενώ η κλίση προς κάποια άλλη είναι μεγάλη. Στην κατεύθυνση με την μικρή κλίση πρέπει να επιλεγεί ένα μεγάλο βήμα, έτσι ώστε να ξεφύγουμε από αυτή την επίπεδη περιοχή, ενώ αντίθετα στην κατεύθυνση με τη μεγάλη κλίση το βήμα πρέπει να είναι μικρό έτσι ώστε να γίνει προσεκτική ανίχνευση του χώρου για την εύρεση κάποιου πιθανού ελαχίστου.

Οι μέθοδοι που χρησιμοποιούν διαφορετικό ρυθμό εκπαίδευσης για κάθε βάρος, συνήθως το επιτυγχάνουν με την κατάλληλη αρχική επιλογή τιμών για κάποιες κρίσιμες ευρετικές παραμέτρους και με την ταυτόχρονη ρύθμιση κάποιων παραμέτρων εκπαίδευσης σε κάθε επανάληψη. Ο σκοπός είναι να επιτύχουν να μειώνουν την τιμή της συνάρτησης σφάλματος σε κάθε μια κατεύθυνση του χώρου των βαρών. Όμως, δεν υπάρχει βεβαιότητα ότι η τιμή της συνάρτησης σφάλματος θα μειώνεται μονότονα σε κάθε επανάληψη και ότι η ακολουθία των σημείων θα συγκλίνει τελικά σε κάποιο ελάχιστο της συνάρτησης E .

Παρακάτω θα παρουσιάσουμε μερικές μεθόδους που χρησιμοποιούν διαφορετικό ρυθμό εκπαίδευσης για κάθε βάρος και θα σχολιάσουμε τα πλεονεκτήματα και τα μειονεκτήματά τους. Στη συνέχεια του κεφαλαίου παρουσιάζουμε ένα νέο θεωρητικό αποτέλεσμα που επιτρέπει την τροποποίηση οποιασδήποτε μεθόδου, έτσι ώστε να αποκτήσει την ιδιότητα της ευρείας σύγκλισης δηλαδή σύγκλιση σε ένα τοπικό ελάχιστο της συνάρτησης σφάλματος από σχεδόν οποιοδήποτε αρχικό σημείο. Τέλος για να δείξουμε την αποτελεσματικότητα της προτεινόμενης τεχνικής συγκρίνουμε δύο γνωστούς αλγόριθμους εκπαίδευσης με τις προτεινόμενες τροποποιήσεις τους. Τα αποτελέσματά μας είναι ιδιαίτερα ικανοποιητικά και δείχνουν ότι η προτεινόμενη στρατηγική έχει θετική επίδραση στη συμπεριφορά των αλγορίθμων που δοκιμάσαμε.

3.2 Μέθοδοι με Διαφορετικό Ρυθμό Εκπαίδευσης για κάθε Βάρος

Το κίνητρο για την δημιουργία μεθόδων που χρησιμοποιούν διαφορετικό ρυθμό εκπαίδευσης για κάθε βάρος είναι το γεγονός ότι ο βέλτιστος ρυθμός εκπαίδευσης κατά μήκος μιας κατεύθυνσης της συνάρτησης σφάλματος, δεν είναι απαραίτητα ικανοποιητικός και για κάποια άλλη κατεύθυνση. Επιπροσθέτως, σχεδόν πάντα στην πράξη, ένας ρυθμός εκπαίδευσης δεν είναι ικανοποιητικός για όλες τις περιοχές του χώρου των βαρών. Έτσι είναι βασικό να επιλεχθεί ο ρυθμός εκπαίδευσης έτσι ώστε να είναι αρκετά μεγάλος για να επιτρέπει την αποφυγή επίπεδων περιοχών με μικρή κλίση, ενώ ταυτόχρονα να αποτρέπει την ταλάντωση της μεθόδου στις απότομες περιοχές με μεγάλη κλίση.

Για αυτόν τον λόγο, μια κοινή προσέγγιση είναι η χρησιμοποίηση διαφορετικού ρυθμού εκπαίδευσης για κάθε βάρος (κατεύθυνση) με σκοπό την αποφυγή της αργής σύγκλισης στις επίπεδες κατευθύνσεις και των ταλαντώσεων στις απότομες κατευθύνσεις, καθώς επίσης και για την αξιοποίηση του έμφυτου παραλληλισμού στον υπολογισμό των τιμών της συνάρτησης σφάλματος $E(w)$ και του διανύσματος των μερικών παραγώγων της $\nabla E(w)$ από τον αλγόριθμο της οπισθοδρομικής διάδοσης του σφάλματος. Στη βιβλιογραφία έχουν προταθεί πάρα πολλοί αλγόριθμοι που είναι παραλλαγές της μεθόδου οπισθοδρομικής διάδοσης του σφάλματος [35, 36, 57, 103, 129, 145], και έχουν τον γενικό επαναληπτικό τύπο:

$$w^{k+1} = w^k - \text{diag}\{\eta_1^k, \dots, \eta_i^k, \dots, \eta_n^k\} \nabla E(w^k). \quad (3.1)$$

Οι αλγόριθμοι που ακολουθούν τον παραπάνω επαναληπτικό τύπο επιχειρούν να μειώ-

σουν την τιμή της συνάρτησης σφάλματος ανιχνεύοντας τοπικά ελάχιστα με μικρά βήματα. Αυτά τα βήματα περιορίζονται συνήθως από ευρετικές παραμέτρους, που εξαρτώνται από το εκάστοτε πρόβλημα, προκειμένου αποφύγουν τις ταλαντώσεις και να εξασφαλίσουν ελαχιστοποίηση της συνάρτησης σφάλματος κατά μήκος κάθε κατεύθυνσης. Αυτό το γεγονός συνήθως έχει ως αποτελέσματα να υπάρχει μια προσπάθεια εξισορρόπησης μεταξύ της σταθερότητας και της ταχύτητας σύγκλισης· όσο πιο δεσμευτικές είναι οι ευρετικές παράμετροι τόσο πιο σίγουρη, αλλά και αργή, είναι η σύγκλιση. Για παράδειγμα, η μέθοδος *delta-bar-delta* [57], η μέθοδος των Silva και Almeida [145], και η μέθοδος *Quickprop* [35] (η μέθοδος *Quickprop* εξετάζεται εκτενώς στο Κεφάλαιο 4) εισάγουν πρόσθετες ευρετικές παραμέτρους εκπαίδευσης που εξαρτώνται από το πρόβλημα, για να αυξήσουν την σταθερότητά τους.

Μια άλλη κοινή προσέγγιση, που χρησιμοποιείται για παράδειγμα από τη μέθοδο οπισθοδρομικής διάδοσης του σφάλματος με προσαρμοστικό ρυθμό εκπαίδευσης για κάθε βάρος [84] και από τη μέθοδο *Rprop* [129], είναι να υιοθετηθούν άνω και κάτω όρια για την τιμή του ρυθμού εκπαίδευσης, που επιλέγονται ευρετικά, και βοηθούν στην αποφυγή της χρήσης εξαιρετικά μικρών ή μεγάλων ρυθμών εκπαίδευσης που συχνά αποπροσανατολίζουν την γενική κατεύθυνση ανίχνευσης. Το κάτω όριο της τιμής του ρυθμού εκπαίδευσης βοηθά στην αποτροπή πιθανής αργής σύγκλισης και το άνω όριο, που εξαρτάται από τη μορφή της συνάρτησης σφάλματος, περιορίζει την επιρροή που μπορεί να έχει ένας συγκριτικά μεγάλος ρυθμός εκπαίδευσης μιας κατεύθυνσης στην τελική (συνισταμένη) κατεύθυνση ανίχνευσης.

Μια γνωστή δυσκολία στην ρύθμιση ευρετικών παραμέτρων για την προσαρμογή του ρυθμού εκπαίδευσης είναι ότι πιθανή αστοχία στην επιλογή τους για μια από τις κατευθύνσεις μπορεί να επηρεάσει την συνισταμένη κατεύθυνση. Έτσι είναι πιθανό η ελαχιστοποίηση να περιοριστεί μόνο σε ένα υπόχωρο του χώρου των βαρών. Τότε, η μέθοδος δεν μπορεί να χρησιμοποιήσει πληροφορίες από όλες τις κατευθύνσεις και η λειτουργία του εκπαιδευμένου ΤΝΔ συνήθως είναι λανθασμένη. Τέλος, είναι πολύ δύσκολο να εξασφαλισθεί θεωρητικά ότι η ακολουθία των βαρών που δημιουργούν οι μέθοδοι που βασίζονται σε ευρετικές παραμέτρους εκπαίδευσης συγκλίνει σε ένα τοπικό ελάχιστο της συνάρτησης σφάλματος E [35, 36, 57, 84, 103, 129, 145].

3.3 Ευρεία Σύγκλιση Αλγορίθμων με Τοπικό Ρυθμό Εκπαίδευσης

Η εκπαίδευση πολυστρωματικών ΤΝΔ μπορεί να θεωρηθεί ως ένα μη γραμμικό πρόβλημα ελαχιστοποίησης, που η αντικειμενική συνάρτηση είναι σύνθεση συνεχών σιγμοειδών συναρτήσεων που έχουν ευρείες επίπεδες περιοχές με αυθαίρετα μικρή κλίση [14, 138].

Οι αλγόριθμοι εκπαίδευσης πρώτης τάξης που ακολουθούν το επαναληπτικό Σχήμα (3.1) συνήθως προσαρμόζουν τους τοπικούς ρυθμούς εκπαίδευσης με τη βοήθεια ευρετικών παραμέτρων που αξιοποιούν, ανάλογα με τον αλγόριθμο, πληροφορίες σχετικά με τις προηγούμενες τιμές του διανύσματος των μερικών παραγώγων της συνάρτησης $E(w)$, ως προς το i -στο βάρος ή/και πληροφορίες σχετικά με τις προηγούμενες τιμές κάθε ρυθμού εκπαίδευσης. Για παράδειγμα η μέθοδος *Quickprop* [35] ακολουθεί ανεξάρτητα βήματα σε κάθε κατεύθυνση από την εξίσωση της χορδής [166] (για ανάλυση της σύγκλισης της *Quickprop* βλ. και Κεφάλαιο 4), ενώ η μέθοδος *Rprop* [129] τροποποιεί τα βάρη σε κάθε κατεύθυνση χρησιμοποιώντας το πρόσημο των συνιστωσών του διανύσματος των μερικών παραγώγων της συνάρτησης σφάλματος.

Είναι σαφές ότι το νέο διάνυσμα των βαρών της Σχέσης (3.1) δεν βρίσκεται πάνω στην αντίθετη κατεύθυνση του διανύσματος των μερικών παραγώγων της συνάρτησης σφάλματος (δηλ. την κατεύθυνση της πιο απότομης καθόδου)· αντίθετα χρησιμοποιείται μια νέα κατεύθυνση ανίχνευσης. Η νέα συνισταμένη κατεύθυνση είναι το γινόμενο του τοπικού ρυθμού εκπαίδευσης για κάθε κατεύθυνση επί την μερική παράγωγο της συνάρτησης σφάλματος $E(w)$ ως προς το αντίστοιχο βάρος, δηλαδή $-\eta_i \partial_i E(w)$. Αυτή η τακτική οδηγεί στη μείωση

του σφάλματος κατά μήκος κάθε κατεύθυνσης με την εκτέλεση των μικρών βημάτων στο χώρο των βαρών, ώστε να εξασφαλιστεί ελαχιστοποίηση της συνάρτησης σφάλματος και σκοπός της είναι τελικά να υπάρξει μονότονη μείωση του σφάλματος κατά μήκος της συνισταμένης κατεύθυνσης ανίχνευσης.

Στα πλαίσια της θεωρίας Βελτιστοποίησης, το πρόβλημα της τροποποίησης μιας επαναληπτικής μεθόδου ελαχιστοποίησης έτοι ώστε να αποκτήσει την ιδιότητα της ευρείας σύγκλισης, περιγράφεται παρακάτω. Έστω ότι $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ είναι η αντικειμενική συνάρτηση που πρέπει να ελαχιστοποιηθεί, χρησιμοποιώντας το ακόλουθο επαναληπτικό σχήμα:

$$x^{k+1} = x^k + \alpha^k d^k, \quad (3.2)$$

όπου d^k είναι μια κατεύθυνση μείωσης και α^k είναι το βήμα που υπολογίζεται με τη βοήθεια μιας μονοδιάστατης γραμμικής ανίχνευσης που ικανοποιεί τις συνθήκες του Wolfe [174, 175]:

$$f(x^k + \alpha^k d^k) - f(x^k) \leq \sigma_1 \alpha^k \nabla f(x^k)^\top d^k, \quad (3.3)$$

$$\nabla f(x^k + \alpha^k d^k)^\top d^k \geq \sigma_2 \nabla f(x^k)^\top d^k, \quad (3.4)$$

όπου $\nabla f(x)$ είναι το διάνυσμα των μερικών παραγώγων της f στο σημείο x , και $0 < \sigma_1 < \sigma_2 < 1$. Τότε, το επόμενο Θεώρημα που προτάθηκε ανεξάρτητα από τους Zoutendijk [180] και Wolfe [174, 175], μπορεί να χρησιμοποιηθεί για να αποδείξουμε την ευρεία σύγκλιση του Σχήματος (3.2).

Θεώρημα 3.1 [174, 175, 180] Έστω ότι $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ είναι κάτω φραγμένη στο $\mathbb{R}^n$ και ότι η f είναι συνεχώς παραγωγίσιμη σε μια γειτονιά $\mathcal{N}$ του συνόλου $\mathcal{L} = \{x : f(x) \leq f(x^0)\}$, όπου x^0 είναι το αρχικό σημείο του επαναληπτικού Σχήματος (3.2). Υποθέτουμε ακόμα ότι το διάνυσμα των μερικών παραγώγων της συνάρτησης f είναι Lipschitz συνεχές, δηλαδή υπάρχει σταθερά $L > 0$ τέτοια ώστε:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|,$$

για όλα τα $x, y \in \mathcal{N}$. Τότε η συνθήκη του Zoutendijk:

$$\sum_{k \geq 1} \cos^2 \theta_k \|\nabla f(x^k)\|^2 < \infty, \quad (3.5)$$

όπου

$$\cos \theta_k = \frac{-\nabla f(x^k)^\top d^k}{\|\nabla f(x^k)\| \|d^k\|}, \quad (3.6)$$

ικανοποιείται.

Παρατήρηση 3.1 Ας υποθέσουμε ότι ένα επαναληπτικό σχήμα της μορφής (3.2) ακολουθεί μια κατεύθυνση μείωσης d^k , η οποία δεν τείνει να γίνει ορθογώνια με την κατεύθυνση του διανύσματος των μερικών παραγώγων της συνάρτησης f , $\nabla f(x^k)$, για την οποία

$$\cos \theta_k \geq \zeta > 0,$$

για όλα τα k . Τότε, από την συνθήκη του Zoutendijk (3.5), ισχύει ότι:

$$\lim_{k \rightarrow \infty} \|\nabla f(x^k)\| = 0, \quad (3.7)$$

που σημαίνει ότι η ακολουθία των διανυσμάτων των μερικών παραγώγων συγκλίνει στο μηδέν.

Για ένα επαναληπτικό σχήμα της μορφής (3.2), το όριο (3.7) είναι το καλύτερο θεωρητικό αποτέλεσμα ευρείας σύγκλισης που μπορούμε να αποδείξουμε (βλ. [95] για μια αναλυτική παρουσίαση). Από τα παραπάνω, γίνεται προφανές ότι δεν δίνεται καμιά εγγύηση ότι το επαναληπτικό σχήμα (3.2) θα συγκλίνει σε ένα σημείο ολικού ελαχίστου x^* . Αντίθετα αποδεικνύεται ότι έχει την ιδιότητα της ευρείας σύγκλισης, δηλαδή να συγκλίνει σε ένα τοπικό ελάχιστο σχεδόν από οποιοδήποτε αρχικό σημείο του χώρου των βαρών [30, 95].

Για να εφαρμόσουμε τώρα το παραπάνω θεωρητικό αποτέλεσμα σε μεθόδους εκπαίδευσης ΤΝΔ, παρατηρούμε αρχικά ότι η συνάρτηση σφάλματος E είναι κάτω φραγμένη, αφού $E(w) \geq 0$ σαν άθροισμα τετραγώνων. Για ένα δεδομένο σύνολο προτύπων εισόδου και μια δεδομένη αρχιτεκτονική του ΤΝΔ, εάν υπάρχει ένα σημείο w^* τέτοιο ώστε $E(w^*) = 0$, τότε το w^* είναι σημείο ολικού ελαχίστου. Με άλλα λόγια, το διάνυσμα w με την μικρότερη συναρτησιακή τιμή (με $E(w) > 0$) είναι το ολικό ελάχιστο.

Για ΤΝΔ που χρησιμοποιούν λείες συναρτήσεις ενεργοποίησης (οι παράγωγοι τους τουλάχιστον μέχρι τάξης p υπάρχουν και είναι συνεχείς), όπως για παράδειγμα η υπερβολική εφαιπομένη, η λογιστική συνάρτηση ενεργοποίησης κ.α., η συνάρτηση σφάλματος E είναι επίσης λεία.

Στην εκπαίδευση ΤΝΔ, αν και είναι δυνατό γενικά να ελέγχουμε τις προϋποθέσεις του Θεωρήματος 3.1, αυτό είναι υπολογιστικά δαπανηρό ιδιαίτερα για μεγάλα ΤΝΔ και στην πράξη παραλείπεται.

Γενικά, κάθε μέθοδος εκπαίδευσης ΤΝΔ της κλάσης της μεθόδου της οπισθοδρομικής διάδοσης του σφάλματος, που έχει την μορφή (3.2) μπορεί να αποκτήσει την ιδιότητα της ευρείας σύγκλισης εάν ισχύουν οι ακόλουθες συνθήκες:

- (i) Η συνισταμένη κατεύθυνση ανίχνευσης d^k είναι κατεύθυνση μείωσης και δεν τείνει να γίνει ορθογώνια με την κατεύθυνση του διανύσματος των μερικών παραγώγων της συνάρτησης σφάλματος (το $\cos \theta_k$ στη Σχέση (3.6) πρέπει να είναι θετικό).
- (ii) Ο ρυθμός εκπαίδευσης α^k ικανοποιεί τις συνθήκες του Wolfe (3.3)–(3.4). Να σημειωθεί ότι αφού η κατεύθυνση d^k είναι κατεύθυνση μείωσης, εάν η E είναι συνεχώς παραγωγίσιμη και κάτω φραγμένη κατά μήκος της ακτίνας $\{w^k + \alpha d^k \mid \alpha > 0\}$, τότε πάντα υπάρχει α^k που ικανοποιεί τις συνθήκες του Wolfe (3.3) και (3.4) [30, 95].

Για παράδειγμα, η γνωστή μέθοδος της οπισθοδρομικής διάδοσης του σφάλματος που χρησιμοποιεί την μέθοδο της πιο απότομης καθόδου με ένα κοινό ρυθμό εκπαίδευσης για όλα τα βάρη που να ικανοποιεί τις συνθήκες του Wolfe (3.3)–(3.4) έχει την ιδιότητα της ευρείας σύγκλισης, γιατί στην περίπτωση αυτή ισχύει $\cos \theta_k = 1 > 0$.

Σχετικά όμως με τις μεθόδους οπισθοδρομικής διάδοσης του σφάλματος με διαφορετικό ρυθμό εκπαίδευσης ανά βάρος (τοπικός ρυθμός εκπαίδευσης), από όσο γνωρίζουμε δεν υπάρχει μέθοδος που να τις εξοπλίζει με την ιδιότητα της ευρείας σύγκλισης. Το ακόλουθο θεώρημα παρέχει στις μεθόδους με τοπικό ρυθμό εκπαίδευσης την ιδιότητα της ευρείας σύγκλισης, εξασφαλίζοντας ότι η κατεύθυνση ανίχνευσης που χρησιμοποιούνται είναι πάντα κατεύθυνση μείωσης. Είναι σημαντικό το γεγονός ότι το θεωρητικό αυτό αποτέλεσμα είναι ανεξάρτητο από την μέθοδο με την οποία προσαρμόζονται οι τοπικοί ρυθμοί εκπαίδευσης και μπορεί να χρησιμοποιηθεί για να εγγυηθεί ευρεία σύγκλιση για *οποιοδήποτε* αλγόριθμο ακολουθεί την παρακάτω στρατηγική:

- (i) όρισε τους $(n - 1)$, έστω $\{\eta_1, \eta_2, \dots, \eta_{i-1}, \eta_{i+1}, \dots, \eta_n\}$, από τους n τοπικούς ρυθμούς εκπαίδευσης, από το σύνολο $\{\eta_1, \eta_2, \dots, \eta_n\}$, όπως ακριβώς δίνονται από οποιονδήποτε αλγόριθμο προσαρμογής, και
- (ii) υπολόγισε αναλυτικά τον εναπομείναντα τοπικό ρυθμό εκπαίδευσης (δηλαδή τον ρυθμό εκπαίδευσης η_i), χρησιμοποιώντας τις τιμές των υπολοίπων, όπως φαίνεται παρακάτω στη Σχέση 3.9.

Στη συνέχεια δίνουμε ένα νέο Θεώρημα για την ευρεία σύγκλιση αλγορίθμων με τοπικό ρυθμό εκπαίδευσης και την απόδειξή του.

Θεώρημα 3.2 [80] Έστω ότι οι προϋποθέσεις του Θεωρήματος 3.1 για την $f(x)$ και τις μερικές παραγώγους της $\nabla f(x)$ ισχύουν για την συνάρτηση σφάλματος $E(w)$ και για το διάνυσμα των μερικών παραγώγων της $\nabla E(w)$. Τότε, για κάθε αρχικό σημείο $w^0 \in \mathbb{R}^n$, η ακολουθία σημείων $\{w^k\}_{k=0}^\infty$, που δημιουργείται από το ακόλουθο επαναληπτικό σχήμα:

$$w^{k+1} = w^k + \alpha^k d^k, \quad (3.8)$$

όπου $d^k = -\text{diag}\{\eta_1^k, \dots, \eta_i^k, \dots, \eta_n^k\} \nabla E(w^k)$ δηλώνει την κατεύθυνση ανίχνευσης, και η_m^k για $m = 1, 2, \dots, i-1, i+1, \dots, n$ είναι αυθαίρετα επιλεγμένοι μικροί θετικοί ρυθμοί εκπαίδευσης, ενώ ο i -οστος τοπικός ρυθμός εκπαίδευσης υπολογίζεται σύμφωνα με τον τύπο:

$$\eta_i^k = -\frac{\delta}{\partial_i E(w^k)} - \frac{1}{\partial_i E(w^k)} \sum_{\substack{j=1 \\ j \neq i}}^n \eta_j^k \partial_j E(w^k), \quad 0 < \delta \ll \infty, \quad \partial_i E(w^k) \neq 0, \quad (3.9)$$

και το $\alpha^k > 0$ ικανοποιεί τις συνθήκες του Wolfe (3.3)-(3.4), έχει την ιδιότητα της ευρείας σύγκλισης σε ένα σημείο τοπικού ελαχίστου της συνάρτησης $E(w)$.

Απόδειξη. Είναι προφανές ότι η συνάρτηση σφάλματος E είναι κάτω φραγμένη στο $\mathbb{R}^n$, σαν άθροισμα τετραγώνων. Η ακολουθία σημείων $\{w^k\}_{k=0}^\infty$ ακολουθεί την κατεύθυνση:

$$d^k = -\text{diag}\{\eta_1^k, \dots, \eta_i^k, \dots, \eta_n^k\} \nabla E(w^k),$$

η οποία είναι κατεύθυνση μείωσης εάν η_m^k , $m = 1, 2, \dots, i-1, i+1, \dots, n$ είναι τυχαία επιλεγμένοι ρυθμοί εκπαίδευσης (δηλαδή θετικοί πραγματικοί αριθμοί) και η_i^k δίνεται από την Σχέση (3.9), αφού

$$\nabla E(w^k)^\top d^k < 0.$$

Επιπρόσθετα, ισχύει ότι:

$$\cos \theta_k = \frac{-\nabla E(w^k)^\top d^k}{\|\nabla E(w^k)\| \|d^k\|} > 0. \quad (3.10)$$

Συνεπώς, από τα παραπάνω είναι προφανές ότι η ακολουθία σημείων $\{w^k\}_{k=0}^\infty$, που δημιουργείται από το επαναληπτικό Σχήμα (3.8), συγκλίνει ευρέως σε κάποιο τοπικό ελάχιστο της συνάρτησης σφάλματος $E(w)$. Το Θεώρημα αποδείχθηκε. $\square$

Παρατήρηση 3.2 Στη Σχέση (3.9), επιλέγουμε μια κατεύθυνση έτσι ώστε η μερική παράγωγος να μην είναι μηδέν. Προφανώς πάντα υπάρχει μια τέτοια κατεύθυνση, γιατί διαφορετικά πρέπει να βρισκόμαστε ήδη σε ένα τοπικό ελάχιστο.

Στο Θεώρημα 3.2 έχουμε εισάγει δύο νέες παραμέτρους. Η παράμετρος δ , $0 < \delta \ll \infty$, χρησιμοποιείται για την αποφυγή τυχόν προβλημάτων λόγω της περιορισμένης ακρίβειας των υπολογισμών που συχνά εμφανίζονται στα πειράματα και θα πρέπει να έχει μικρή τιμή, ανάλογη της τετραγωνικής ρίζας της ακρίβειας της μηχανής. Στα πειράματα που παρουσιάζουμε στην επόμενη ενότητα, έχουμε επιλέξει την τιμή $\delta = 10^{-6}$.

Αναφορικά με την παράμετρο α^k , στην πράξη, προτείνεται για τις εφαρμογές η τιμή $\alpha^k = 1$ για όλα τα k . Αυτή η τιμή έχει σαν αποτέλεσμα το βήμα της ελαχιστοποίησης κατά μήκος της συνισταμένης κατεύθυνσης αρχικά να εξαρτάται μόνο από τις τιμές των τοπικών ρυθμών εκπαίδευσης. Το μήκος του βήματος στη συνέχεια μπορεί να ελεγχθεί με κατάλληλη προσαρμογή του α^k έτσι ώστε να ικανοποιούνται οι συνθήκες του Wolfe. Για το σκοπό αυτό μπορούμε να χρησιμοποιήσουμε μια απλή στρατηγική οπισθοδρόμησης για την μείωση του

α^k κατά ένα παράγοντα μείωσης $1/q$, όπου $q > 1$. Αυτό έχει σαν αποτέλεσμα το α^k να μειώνεται κατά τον μεγαλύτερο αριθμό από την ακολουθία $\{q^{-m}\}_{m=1}^{\infty}$ [96]. Η επιλογή του q δεν είναι κρίσιμη για την επιτυχημένη εκπαίδευση, όμως επηρεάζει τον συνολικό αριθμό των υπολογισμών της συνάρτησης σφάλματος μέχρι να ικανοποιηθούν οι συνθήκες του Wolfe. Η τιμή $q = 2$ προτείνεται γενικά στη βιβλιογραφία [7, 96] και πράγματι βρέθηκε να μην παρουσιάζει προβλήματα στα πειράματα που πραγματοποιήσαμε.

Τέλος, σχετικά με τις συνθήκες του Wolfe (3.3)–(3.4), πρέπει να παρατηρήσουμε ότι στην περίπτωση της εκπαίδευσης ΤΝΔ, η ανισότητα (3.3) εξασφαλίζει ότι η τιμή της συνάρτησης σφάλματος μειώνεται ικανοποιητικά σε κάθε επανάληψη, ενώ η ανισότητα (3.4) αποτρέπει την χρήση πολύ μικρών βημάτων ελαχιστοποίησης. Συνεπώς, στην προσπάθειά μας να ικανοποιήσουμε την συνθήκη (3.3) είναι σημαντικό να βεβαιωθούμε ότι το α^k δεν μειώνεται περισσότερο από όσο χρειάζεται, με αποτέλεσμα να μην ικανοποιείται η συνθήκη (3.4).

Στην k -στη επανάληψη το διάνυσμα των μερικών παραγώγων της συνάρτησης σφάλματος υπολογίζεται στην αρχή της επανάληψης για τον υπολογισμό του νέου διανύσματος των βαρών, w^{k+1} . Έτσι, η συνθήκη (3.4) δεν μπορεί να ελεγχθεί άμεσα, αφού αυτό θα απαιτούσε ένα επιπλέον υπολογισμό του διανύσματος των μερικών παραγώγων. Το πρόβλημα αυτό μπορεί να λυθεί εύκολα (βλ. και [30]) αντικαθιστώντας την δεύτερη συνθήκη (3.4), με την ακόλουθη σχέση:

$$E(w^k + \alpha^k d^k) - E(w^k) \geq \sigma_2 \alpha^k \nabla E(w^k)^\top d^k, \quad (3.11)$$

και συνεπώς αποφεύγουμε τον πρόσθετο υπολογιστικό κόπο.

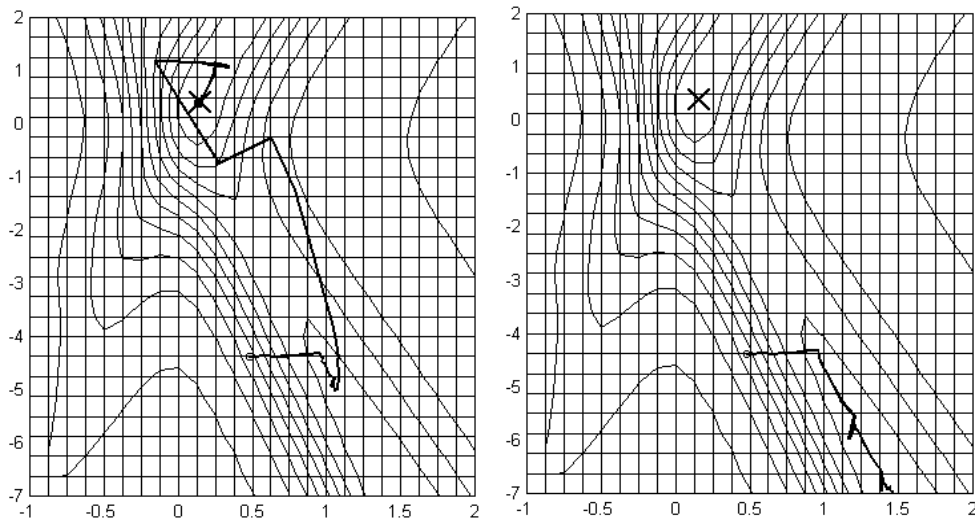
Στο σημείο αυτό είναι χρήσιμο να απεικονίσουμε την συμπεριφορά της προτεινόμενης τεχνικής σε ένα απλό πρόβλημα εκπαίδευσης, το οποίο αφορά την περίπτωση ενός μόνο νευρώνα με δύο βάρη και λογιστική συνάρτηση ενεργοποίησης [82]. Αυτό το απλό ΤΝΔ εκπαιδεύεται πρώτα χρησιμοποιώντας τη μέθοδο Quickprop και στη συνέχεια την ευρέως συγκλίνουσα τροποποίησή της. Η τροποποιημένη μέθοδος χρησιμοποιεί τον θετικό ρυθμό εκπαίδευσης η_1^k , όπως υπολογίζεται από την μέθοδο Quickprop και υπολογίζει τον η_2^k σύμφωνα με τη Σχέση (3.9). Ξεκινώντας από το ίδιο αρχικό σημείο, η ευρέως συγκλίνουσα τροποποιημένη Quickprop εντόπισε με επιτυχία το επιθυμητό ελάχιστο (Σχήμα 3.1, αριστερά), ενώ η κλασική Quickprop μέθοδος (Σχήμα 3.1, δεξιά) δημιουργεί μια τροχιά στο χώρο των βαρών που οδηγεί σε ένα ανεπιθύμητο ακρότατο.

Στο Σχήμα 3.2 παρουσιάζουμε μια τυπική γραφική παράσταση της μείωσης των τιμών της συνάρτησης σφάλματος για το πρόβλημα της ισοτιμίας 3-bit (βλ. Παράρτημα Α.2 και [50, 133]), ξεκινώντας από το ίδιο αρχικό σημείο. Με συνεχή γραμμή βλέπουμε την μέθοδο Quickprop και με διακεκομμένη γραμμή την ευρέως συγκλίνουσα τροποποίησή της. Είναι φανερό ότι η τροποποιημένη μέθοδος καταφέρνει με επιτυχία να εντοπίσει ένα επιθυμητό ελάχιστο ($E(w) \leq 10^{-10}$), ενώ η μέθοδος Quickprop παγιδεύεται σε κάποιο ανεπιθύμητο τοπικό ελάχιστο με μεγαλύτερη συναρτησιακή τιμή.

3.4 Αποτελέσματα των Προσομοιώσεων

Σε αυτή την ενότητα παρουσιάζουμε αποτελέσματα από την κλασική μέθοδο των Silva-Almeida [145] και την κλασική μέθοδο Quickprop [35], καθώς επίσης και αποτελέσματα των τροποποιήσεών τους. Έχουμε λοιπόν εξοπλίσει τους δύο αυτούς αλγόριθμους πρώτης τάξης με τη στρατηγική του Θεωρήματος 3.2 με σκοπό η τροποποιημένη νέα μέθοδος να έχει την ιδιότητα της ευρείας σύγκλισης. Όλες οι μέθοδοι έχουν υλοποιηθεί και συγκριθεί ως προς τον απαιτούμενο αριθμό συναρτησιακών υπολογισμών, υπολογισμών του διανύσματος των μερικών παραγώγων, καθώς επίσης και ως προς την επιτυχία τους.

Τα αποτελέσματά μας δείχνουν ότι οι προτεινόμενες στρατηγικές έχουν την αναμενόμενη συμπεριφορά και σταθερότητα. Παρακάτω θα παρουσιάσουμε αναλυτικά αποτελέσματα από



Σχήμα 3.1: Απεικόνιση της μεθόδου Quickprop για την εκπαίδευση ενός απλού ΤΝΔ με δύο βάρη (με x σημειώνεται το ελάχιστο). Η τροποποιημένη μέθοδος συγκλίνει στο επιθυμητό ελάχιστο (αριστερά), ενώ η κλασική μέθοδος συγκλίνει σε ένα ανεπιθύμητο ακρότατο (δεξιά)

100 προσομοιώσεις των μεθόδων Silva-Almeida, Quickprop και των τροποποιήσεών τους σε τρία προβλήματα εκπαίδευσης, χρησιμοποιώντας τα ίδια αρχικά βάρη που είχαν τυχαία επιλεγεί από την ομοιόμορφη κατανομή στο διάστημα $[-1, 1]$. Ονομάζουμε τις νέες μεθόδους G-Silva-Almeida και G-Quickprop, αντίστοιχα. Οι κλασικές μέθοδοι και οι τροποποιήσεις τους χρησιμοποιούν ακριβώς τις ίδιες παραμέτρους εκπαίδευσης και συγκρίνονται κάτω από τις ίδιες συνθήκες.

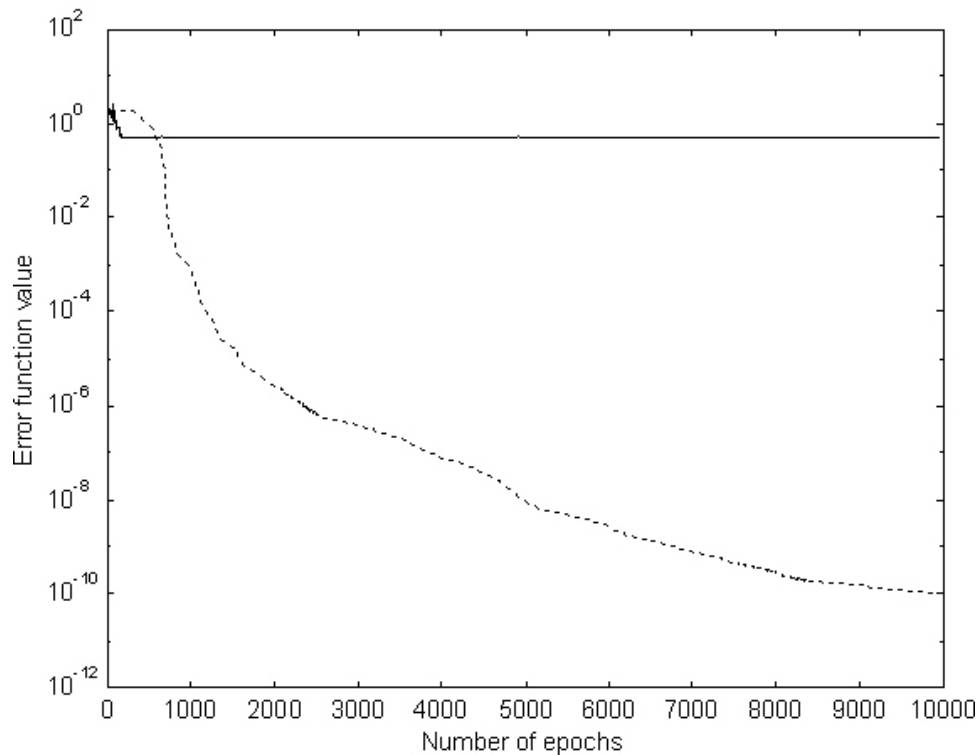
Η ευρετική παράμετρος, *μέγιστος παράγοντας αύξησης* (maximum growth factor) της μεθόδου Quickprop έχει την προτεινόμενη τιμή $m = 1.75$ [35]. Οι παράμετροι αύξησης και μείωσης του ρυθμού εκπαίδευσης (learning rate increment and decrement factors) της μεθόδου Silva-Almeida ρυθμίστηκαν κατάλληλα και οι τιμές που έλαβαν είναι $u = 1.02$ και $d = 0.5$, αντίστοιχα. Αξίζει να σημειωθεί ότι σε όλα τα πειράματα που αναφέρουμε εδώ, η Σχέση (3.9) εφαρμόστηκε κυκλικά στους ρυθμούς εκπαίδευσης, δηλαδή στην k -οστή επανάληψη ήταν $i = k \bmod n$.

3.4.1 Αναγνώριση αριθμών

Το πρόβλημα αναγνώρισης των αριθμών (βλ. και Παράρτημα Α.7) είναι το πρώτο μας πείραμα. Ένα ΤΝΔ με 64 νευρώνες εισόδου, 6 κρυφούς νευρώνες και 10 νευρώνες εξόδου (συνολικά 444 βάρη και 16 πολώσεις) εκπαιδεύτηκε για να αναγνωρίζει τους αριθμούς από το 0 έως το 9 [82]. Το ΤΝΔ βασίζεται στη λογιστική συνάρτηση ενεργοποίησης.

Τα αποτελέσματα των μεθόδων παρουσιάζονται στους Πίνακες 3.1 και 3.2, όπου: μ συμβολίζει το μέσο αριθμό των υπολογισμών της συνάρτησης σφάλματος ή του διανύσματος των μερικών της παραγώγων, σ συμβολίζει την αντίστοιχη τυπική απόκλιση, Min/Max συμβολίζει τον ελάχιστο/μέγιστο αριθμό των υπολογισμών της συνάρτησης σφάλματος ή του διανύσματος των μερικών της παραγώγων, D υποδηλώνει ότι ο αλγόριθμος απέκλινε σε όλες τις δοκιμές, και % είναι το ποσοστό επιτυχίας, δηλαδή το ποσοστό των προσομοιώσεων που συνέκλιναν σε κάποιο επιθυμητό ελάχιστο.

Η συνθήκη τερματισμού για όλους τους αλγόριθμους ήταν να βρεθεί τιμή της συνάρτησης σφάλματος: $E \leq 10^{-1}$ στην πρώτη περίπτωση (Πίνακας 3.1), και $E \leq 10^{-2}$ στη δεύτερη περίπτωση (Πίνακας 3.2). Και στις δύο περιπτώσεις ο μέγιστος επιτρεπτός αριθμός



Σχήμα 3.2: Τυπική γραφική παράσταση της μείωσης των τιμών της συνάρτησης σφάλματος για το πρόβλημα της ισοτιμίας των 3-bit, ξεκινώντας από το ίδιο αρχικό σημείο. Με συνεχή γραμμή βλέπουμε την μέθοδο Quickprop και με διακεκομμένη γραμμή την ευρέως συγκλίνουσα τροποποίησή της

υπολογισμών της συνάρτησης σφάλματος ήταν 5000.

Όπως φαίνεται στον Πίνακα 3.1 η τροποποιημένη Quickprop μέθοδος με την ιδιότητα της ευρείας σύγκλισης (G-Quickprop) είναι ταχύτερη και πιο αξιόπιστη από την κλασική Quickprop μέθοδο που απέτυχε να συγκλίνει σε όλες τις περιπτώσεις. Στο ίδιο πρόβλημα, η μέθοδος των Silva-Almeida, αν και είναι ταχύτερη από την τροποποιημένη, αποτυγχάνει να συγκλίνει σε κάποιο επιθυμητό ελάχιστο σε 43 από τα 100 πειράματα. Αυτό ήταν το αποτέλεσμα μιας αστάθειας της μεθόδου, λόγω της εκθετικής αύξησης του ρυθμού εκπαίδευσης σε διαδοχικές επαναλήψεις. Αυτή η συμπεριφορά έχει σαν επακόλουθο μεγάλα βήματα ελαχιστοποίησης κατά μήκος κάποιων κατευθύνσεων, που επιβάλλουν μεγάλες τροποποιήσεις στα αντίστοιχα βάρη, αναγκάζοντας τους νευρώνες εξόδου σε κορεσμό και κατά συνέπεια τη μέθοδο να συγκλίνει σε κάποιο ανεπιθύμητο ακρότατο ή να αποκλίνει εντελώς. Από την άλλη μεριά, η τροποποιημένη μέθοδος G-Silva-Almeida ξεπερνά με επιτυχία αυτά τα προβλήματα, αφού υπολογίζει αναλυτικά τον i -στο ρυθμό εκπαίδευσης σύμφωνα με το Θεώρημα 3.2. Έτσι έχει μόνο 1% ποσοστό αποτυχίας, λόγω του ότι ξεπέρασε το ανώτερο όριο των υπολογισμών της συνάρτησης σφάλματος.

Συγκρίνοντας τους Πίνακες 3.1 και 3.2 βλέπουμε ότι οι τροποποιημένες μέθοδοι έχουν συνεπή και προβλέψιμη συμπεριφορά, ενώ η απόδοση της κλασικής μεθόδου των Silva-Almeida επιδεινώνεται καθώς αυξάνεται η ακρίβεια με την οποία ζητάμε την λύση. Η προσαρμογή του ρυθμού εκπαίδευσης από την μέθοδο Silva-Almeida έχει σαν αποτέλεσμα τη γρήγορη σύγκλιση σε 26 από τα πειράματα, αλλά στα υπόλοιπα δεν επέτυχε να εντοπίσει κάποιο ελάχιστο με την επιθυμητή ακρίβεια. Έτσι στον Πίνακα 3.2 το ποσοστό επιτυχίας της μεθόδου Silva-Almeida παρουσιάζεται μειωμένο λόγω σύγκλισης σε ανεπιθύμητα τοπικά ελάχιστα. Πρέπει να σημειωθεί ότι τα αποτελέσματα των Πινάκων 3.1 και 3.2 παράχθη-

Πίνακας 3.1: Αποτελέσματα από το πρόβλημα αναγνώρισης αριθμών ($E \leq 10^{-1}$)

Αλγόριθμος	Συναρτησιακοί Υπολογισμοί			Υπολογισμοί μερικών παραγώγων			Επιτυχία %
	μ	σ	Min/Max	μ	σ	Min/Max	
Silva-Almeida	124.21	10.557	109/151	124.21	10.557	109/151	57
G-Silva-Almeida	403.21	114.352	145/798	711.92	298.395	148/1428	99
Quickprop	D	D	D	D	D	D	0
G-Quickprop	82.42	77.297	26/485	172.31	202.836	26/1023	99

Πίνακας 3.2: Αποτελέσματα από το πρόβλημα αναγνώρισης αριθμών ($E \leq 10^{-2}$)

Αλγόριθμος	Συναρτησιακοί Υπολογισμοί			Υπολογισμοί μερικών παραγώγων			Επιτυχία %
	μ	σ	Min/Max	μ	σ	Min/Max	
Silva-Almeida	218.23	9.774	204/237	218.23	9.774	204/237	26
G-Silva-Almeida	712.92	235.478	335/1526	1423.41	796.503	335/4556	99
Quickprop	D	D	D	D	D	D	0
G-Quickprop	160.06	147.197	35/641	372.36	443.809	35/1778	100

καν χρησιμοποιώντας τις ίδιες αρχικές συνθήκες για όλους τους αλγόριθμους· το μόνο που άλλαξε είναι η επιθυμητή ακρίβεια, που από $E \leq 10^{-1}$ έγινε $E \leq 10^{-2}$.

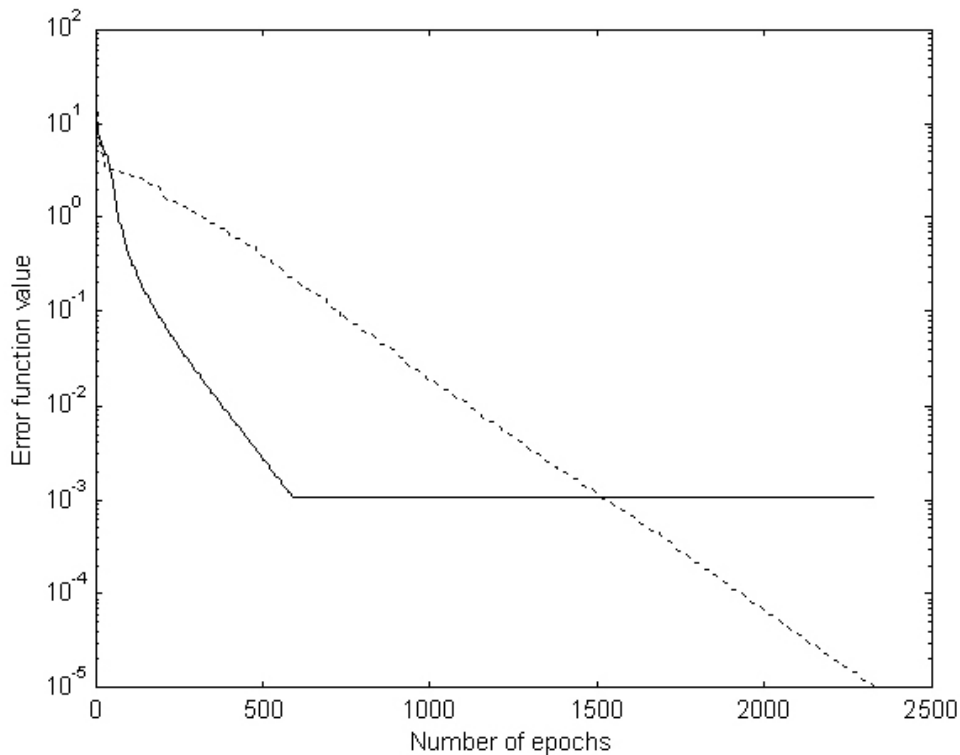
Μια τυπική γραφική παράσταση της μείωσης των τιμών της συνάρτησης σφάλματος, για την κλασική και την τροποποιημένη μέθοδο Silva-Almeida, παρουσιάζουμε στο Σχήμα 3.3. Στην περίπτωση αυτή εκπαιδεύουμε εξαντλητικά το ΤΝΔ, ξεκινώντας τις μεθόδους από το ίδιο αρχικό σημείο. Με συνεχή γραμμή βλέπουμε την μέθοδο Silva-Almeida και με διακεκομμένη γραμμή την ευρέως συγκλίνουσα τροποποίησή της. Είναι φανερό ότι η τροποποιημένη μέθοδος καταφέρνει με επιτυχία να εντοπίσει ένα επιθυμητό ελάχιστο ($E(w) \simeq 10^{-5}$), ενώ η μέθοδος Silva-Almeida παγιδεύεται σε κάποιο ανεπιθύμητο τοπικό ελάχιστο με μεγαλύτερη συναρτησιακή τιμή.

3.4.2 Προσέγγιση μιας συνεχούς συνάρτησης

Το δεύτερο πρόβλημα που εξετάζουμε είναι αυτό της προσέγγισης μιας συνεχούς συνάρτησης (βλ. και Παράρτημα Α.5). Η συνεχής συνάρτηση $f(x) = \sin(x) \cos(2x)$ προσεγγίζεται από ένα 1-15-1 ΤΝΔ (30 βάρη και 16 πολώσεις), χρησιμοποιώντας 20 σημεία από το διάστημα $[0, 2\pi]$. Η συνθήκη τερματισμού ήταν να βρεθεί ελάχιστο με τιμή $E \leq 0.1$, με μέγιστο αριθμό υπολογισμών της συνάρτησης σφάλματος 10000. Το ΤΝΔ βασίστηκε σε κρυφούς νευρώνες που χρησιμοποιούν την υπερβολική εφαπτομένη και σε γραμμικούς νευρώνες εξόδου. Τα συγκριτικά αποτελέσματα παρουσιάζονται στον Πίνακα 3.3.

Η μέθοδος των Silva-Almeida παρουσιάζει το μικρότερο ποσοστό επιτυχίας, λόγω σύγκλισης σε ανεπιθύμητα τοπικά ελάχιστα και καταφέρνει να συγκλίνει μόνο 11 φορές, αν και χρησιμοποιήσαμε τις καλύτερες δυνατές τιμές για τις ευρετικές της παραμέτρους. Στο ίδιο πρόβλημα, η τροποποιημένη μέθοδος Quickprop ξεπερνά την κλασική μέθοδο Quickprop στον αριθμό των επιτυχιών, αφού έχει 99% επιτυχία και η Quickprop έχει μόλις 27%.

Σε πρόσθετα πειράματα με τις ίδιες αρχικές συνθήκες και αλλαγή της επιθυμητής ακρίβειας σε $E \leq 10^{-2}$, οι αλγόριθμοι είχαν παρόμοια συμπεριφορά με το πρόβλημα της αναγνώρισης αριθμών και οι τροποποιημένες μέθοδοι είχαν την αναμενόμενη συμπεριφορά και 100% επιτυχία.



Σχήμα 3.3: Τυπική γραφική παράσταση της μείωσης των τιμών της συνάρτησης σφάλματος για το πρόβλημα της αναγνώρισης αριθμών, ξεκινώντας από το ίδιο αρχικό σημείο. Με συνεχή γραμμή βλέπουμε την μέθοδο Silva-Almeida και με διακεκομμένη γραμμή την ευρέως συγκλίνουσα τροποποίησή της

3.4.3 Αναγνώριση ανωμαλιών σε κολονοσκοπήσεις

Το πρόβλημα της αναγνώρισης ανωμαλιών σε κολονοσκοπήσεις χρησιμοποιήθηκε για να ελέγξουμε την επιρροή των αρχικών τιμών των βαρών στην διαδικασία της εκπαίδευσης του ΤΝΔ. Για το λόγο αυτό ένα 16-40-2 ΤΝΔ (720 βάρη και 42 πολώσεις), με τη λογιστική συνάρτηση ενεργοποίησης, εκπαιδεύτηκε να αναγνωρίζει ανωμαλίες σε κολονοσκοπήσεις. Οι ανωμαλίες ήταν μακροσκοπικά Τύπου III και Τύπου V. Στο Σχήμα 3.4 (πάνω αριστερά) φαίνεται μια ανωμαλία που είναι Τύπου III, ενώ στο Σχήμα 3.4 (κάτω αριστερά) απεικονίζεται μια ανωμαλία Τύπου V [67].

Τα πρότυπα εξάχθηκαν από εικόνες 10 φυσιολογικών και 10 καρκινικών ιστών και η εκπαίδευση σταμάτησε όταν το σφάλμα ταξινόμησης του ΤΝΔ έγινε 3% (για τις τεχνικές λεπτομέρειες της εξαγωγής των προτύπων, βλ. την εργασία [58]). Κάναμε 100 πειράματα χρησιμοποιώντας την μέθοδο Quickprop και την τροποποίησή της, αρχικοποιώντας τα βάρη με τιμές από την ομοιόμορφη κατανομή από έξι διαφορετικά διαστήματα.

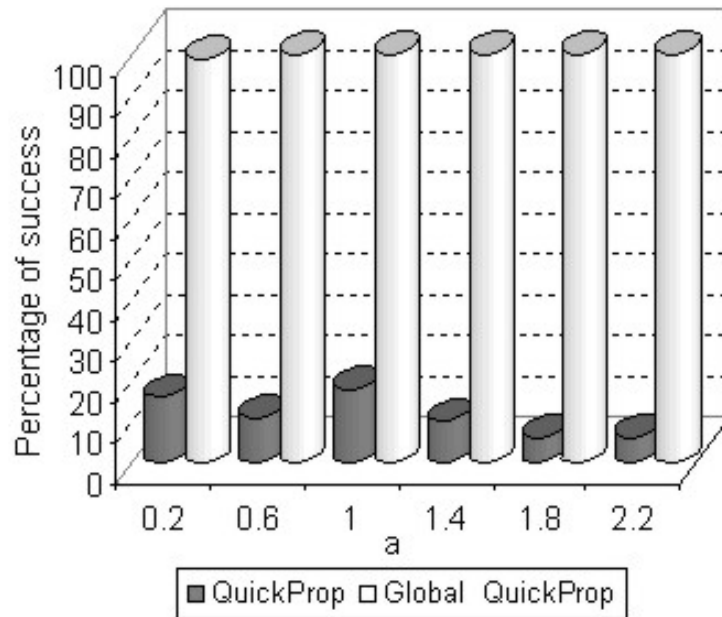
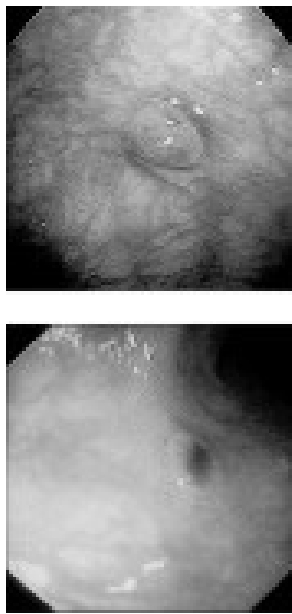
Στο Σχήμα 3.4 (δεξιά) απεικονίζεται η γραφική παράσταση του ποσοστού επιτυχίας σχετικά με τα έξι διαφορετικά διαστήματα, $(-a, a)$, αρχικοποίησης των βαρών, όπου $a \in \{0.2, 0.6, 1, 1.4, 1.8, 2.2\}$. Είναι προφανές ότι η μέθοδος G-Quickprop παρουσιάζεται σημαντικά καλύτερη από την κλασική μέθοδο σε όλα τα πειράματά μας.

3.5 Συμπεράσματα – Συνεισφορά

Σε αυτό το κεφάλαιο προτείναμε και αποδείξαμε ένα νέο θεωρητικό αποτέλεσμα που υποστηρίζει την ανάπτυξη αιτιοκρατικών αλγορίθμων εκπαίδευσης ΤΝΔ ευρείας σύγκλισης

Πίνακας 3.3: Αποτελέσματα από το πρόβλημα προσέγγισης μιας συνεχούς συνάρτησης

Αλγόριθμος	Συναρτησιακοί Υπολογισμοί			Υπολογισμοί μερικών παραγώγων			Επιτυχία %
	μ	σ	Min/Max	μ	σ	Min/Max	
Silva-Almeida	23.11	116.18	84/150	23.11	116.18	84/150	11
G-Silva-Almeida	352.44	105.21	48/764	688.26	197.02	48/2354	99
Quickprop	362.81	268.55	58/953	362.81	268.55	58/953	27
G-Quickprop	176.51	119.98	40/694	252.10	179.31	50/1033	99

**Σχήμα 3.4:** Οι εικόνες για το πρόβλημα αναγνώρισης ανωμαλιών σε κολονοσκοπήσεις (αριστερά). Το ποσοστό επιτυχίας σε σχέση με διάστημα αρχικοποίησης των βαρών $(-a, a)$, (δεξιά).

με τοπικούς ρυθμούς εκπαίδευσης. Το θεώρημα παρέχει σε οποιαδήποτε μέθοδο την ιδιότητα της ευρείας σύγκλισης, με την προϋπόθεση να ακολουθεί την τεχνική προσαρμογής της κατεύθυνσης ανίχνευσης και ρύθμισης του ρυθμού εκπαίδευσης του Θεωρήματος 3.2.

Στη συνέχεια, εξετάστηκαν και συγκρίθηκαν δύο πολύ γνωστοί αλγόριθμοι με τις ευρέως συγκλίνουσες τροποποιήσεις τους. Οι νέες μέθοδοι, σύμφωνα με τα αποτελέσματά μας, έχουν σημαντικά βελτιωμένα ποσοστά επιτυχίας και είναι ικανές να ανιχνεύσουν ελάχιστα με μεγαλύτερη ακρίβεια. Βέβαια απαιτούν πρόσθετους υπολογισμούς της συνάρτησης σφάλματος και του διανύσματος των μερικών της παραγώγων, όπως παρατηρήσαμε για παράδειγμα στην τροποποιημένη Silva-Almeida μέθοδο.

Τα αποτελέσματα δείχνουν ότι η στρατηγική του Θεωρήματος 3.2 έχει την αναμενόμενη συμπεριφορά και πρακτική εφαρμογή, αφού αυξάνει σημαντικά τη σθεναρότητα και την πιθανότητα επιτυχίας. Σύμφωνα με την εμπειρία μας, η αύξηση του ποσοστού επιτυχίας, ιδιαίτερα σε πραγματικά προβλήματα, δικαιολογεί τον ελάχιστο πρόσθετο προγραμματιστικό κόπο για την υλοποίηση της στρατηγικής του Θεωρήματος 3.2.

Μαθηματική Θεμελίωση της Μεθόδου Quickprop και μια Νέα Τροποποίησή της

Αυτός που αγαπάει μόνο την Πράξη
χωρίς τη Θεωρία,
είναι σαν τον ναυτικό που σαλπάρει
χωρίς πηδάλιο και πυξίδα, και ποτέ
δεν ξέρει προς τα που να στραφεί.

—Leonardo da Vinci (1452–1519)

Σε αυτό το κεφάλαιο θα παρουσιάσουμε την γνωστή μέθοδο εκπαίδευσης ΤΝΔ Quickprop [35, 36] και ένα μαθηματικό πλαίσιο για την ανάλυση της σύγκλισής της [165, 166]. Επιπλέον, προτείνουμε μια τροποποίηση αυτής της μεθόδου που παρουσιάζει αυξημένη ταχύτητα και σταθερότητα σύγκλισης, και, ταυτόχρονα, επιλύει το δύσκολο πρόβλημα της επιλογής των ευρετικών παραμέτρων της μεθόδου. Τα αποτελέσματα των πειραμάτων μας δείχνουν ότι τα αυξημένα ποσοστά σύγκλισης που επιτυγχάνονται από τον προτεινόμενο αλγόριθμο, δεν επηρεάζουν σε καμιά περίπτωση την ικανότητα και τη σταθερότητα γενίκευσής του.

4.1 Εισαγωγή

Η μέθοδος Quickprop (ή Qprop) [35, 36] είναι μια πολύ δημοφιλής μέθοδος εκπαίδευσης ΤΝΔ ανά ομάδα προτύπων εισόδου, λόγω της ταχύτητας σύγκλισης που παρουσιάζει. Επίσης, όπως είναι γνωστό, μακριά από τη γειτονιά ενός ελάχιστου η μορφολογία του χώρου των βαρών σε ορισμένες περιπτώσεις αναγκάζει την μέθοδο Qprop να αποκλίνει, να έχει χαμηλά ποσοστά επιτυχίας και γενικά η μέθοδος παρουσιάζει προβλήματα σταθερότητας. Έτσι, για να αντιμετωπιστεί αυτό το πρόβλημα χρησιμοποιούνται ευρετικές παράμετροι, που εξαρτώνται από την εκάστοτε εφαρμογή.

Σε αυτό το κεφάλαιο κάνουμε μια εισαγωγή στη μέθοδο Qprop και δείχνουμε ότι είναι μια γενίκευση της μεθόδου χορδής για μη γραμμικές εξισώσεις, που εφαρμόζεται στην κλίση της συνάρτησης σφάλματος. Επιπλέον, παρουσιάζουμε μια τροποποίηση αυτού του αλγορίθμου που έχει βελτιωμένη ταχύτητα σύγκλισης και σταθερότητα, και συγχρόνως δεν απαιτεί τη χρήση δύσκολων στην επιλογή ευρετικών παραμέτρων εκπαίδευσης. Τέλος, αποδεικνύουμε ένα θεώρημα για τη σύγκλιση της τροποποιημένης αυτής μεθόδου.

4.2 Μέθοδοι Χορδής

Έστω $F = (f_1, f_2, \dots, f_n) : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ είναι μια Fréchet-παραγωγίσιμη συνάρτηση και x^* είναι μια λύση του μη γραμμικού συστήματος εξισώσεων:

$$F(x) = \Theta^n \equiv (0, 0, \dots, 0), \quad (4.1)$$

στο εσωτερικό της περιοχής $\mathcal{D}$.

Η πιο γνωστή μέθοδος για την αριθμητική προσέγγιση του x^* είναι η μέθοδος του Newton. Δεδομένης μια αρχικής προσέγγισης x^0 , η μέθοδος του Newton υπολογίζει μια ακολουθία από σημεία $\{x^k\}_{k=0}^\infty$, επιλύοντας την ακόλουθη εξίσωση του Newton:

$$F'(x^k) (x^{k+1} - x^k) = -F(x^k). \quad (4.2)$$

Εάν το x^0 είναι αρκετά κοντά στη λύση x^* , F είναι συνεχώς παραγωγίσιμη σε μια γειτονιά του x^* και η Ιακωβιανή $F'(x^*)$ δεν είναι ιδιάζουσα (singular), τότε η ακολουθία των σημείων $\{x^k\}$ της μεθόδου του Newton συγκλίνει τετραγωνικά στο x^* . Επιπροσθέτως, κάτω από τις ίδιες συνθήκες, κάθε ακολουθία σημείων $\{y^k\}$ που συγκλίνει στο x^* υπεργραμμικά είναι στενά συνδεδεμένη με τη μέθοδο του Newton από το γεγονός ότι η σχετική διαφορά μεταξύ $y^{k+1} - y^k$ και η διόρθωση της μεθόδου Newton $-F'(y^k)^{-1}F(y^k)$ θα τείνουν στο μηδέν [27].

Αν και η τετραγωνική σύγκλιση της μεθόδου Newton είναι μια επιθυμητή ιδιότητα, εντούτοις η μέθοδος εξαρτάται από την αρχική προσέγγιση και απαιτεί γενικά $(n^2 + n)$ συναρτησιακούς υπολογισμούς καθώς και την επίλυση ενός $n \times n$ γραμμικού συστήματος σε κάθε επανάληψη.

Για το λόγο αυτό αναπτύχθηκαν οι quasi-Newton μέθοδοι που απαιτούν λιγότερο υπολογιστικό κόπο, ενώ διατηρούν κάποιες από τις ιδιότητες σύγκλισης της μεθόδου Newton. Έτσι, με διαδοχικές επαναλήψεις υπολογίζουν προσεγγίσεις του x^* και της Ιακωβιανής στη λύση $F'(x^*)$. Εάν x^k είναι η τρέχουσα προσέγγιση της λύσης και B_k η τρέχουσα προσέγγιση της Ιακωβιανής, μετά από τον υπολογισμό του x^{k+1} , η B_k ανασχηματίζεται για να διαμορφωθεί η επόμενη προσέγγιση B_{k+1} . Η μέθοδος κατασκευής της προσέγγισης B_{k+1} καθορίζει και το είδος της quasi-Newton μεθόδου. Δεδομένου του αρχικού σημείου x^0 , η μέθοδος υπολογίζει την ακολουθία των σημείων $\{x^k\}_{k=0}^\infty$, επιλύοντας την ακόλουθη εξίσωση (που ονομάζεται quasi-Newton ή secant) [30]:

$$B_{k+1} (x^{k+1} - x^k) = F(x^{k+1}) - F(x^k). \quad (4.3)$$

Τα πλεονεκτήματα των quasi-Newton μεθόδων είναι ότι απαιτούν μόνο n συναρτησιακούς υπολογισμούς για κάθε επανάληψη. Ως εκ τούτου, εάν μπορεί να βρεθεί μια καλή αρχική προσέγγιση της $F'(x^*)$, τότε οι μέθοδοι αυτές πλεονεκτούν από την άποψη των απαιτούμενων συναρτησιακών υπολογισμών σε σύγκριση με την μέθοδο Newton. Στις περισσότερες quasi-Newton μεθόδους οι παράγωγοι δεν υπολογίζονται σε κάθε επανάληψη. Βέβαια, ο τοπικός ρυθμός σύγκλισης τείνει να είναι υπεργραμμικός αντί τετραγωνικός για τις περισσότερες από αυτές τις μεθόδους.

Μια από τις πιο γνωστές προσεγγίσεις της Ιακωβιανής είναι αυτή που πρότεινε ο Broyden [20]. Η μέθοδος του Broyden έχει τοπικά υπεργραμμική σύγκλιση και έτσι θεωρείται μια καλή εναλλακτική μέθοδος. Ο αλγόριθμος του Broyden για την επίλυση του (4.1) έχει την ακόλουθη γενική μορφή [30]. Δεδομένου του αρχικού σημείου x^0 και ενός, μη ιδιάζοντα, πίνακα B_0 , η μέθοδος παράγει μια ακολουθία από βήματα s_k σύμφωνα με τα ακόλουθα:

Για $k = 0, 1, \dots$ μέχρι η μέθοδος να συγκλίνει:

Λύσε το $B_k s_k = -F(x^k)$ για s_k ,

Θέσε $x^{k+1} = x^k + s_k$,

$$\begin{aligned}\Theta\acute{\epsilon}\sigma\epsilon\ z^k &= F(x^{k+1}) - F(x^k), \\ \Theta\acute{\epsilon}\sigma\epsilon\ B_{k+1} &= B_k + \frac{\left(z^k - B_k s_k\right) s_k^\top}{s_k^\top s_k}.\end{aligned}$$

Η μέθοδος του Broyden είναι πολύ δημοφιλής στην πράξη για δύο βασικούς λόγους: (α) απαιτεί γενικά λιγότερους συναρτησιακούς υπολογισμούς από την μέθοδο Newton με πεπερασμένες διαφορές, και (β) μπορεί να υλοποιηθεί έτσι ώστε να απαιτεί μόνο $O(n^2)$ πράξεις ανά επανάληψη [29, σελ. 27-29].

4.3 Η Μέθοδος Quicprop

Σε αυτή την ενότητα περιγράφουμε την μέθοδο Qprop και δείχνουμε ότι ανήκει στην οικογένεια των μεθόδων χορδής (secant methods). Είναι γνωστό ότι στα προβλήματα ελαχιστοποίησης όλα τα τοπικά ελάχιστα w^* μιας συνεχώς παραγωγίσιμης συνάρτησης σφάλματος E ικανοποιούν την αναγκαία συνθήκη:

$$\nabla E(w^*) = \Theta^n, \quad (4.4)$$

όπου ∇E είναι η κλίση της συνάρτησης σφάλματος E . Η εξίσωση (4.4) αντιπροσωπεύει ένα σύστημα n μη γραμμικών εξισώσεων, που η λύση τους δίνει το w^* . Έτσι μια προσέγγιση του προβλήματος της ελαχιστοποίησης της συνάρτησης σφάλματος E είναι το επιλυθεί το σύστημα (4.4), με την προϋπόθεση να εξασφαλιστεί ότι η λύση που βρίσκεται πράγματι αντιστοιχεί σε ελάχιστο της συνάρτησης σφάλματος. Αυτό είναι ανάλογο με την επίλυση του ακόλουθου συστήματος εξισώσεων:

$$\begin{aligned}\partial_1 E(w_1, w_2, \dots, w_n) &= 0, \\ \partial_2 E(w_1, w_2, \dots, w_n) &= 0, \\ &\vdots \\ \partial_n E(w_1, w_2, \dots, w_n) &= 0,\end{aligned} \quad (4.5)$$

όπου $\partial_i E$ συμβολίζει την i -στη συντεταγμένη του διανύσματος κλίσης ∇E .

Στο κλασικό επαναληπτικό σχήμα της μεθόδου Qprop, όπως δόθηκε στο [35], η τιμή του i -σπου βάρους δίνεται από την ακόλουθη σχέση:

$$w_i^{k+1} = w_i^k - \left\{ \frac{\partial_i E(w^k) - \partial_i E(w^{k-1})}{w_i^k - w_i^{k-1}} \right\}^{-1} \partial_i E(w^k).$$

Χρησιμοποιώντας συμβολισμό πινάκων, η παραπάνω σχέση γράφεται ως εξής:

$$w^{k+1} = w^k - B_k^{-1} \nabla E(w^k),$$

όπου ο πίνακας B_k είναι ένας διαγώνιος πίνακας όπου τα στοιχεία του $[b_{ii}^k]$, $i = 1, 2, \dots, n$ δίνονται ως εξής:

$$b_{ii}^k = \frac{\partial_i E(w^k) - \partial_i E(w^{k-1})}{w_i^k - w_i^{k-1}}.$$

Είναι προφανές ότι ο πίνακας B_k ικανοποιεί την ακόλουθη εξίσωση χορδής:

$$B_k (w^k - w^{k-1}) = \nabla E(w^k) - \nabla E(w^{k-1}), \quad (4.6)$$

και έτσι η μέθοδος Qprop ανήκει στη κατηγορία μεθόδων quasi-Newton.

Χρησιμοποιώντας το πλαίσιο αυτό, παρουσιάζουμε την παρακάτω τροποποίηση της μεθόδου Quickprop:

$$w_i^{k+1} = w_i^k - \eta_i \left\{ \frac{\partial_i E(w^k) - \partial_i E(w^{k-1})}{w_i^k - w_i^{k-1}} \right\}^{-1} \partial_i E(w^k),$$

όπου η_i είναι αυθαίρετοι μη μηδενικοί πραγματικοί αριθμοί. Η τροποποιημένη μέθοδος ικανοποιεί την εξίσωση χορδής και συνεπώς ανήκει και αυτή στη κατηγορία μεθόδων quasi-Newton. Βασίζόμενοι στην ανωτέρω ανάλυση είναι πλέον φανερό ότι η μέθοδος Qprop, καθώς επίσης και η ανωτέρω τροποποίησή της, έχει τις ιδιότητες σύγκλισης των μεθόδων χορδής [30, 96, 125].

Γενικά, ο πίνακας B_k μπορεί να έχει και μη θετικά στοιχεία. Αυτό έχει σαν αποτελέσματα ένα μη θετικά ορισμένο πίνακα, που στην πράξη σημαίνει ότι η μέθοδος μπορεί να δώσει αρνητικά ή μηδενικά βήματα στις αντίστοιχες κατευθύνσεις. Για να απαλειφθεί αυτό το πρόβλημα, έχει εισαχθεί μια ευρετική παράμετρος αποκαλούμενη «μέγιστος παράγοντας αύξησης» (maximum growth factor) [35] και ελέγχει τις αυξομειώσεις των βημάτων της μεθόδου.

4.4 Αλγόριθμοι Ευρείας Σύγκλισης με Προσαρμοστικό Ρυθμό Εκπαίδευσης

Ένας αλγόριθμος εκπαίδευσης μπορεί να αποκτήσει την ιδιότητα της ευρείας σύγκλισης με κατάλληλη προσαρμογή του ρυθμού εκπαίδευσης, κατά τέτοιο τρόπο ώστε το σφάλμα του ΤΝΔ να ελαχιστοποιείται ακριβώς κατά μήκος της παρούσας κατεύθυνσης ανίχνευσης σε κάθε επανάληψη, δηλαδή $E(w^{k+1}) < E(w^k)$. Για το σκοπό αυτό, απαιτείται μια επαναληπτική αναζήτηση, η οποία είναι συχνά ακριβή από την άποψη των υπολογισμών της συνάρτησης σφάλματος. Πρέπει να σημειωθεί ότι η ανωτέρω σχετικά απλή συνθήκη δεν εγγυάται την ευρεία σύγκλιση για γενικές συναρτήσεις, δηλαδή την σύγκλιση σε ένα τοπικό σημείο ελαχίστου από οποιοδήποτε αρχικό σημείο (για μια γενική συζήτηση σχετικά με τις ευρέως συγκλίνουσες μεθόδους βλ. [30]).

Η χρήση αλγορίθμων για την προσαρμογή του ρυθμού εκπαίδευσης που επιβάλλουν μονότονη μείωση του σφάλματος χρησιμοποιώντας ακατάλληλες τιμές για κρίσιμες ευρετικές παραμέτρους εκπαίδευσης μπορούν να επιβραδύνουν σημαντικά το ρυθμό σύγκλισης, ή ακόμα και να οδηγήσουν τη μέθοδο σε απόκλιση και πρόωρο κορεσμό (premature saturation) [68, 130]. Επιπροσθέτως, στην περίπτωση χρήσης ευρετικών παραμέτρων εκπαίδευσης δεν είναι δυνατή η δημιουργία αλγορίθμων ευρείας σύγκλισης. Μια καλύτερη αντιμετώπιση του προβλήματος είναι η προσαρμογή του ρυθμού εκπαίδευσης έτσι ώστε η τιμή της συνάρτησης σφάλματος να μειώνεται αρκετά σε κάθε επανάληψη και η μείωση αυτή να συνοδεύεται από αλλαγή της τιμής του τρέχοντος σημείου w .

Ακολουθώντας τον παραπάνω συλλογισμό, για το επαναληπτικό σχήμα:

$$w^{k+1} = w^k + \eta^k \varphi^k, \quad (4.7)$$

όπου φ^k είναι η κατεύθυνση ανίχνευσης, όπως είδαμε και στο Κεφάλαιο 3, μπορούν να χρησιμοποιηθούν οι συνθήκες του Wolfe:

$$E(w^{k+1}) - E(w^k) \leq \sigma_1 \eta^k \langle \nabla E(w^k), \varphi^k \rangle, \quad (4.8)$$

$$\langle \nabla E(w^{k+1}), \varphi^k \rangle \geq \sigma_2 \langle \nabla E(w^k), \varphi^k \rangle, \quad (4.9)$$

όπου $0 < \sigma_1 < \sigma_2 < 1$ και $\langle \cdot, \cdot \rangle$ συμβολίζει το εσωτερικό γινόμενο στο $\mathbb{R}^n$.

Η πρώτη ανισότητα εξασφαλίζει ότι η τιμή της συνάρτησης σφάλματος θα μειώνεται αρκετά σε κάθε επανάληψη, ενώ η δεύτερη ανισότητα εμποδίζει τον ρυθμό εκπαίδευσης να γίνει πολύ μικρός, με αποτέλεσμα την πρόωρη σύγκλιση της μεθόδου. Μπορεί ναδειχθεί ότι αν φ^k είναι μια κατεύθυνση μείωσης και η συνάρτηση σφάλματος E είναι συνεχώς παραγωγίσιμη και κάτω φραγμένη κατά μήκος της ακτίνας $\{w^k + \eta\varphi^k \mid \eta > 0\}$, τότε πάντα υπάρχει κάποιος ρυθμός εκπαίδευσης που να ικανοποιεί τις Σχέσεις (4.8)–(4.9) [95, 174, 175].

Το ακόλουθο Θεώρημα, που δόθηκε από τον Wolfe [30, 95, 174, 175], αποδεικνύει ότι εάν η συνάρτηση σφάλματος E είναι κάτω φραγμένη, τότε η ακολουθία σημείων $\{w^k\}_{k=0}^\infty$ που δημιουργείται από οποιονδήποτε αλγόριθμο που ακολουθεί μια κατεύθυνση μείωσης φ^k της οποίας η γωνία θ_k με $-\nabla E(w^k)$ είναι:

$$\cos \theta_k = \frac{\langle -\nabla E(w^k), \varphi^k \rangle}{\|\nabla E(w^k)\| \|\varphi^k\|} \geq \delta > 0, \quad (4.10)$$

και επίσης ικανοποιεί τις συνθήκες του Wolfe, τότε ισχύει ότι είτε $\nabla E(w^k) = 0$ για κάποιο k , ή $\lim_{k \rightarrow \infty} \nabla E(w^k) = 0$ [29].

Θεώρημα 4.1 [30, 95, 174, 175] Έστω ότι η συνάρτηση σφάλματος $E : \mathbb{R}^n \rightarrow \mathbb{R}$ είναι συνεχώς παραγωγίσιμη στο $\mathbb{R}^n$ και ότι ∇E είναι Lipschitz συνεχής στο $\mathbb{R}^n$. Τότε, για οποιοδήποτε $w^0 \in \mathbb{R}^n$, είτε η E δεν είναι κάτω φραγμένη, ή υπάρχει ακολουθία σημείων $\{w^k\}_{k=0}^\infty$ που ικανοποιούν τις συνθήκες του Wolfe (4.8)–(4.9) και είτε:

- (i) $\langle \nabla E(w^k), (w^{k+1} - w^k) \rangle < 0$, ή
- (ii) $\nabla E(w^k) = 0$, και $w^{k+1} - w^k = 0$,

για κάθε $k > 0$. Επιπροσθέτως, για κάθε τέτοια ακολουθία, είτε:

- (a) $\nabla E(w) \neq 0$ για κάποιο $k \geq 0$, ή
- (β) $\lim_{k \rightarrow \infty} E(w^k) = -\infty$, ή
- (γ) $\lim_{k \rightarrow \infty} \langle \nabla E(w^k), (w^{k+1} - w^k) \rangle / \|w^{k+1} - w^k\| = 0$.

Για ένα ανάλογο θεωρητικό αποτέλεσμα σύγκλισης βρίσκεται στο [30, σελ. 123], όπου η ακολουθία $\{w^k\}_{k=0}^\infty$ συγκλίνει q -υπεργραμμικά σε ένα ελάχιστο w^* .

Στην πράξη, η συνθήκη (4.9) δεν είναι απαραίτητη και μπορεί να αντικατασταθεί από μια στρατηγική οπισθοδρόμησης (backtracking) που θα έχει σαν σκοπό την αποφυγή των πολύ μικρών βημάτων [84]. Μια απλή τέτοια στρατηγική για την ρύθμιση του βήματος της ελαχιστοποίησης, έτσι ώστε να ικανοποιεί τις συνθήκες (4.8)–(4.9) σε κάθε επανάληψη, είναι να μειώνεται ο ρυθμός εκπαίδευσης κατά ένα παράγοντα μείωσης $1/q$, όπου $q > 1$ [84, 96]. Επίσης μπορεί να αποδειχθεί ότι εάν η Σχέση (4.9) αντικατασταθεί από την ακόλουθη σχέση:

$$E(w^k + \eta^k \varphi^k) - E(w^k) \geq \sigma_2 \eta^k \langle \nabla E(w^k), \varphi^k \rangle, \quad (4.11)$$

όπου $\sigma_2 \in (\sigma_1, 1)$, τότε το Θεώρημα 4.1 εξακολουθεί να ισχύει [30].

4.5 Η Τροποποιημένη Μέθοδος Quickprop

Για να αποφύγουμε την επίπονη ρύθμιση των ευρετικών παραμέτρων εκπαίδευσης της μεθόδου Qprop και για να εγγυηθούμε την επιθυμητή ιδιότητα ο πίνακας B_k να είναι πάντα

Θετικά ορισμένος, προτείνουμε την ακόλουθη τροποποίηση της μεθόδου Qprop:

$$w_i^{k+1} = w_i^k - \eta \left\{ \frac{|\partial_i E(w^k) - \partial_i E(w^{k-1})|}{|w_i^k - w_i^{k-1}|} \right\}^{-1} \partial_i E(w^k),$$

όπου ο συντελεστής η μπορεί να ρυθμιστεί κατάλληλα. Με αυτό τον τρόπο, το μήκος του βήματος της ελαχιστοποίησης τροποποιείται έτσι ώστε να ικανοποιεί τις συνθήκες του Wolfe, ενώ παράλληλα τα βάρη ανανεώνονται κατά μήκος μια κατεύθυνσης μείωσης.

Ακολουθεί μια περιγραφή της τροποποιημένης μεθόδου Quickprop (MQprop), όπου MIT είναι ο μέγιστος επιτρεπτός αριθμός επαναλήψεων και ε είναι η επιθυμητή ακρίβεια.

Αλγόριθμος 4.1 Η Τροποποιημένη Μέθοδος Quickprop - MQprop

1. Είσοδος: $\{E; w^0; \eta^0; (\lambda_1^0, \lambda_2^0, \dots, \lambda_n^0); MIT; \varepsilon\}$.
2. Θέσε $k = -1$.
3. Εάν $k < MIT$, αντικατέστησε το k με $k + 1$, θέσε $\eta = \eta^0$, και πήγαινε στο επόμενο Βήμα· αλλιώς, πήγαινε στο Βήμα 8.
4. Εάν $k \geq 1$ και $\Lambda_i^k = |\partial_i E(w^k) - \partial_i E(w^{k-1})|/|w_i^k - w_i^{k-1}| \neq 0$, για κάθε $i = 1, \dots, n$, θέσε $\lambda_i^k = 1/\Lambda_i^k$, αλλιώς θέσε $\lambda_i^k = \lambda_i^0$.
5. Ρύθμισε κατάλληλα το η .
6. Θέσε $w^{k+1} = w^k - \eta \text{diag}\{\lambda_1^k, \lambda_2^k, \dots, \lambda_n^k\} \nabla E(w^k)$.
7. Εάν $\|\nabla E(w^k)\| \leq \varepsilon$ πήγαινε στο Βήμα 8· αλλιώς πήγαινε στο Βήμα 3.
8. Έξοδος: $\{w^k; E(w^k); \nabla E(w^k)\}$.

Ας υποθέσουμε τώρα ότι η ρύθμιση του η στο Βήμα 5 του Αλγόριθμου 4.1 γίνεται με βάση τις Σχέσεις (4.8) και (4.9). Το επόμενο Θεώρημα αποδεικνύει ότι εάν η συνάρτηση σφάλματος E είναι κάτω φραγμένη, τότε η ακολουθία σημείων $\{w^k\}_{k=0}^\infty$ που παράγεται από τον Αλγόριθμο 4.1 συγκλίνει σε ένα σημείο w^* για το οποίο ισχύει $\nabla E(w^*) = 0$.

Θεώρημα 4.2 [166] Έστω ότι η συνάρτηση σφάλματος $E : \mathbb{R}^n \rightarrow \mathbb{R}$ είναι συνεχώς παραγωγίσμη και κάτω φραγμένη στο $\mathbb{R}^n$ και έστω ότι ∇E είναι Lipschitz συνεχής στο $\mathbb{R}^n$. Τότε, για οποιοδήποτε αρχικό σημείο $w^0 \in \mathbb{R}^n$, για την ακολουθία σημείων $\{w^k\}_{k=0}^\infty$, που παράγεται από τον Αλγόριθμο 4.1, ικανοποιώντας τις συνθήκες (4.8)–(4.9), ισχύει ότι $\lim_{k \rightarrow \infty} \nabla E(w^k) = 0$.

Απόδειξη. Η ακολουθία $\{w^k\}_{k=0}^\infty$ ακολουθεί την κατεύθυνση

$$\varphi^k(w^k) = -\text{diag}\{1/\Lambda_1^k, \dots, 1/\Lambda_n^k\} \nabla E(w^k),$$

η οποία είναι κατεύθυνση μείωσης αφού

$$\langle \nabla E(w^k), \varphi^k(w^k) \rangle < 0.$$

Επιπρόσθετα, η προϋπόθεση για την γωνία θ_k ικανοποιείται αφού είναι προφανές ότι χρησιμοποιώντας την Σχέση (4.10) παίρνουμε $\cos \theta_k > 0$. Συνεπώς, από το Θεώρημα 4.1 ισχύει ότι $\lim_{k \rightarrow \infty} \nabla E(w^k) = 0$. Το Θεώρημα αποδείχθηκε. $\square$

Παρατήρηση 4.1 Να σημειωθεί ότι για ΤΝΔ που χρησιμοποιούν στιγμοειδείς συναρτήσεις ενεργοποίησης η υπόθεση για συνεχώς παραγωγίσμη συνάρτηση σφάλματος είναι περιττή. Επίσης, στην περίπτωση αυτή η συνάρτηση σφάλματος είναι πάντα κάτω φραγμένη (σαν άθροισμα τετραγώνων) στο $\mathbb{R}^n$.

4.6 Πειραματικά Αποτελέσματα

Στην ενότητα αυτή παρουσιάζουμε αποτελέσματα από τρία προβλήματα εκπαίδευσης ΤΝΔ. Έχουμε δοκιμάσει και συγκρίνει τις ακόλουθες μεθόδους:

- (i) τη μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος με σταθερό ρυθμό εκπαίδευσης (BP) [133].
- (ii) τη μέθοδο της πιο απότομης καθόδου με γραμμική ανίχνευση για τον ρυθμό εκπαίδευσης (Steepest Descent with Line Search), που προτάθηκε από τον Polak [125] (SDLS).
- (iii) τη μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος με σταθερό ρυθμό εκπαίδευσης και ορμή (momentum) (BPM) [57, 133].
- (iv) τη μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος με προσαρμοστικό ρυθμό εκπαίδευσης και ορμή (ABP) [159].
- (v) τη μέθοδο των Fletcher-Reeves (FR) [40].
- (vi) τη μέθοδο των Polak-Ribiere (PR) [40].
- (vii) τη μέθοδο των Polak-Ribiere δεσμευμένη από τη μέθοδο FR (PR-FR) [40].
- (viii) την μέθοδο Quickprop (Qprop) [35].
- (ix) την τροποποιημένη μέθοδο Quickprop (MQprop).

Για την υλοποίηση των μεθόδων FR, PR, και PR-FR, χρησιμοποιούμε την γραμμική ανίχνευση του Armijo όπως έχει τροποποιηθεί από τον Polak [125]. Στο Βήμα 5 της μεθόδου MQprop (βλ. Αλγόριθμο 4.1), χρησιμοποιούμε μια απλή στρατηγική οπισθοδρόμησης για την ρύθμιση του η : το η μειώνεται κατά ένα παράγοντα $1/q$, όπου $q = 2$ [84].

Για όλα τα προβλήματα, στα αποτελέσματα που παραθέτουμε στους ακόλουθους πίνακες παρουσιάζουμε τον μέσο αριθμό των επαναλήψεων (μ_{IT}) που απαιτήθηκαν από την κάθε μέθοδο για την σύγκλιση σε κάποιο τοπικό ελάχιστο, τον μέσο αριθμό των υπολογισμών της συνάρτησης σφάλματος (μ_{FE}) και τον αριθμό των επιτυχημένων πειραμάτων σε ένα σύνολο 1000 ανεξάρτητων δοκιμών (Επιτυχία).

Είναι γνωστό ότι η επιλογή του αρχικού διανύσματος των βαρών είναι πολύ σημαντική στην εκπαίδευση ΤΝΔ. Πολύ μικρά αρχικά βάρη έχουν σαν αποτέλεσμα πολύ μικρά βήματα, και τελικά η ελαχιστοποίηση προς κάποιες κατευθύνσεις είναι πρακτικά αδύνατη. Έτσι για την εκπαίδευση απαιτούνται πολύ περισσότερες επαναλήψεις [133]. Στην χειρότερη περίπτωση, η μέθοδος μπορεί να παγιδευτεί σε κάποιο ανεπιθύμητο τοπικό ελάχιστο με αποτέλεσμα η εκπαίδευση να σταματήσει. Από την άλλη μεριά, μεγάλες αρχικές τιμές επιταχύνουν αρχικά την διαδικασία εκπαίδευσης, αλλά πολλές φορές οδηγούν κάποιους νευρώνες σε κορεσμό και παράγουν νέα σημεία με πολύ μικρές (σχεδόν μηδενικές) τιμές της κλίσης τους. Σε αυτές τις περιπτώσεις η εκπαίδευση είναι εξαιρετικά αργή [81]. Έτσι, είναι συνηθισμένη τακτική στην εκπαίδευση ΤΝΔ, τα αρχικά βάρη να επιλέγονται από την ομοιόμορφη κατανομή στο διάστημα $(w_{\min}, w_{\max})$, όπου συνήθως είναι $w_{\min} = -w_{\max}$. Για να συγκρίνουμε τις διάφορες μεθόδους επιλέχθηκαν 1000 αρχικές τιμές για τα βάρη, με συνιστώσες από το παραπάνω διάστημα $(-1, +1)$.

4.6.1 Αποκλειστικό-ΕΙΤΕ

Το κλασικό πρόβλημα του αποκλειστικό-ΕΙΤΕ (βλ. Παράρτημα Α.1) γνωστό για τα πολλά τοπικά ελάχιστα που παρουσιάζει [57, 133], θα είναι το πρώτο πρόβλημα που θα δοκιμάσουμε. Η συνθήκη τερματισμού για όλους τους αλγόριθμους που δοκιμάσαμε είναι να βρεθεί ένα τοπικό ελάχιστο που να έχει συναρτησιακή τιμή $E \leq 0.04$. Τα αποτελέσματά συνοψίζονται στον Πίνακα 4.1.

Στο παράδειγμα αυτό ο αριθμός των επιτυχιών εξαρτάται από τα πολλά ανεπιθύμητα τοπικά ελάχιστα. Έτσι οι μέθοδοι FR, PR και PR-FR συνέκλιναν συνήθως σε ανεπιθύμητα τοπικά ελάχιστα, δηλαδή τοπικά ελάχιστα με συναρτησιακή τιμή $E > 0.04$. Στην πράξη αυτό συνήθως σημαίνει ότι το εκπαιδευμένο ΤΝΔ δεν καταφέρνει να ταξινομήσει σωστά κάποια από τα πρότυπα εισόδου. Η μέθοδος που προτείνουμε (MQprop) έχει καλύτερο ποσοστό επιτυχίας από τις μεθόδους FR, PR και PR-FR. Επίσης, είναι γρηγορότερη από τις μεθόδους BP, SDLS, BPM και FR, αφού απαιτεί λιγότερους υπολογισμούς της τιμής της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της για να καταφέρει να εκπαιδεύσει το ΤΝΔ. Οι μέθοδοι PR και PR-FR απαιτούν λιγότερους συναρτησιακούς υπολογισμούς, αλλά έχουν μικρότερο ποσοστό επιτυχίας από την MQprop. Τέλος, πρέπει να σημειώσουμε ότι η κλασική μέθοδος Qprop χωρίς την κατάλληλη και χρονοβόρο ρύθμιση ευρετικών παραμέτρων εκπαίδευσης δεν καταφέρνει να εκπαιδεύσει το ΤΝΔ.

Πίνακας 4.1: Αποτελέσματα από το πρόβλημα του αποκλειστικό-ΕΙΤΕ

Αλγόριθμος	μ_{IT}	μ_{FE}	Επιτυχία
BP	549	1098	810/1000
SDLS	64	435	810/1000
BPM	803	1606	810/1000
ABP	157	314	810/1000
FR	84	282	130/1000
PR	21	169	380/1000
PR-FR	22	171	410/1000
MQprop	52	234	810/1000

4.6.2 Ταξινόμηση υφής

Το πρόβλημα ταξινόμησης υφής [82] είναι το δεύτερο πρόβλημα που παρουσιάζουμε. Ένα 16-8-12 ΤΝΔ (με 244 μεταβλητές) εκπαιδεύεται για να ταξινομή 12 εικόνες διαφορετικής υφής (βλ. Παράρτημα Α.8).

Λεπτομερή αποτελέσματα σχετικά με την απόδοση των αλγορίθμων παρουσιάζουμε στον Πίνακα 4.2. Η συνθήκη τερματισμού για όλους τους αλγόριθμους είναι να επιτύχουν σφάλμα ταξινόμησης $CE < 3\%$, δηλαδή το εκπαιδευμένο ΤΝΔ να ταξινομή επιτυχώς τουλάχιστον 117 από τα 120 πρότυπα. Και σε αυτό το παράδειγμα η κλασική μέθοδος Qprop χωρίς ευρετικές παραμέτρους έχει πολύ κακή απόδοση, καθώς συνήθως αποκλίνει, και έτσι δεν συμπεριλαμβάνεται στον Πίνακα 4.2.

Πίνακας 4.2: Αποτελέσματα από το πρόβλημα ταξινόμησης υφής

Αλγόριθμος	μ_{IT}	μ_{FE}	Επιτυχία
BP	15839	31678	960/1000
SDLS	13256	26517	965/1000
BPM	12422	24844	940/1000
ABP	560	1120	1000/1000
FR	1624	12674	250/1000
PR	140	810	990/1000
PR-FR	145	1005	996/1000
MQprop	406	1228	1000/1000

Εκτός από την ταχύτητα και το ποσοστό επιτυχίας σε αυτό το πρόβλημα ελέγχουμε και την ικανότητα γενίκευσης των μεθόδων. Τα επιτυχώς εκπαιδευμένα ΤΝΔ καλούνται να ταξινομήσουν κάποια άγνωστα πρότυπα ελέγχου. Για τον υπολογισμό της γενίκευσης των ΤΝΔ χρησιμοποιούμε τον κανόνα του μεγίστου (max rule), δηλαδή ένα πρότυπο ελέγχου θεωρείται σωστά ταξινομημένο, αν η τιμή του αντίστοιχου νευρώνα εξόδου είναι η μεγαλύτερη από όλους τους άλλους νευρώνες εξόδου. Η μέση γενίκευση για κάθε αλγόριθμο είναι: BP=90.0%, SDLS=90.0%, BPM=90.0%, ABP=93.5%, FR=92.0%, PR=92.6%, PR-FR=93.5%, και MQprop=94.0%.

Συμπερασματικά μπορούμε να πούμε ότι η μέθοδος PR απαιτεί, σε σύγκριση με τις υπόλοιπες μεθόδους, τους λιγότερους υπολογισμούς της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της για την εκπαίδευση του ΤΝΔ. Από την άλλη μεριά οι μέθοδοι ABP και MQprop παρουσιάζουν σθεναρότερη συμπεριφορά, καθώς έχουν μεγαλύτερο ποσοστό επιτυχίας και επίσης πολύ καλή γενίκευση.

4.6.3 Αναγνώριση αριθμών

Σε αυτό το πείραμα [82, 148] ένα 64–6–10 ΤΝΔ (444 βάρη και 16 πολώσεις) εκπαιδεύεται να αναγνωρίζει τους αριθμούς από 0 έως 9 (βλ. Παράρτημα Α.7). Η συνθήκη τερματισμού για όλες τις μεθόδους είναι να βρεθεί ένα τοπικό ελάχιστο με συναρτησιακή τιμή $E \leq 0.001$. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 4.3. Είναι φανερό ότι η μέθοδος MQprop είναι ταχύτερη από όλες τις άλλες μεθόδους που δοκιμάσαμε και έχει ποσοστό επιτυχίας 100%.

Πίνακας 4.3: Αποτελέσματα από το πρόβλημα αναγνώρισης των αριθμών

Αλγόριθμος	μ_{IT}	μ_{FE}	Επιτυχία
BP	14489	28978	660/1000
SDLS	12225	24454	990/1000
BPM	10142	20284	540/1000
ABP	1975	3950	910/1000
FR	620	3121	420/1000
PR	649	2124	960/1000
PR-FR	750	3473	1000/1000
MQprop	159	739	1000/1000

4.7 Συμπεράσματα – Συνεισφορά

Σε αυτό το κεφάλαιο περιγράψαμε την γνωστή μέθοδο εκπαίδευσης ΤΝΔ QuickProp και δώσαμε ένα θεωρητικό αποτέλεσμα σχετικά με την σύγκλισή της. Στη συνέχεια προτείναμε μια τροποποίησή της (MQprop), που δεν απαιτεί την συχνά δύσκολη ρύθμιση των ευρετικών και εξαρτωμένων από το πρόβλημα παραμέτρων εκπαίδευσης της μεθόδου QProp. Αυτές οι ευρετικές παράμετροι, αφού τους δοθούν κατάλληλες τιμές, βοηθούν την κλασική μέθοδο QProp να έχει σταθερότερη σύγκλιση και να αποκλίνει σπανιότερα.

Επίσης, αποδεικνύεται ένα νέο θεώρημα που εγγυάται την σύγκλιση της προτεινόμενης τροποποίησης. Η νέα αυτή μέθοδος συγκρίθηκε με πολλές άλλες μεθόδους εκπαίδευσης ΤΝΔ και το συμπέρασμα είναι ότι παρουσιάζει ταχύτατη, ομαλή και σθεναρή σύγκλιση, με αποτέλεσμα να έχει μεγαλύτερα ποσοστά επιτυχίας. Τέλος, η αύξηση στην ταχύτητα σύγκλισης, δεν έχει αρνητικές συνέπειες στην ικανότητα γενίκευσης των εκπαιδευμένων ΤΝΔ, αφού αυτά είχαν πολύ υψηλή γενίκευση σε όλα τα προβλήματα που δοκιμάσαμε.

Μέρος ΙΙΙ

Μέθοδοι Ολικής Βελτιστοποίησης για την Εκπαίδευση ΤΝΔ

Εκπαίδευση Τεχνητών Νευρωνικών Δικτύων με Ακέραια Βάρη

Δεν είναι το ισχυρότερο είδος αυτό
που επιβιώνει, ούτε το πιο έξυπνο·
είναι το πιο ευπροσάρμοστο στις αλλαγές.

—Charles Robert Darwin (1809-1882)

Σε αυτό το κεφάλαιο θα μελετήσουμε με την εφαρμογή Εξελικτικών Αλγόριθμων (Evolutionary Algorithms – EA) για την εκπαίδευση ΤΝΔ με άκεραια βάρη. Τα ΤΝΔ με άκεραια βάρη είναι πιο εύκολα υλοποιήσιμα σε υλικό (hardware) σε σχέση με τα δίκτυα που χρησιμοποιούν βάρη πραγματικών αριθμών. Στόχος του κεφαλαίου είναι παράλληλα με τους νέους αλγόριθμους εκπαίδευσης που αναπτύχθηκαν να παρουσιάσει και μια πιο γενική εικόνα των ΕΑ και των εφαρμογών τους [117, 119, 120].

5.1 Εισαγωγή

Τα ΤΝΔ μπορούν να εξομοιωθούν με τη χρήση προγραμμάτων υπολογιστών. Όμως για να μπορέσουν να χρησιμοποιηθούν σε πραγματικές εφαρμογές όπου απαιτείται πολύ μικρός χρόνος εκτέλεσης και σε εφαρμογές πραγματικού χρόνου (real time), είναι απαραίτητη η υλοποίηση σε υλικό (hardware). Επίσης, η πλέον φυσική υλοποίηση είναι η παράλληλη υλοποίηση, λόγω της ανεξαρτησίας που παρουσιάζουν οι υπολογισμοί σε κάθε νευρώνα σε σχέση με τους υπολογισμούς που γίνονται στους υπόλοιπους νευρώνες του ίδιου στρώματος.

Το πρόβλημα είναι ότι οι περισσότεροι αλγόριθμοι εκπαίδευσης ΤΝΔ, παράγουν δίκτυα που έχουν βάρη πραγματικών αριθμών. Η υλοποίηση σε ψηφιακό υλικό τέτοιων δικτύων είναι δύσκολη αλλά και ακριβή. Ένα δεύτερο πρόβλημα είναι η αποθήκευση των πραγματικών βαρών, αφού για να μπορεί το ΤΝΔ να λειτουργήσει σωστά απαιτείται αποθήκευση πολλών δεκαδικών ψηφίων τους.

Αντιθέτως, τα ΤΝΔ που έχουν εκπαιδευτεί με άκεραια βάρη είναι πιο εύκολο να υλοποιηθούν τόσο με ηλεκτρονικά όσο και με οπτικά κυκλώματα και η αποθήκευση των άκεραιων βαρών είναι πολύ πιο απλή.

5.2 Εκπαίδευση με Διαφοροεξελικτικούς Αλγόριθμους

Σε μία πρόσφατη εργασία τους οι Storn και Price [153] παρουσίασαν μία νέα κλάση μεθόδων ελαχιστοποίησης και ονόμασαν αυτούς τους αλγόριθμους Διαφοροεξελικτικούς (Differential Evolution – DE). Οι αλγόριθμοι αυτοί ανήκουν στην γενικότερη κατηγορία των ΕΑ.

Οι ΔιαφοροΕξελικτικοί Αλγόριθμοι (ΔΕΑ) σχεδιάστηκαν για να μπορούν να αντιμετωπίσουν δύσκολα προβλήματα ελαχιστοποίησης, όπου η αντικειμενική συνάρτηση είναι μη γραμμική, πιθανά μη παραγωγίσιμη και έχει πολλά τοπικά, αλλά και ολικά ελάχιστα. Για να επιτευχθεί αυτός ο στόχος, οι ΔΕΑ είναι μια κλάση στοχαστικών παράλληλων μεθόδων που δανείζεται ιδέες από την ευρύτερη κλάση των Εξελικτικών Αλγορίθμων, και απαιτούν από τον χρήστη να ρυθμίσει λίγες και σχετικά εύκολες στην επιλογή παραμέτρους. Τα αποτελέσματα της εργασίας [153] δείχνουν ότι οι ΔΕΑ έχουν καλές ιδιότητες σύγκλισης (ταχύτητα, ποσοστό επιτυχίας, εύρεση του ολικού ελαχίστου) και ότι ξεπερνούν σε επίδοση άλλους ΕΑ.

Για να μπορούν να χρησιμοποιηθούν οι ΔΕΑ στην εκπαίδευση ΤΝΔ με ακέραια βάρη, ξεκινάμε τον αλγόριθμο με ένα συγκεκριμένο αριθμό (NP) από N -διάστατα ακέραια διανύσματα βαρών, σαν αρχικό *πληθυσμό*, και τα εξελίσσουμε καθώς προχωρά ο χρόνος (επανάληψεις) μέσω των τελεστών των ΔΕΑ.

Το πλήθος των αρχικών N -διάστατων ακέραιων διανυσμάτων παραμένει σταθερό καθ' όλη την διάρκεια της εκπαίδευσης. Ο πληθυσμός των βαρών αρχικοποιείται με τυχαίους ακέραιους από το διάστημα $[-\Delta, \Delta]$, ακολουθώντας την ομοιόμορφη κατανομή. Σε κάθε επανάληψη του αλγορίθμου, που ονομάζεται και *γενιά* (generation), δημιουργούνται νέα διανύσματα βαρών (reproduction) με τον κατάλληλο συνδυασμό τυχαία επιλεγμένων διανυσμάτων από τον τρέχοντα πληθυσμό και το αποτέλεσμα στρογγυλοποιείται στον πλησιέστερο ακέραιο. Αυτή η διαδικασία ονομάζεται *τελεστής μετάλλαξης* (mutation operator).

Τα νέα αυτά ακέραια διανύσματα «αναμειγνύονται» κατάλληλα με ένα προκαθορισμένο ακέραιο διάνυσμα του πληθυσμού — το *διάνυσμα στόχο* (target vector) — και η διαδικασία αυτή ονομάζεται *τελεστής ανασυνδυασμού*¹ (crossover operator). Το αποτέλεσμα αυτής της διαδικασίας είναι ένα ακέραιο διάνυσμα που ονομάζεται *δοκιμαστικό* (trial vector). Το δοκιμαστικό διάνυσμα γίνεται αποδεκτό για την επόμενη γενιά αν και μόνο αν βελτιώνει (μειώνει) την προηγούμενη τιμή της συνάρτησης σφάλματος E του ΤΝΔ. Η τελευταία διαδικασία ονομάζεται και *τελεστής επιλογής* (selection operator). Συνοπτικά, η διαδικασία εφαρμογής των ΔΕΑ περιγράφεται από το ακόλουθο αλγοριθμικό σχήμα:

```

0 : Αρχικοποίηση του ΤΝΔ;
1 : For  $j = 1, maxgen$ 
2 :   For  $i = 1, NP$ 
3 :     MUTATION( $w_j^i$ )  $\rightarrow$  mutant_vector;
4 :     CROSSOVER(mutant_vector)  $\rightarrow$  trial_vector;
5 :     If  $E(trial\_vector) \leq E(w_j^i)$ ;
6 :       αποδοχή του trial_vector για την επόμενη γενιά;
7 :     EndIf
8 :   EndFor
9 :   If (Συνθήκη Τερματισμού) Terminate;
10: EndFor

```

Στην συνέχεια θα ορίσουμε τους παραπάνω βασικούς τελεστές που χρησιμοποιούνται στην εκπαίδευση ΤΝΔ.

5.2.1 Ο τελεστής μετάλλαξης

Ο πρώτος τελεστής που θα περιγράψουμε είναι ο τελεστής μετάλλαξης. Συγκεκριμένα, για κάθε διάνυσμα βαρών w_g^i , $i = 1, \dots, NP$, όπου g δηλώνει την τρέχουσα γενιά, δημιουργείται ένα νέο διάνυσμα v_{g+1}^i , που ονομάζεται *μεταλλαγμένο διάνυσμα* (mutant vector). Το

¹Ο ανασυνδυασμός στην Γενετική Ανάλυση ονομάζεται και χιασματυπία.

νέο αυτό διάνυσμα υπολογίζεται ακολουθώντας μία από τις παρακάτω σχέσεις:

$$v_{g+1}^i = w_g^{r1} + \mu (w_g^{r1} - w_g^{r2}), \quad (5.1)$$

$$v_{g+1}^i = w_g^{best} + \mu (w_g^{r1} - w_g^{r2}), \quad (5.2)$$

$$v_{g+1}^i = w_g^{r1} + \mu (w_g^{r2} - w_g^{r3}), \quad (5.3)$$

$$v_{g+1}^i = w_g^i + \mu (w_g^{best} - w_g^i) + \mu (w_g^{r1} - w_g^{r2}), \quad (5.4)$$

$$v_{g+1}^i = w_g^{best} + \mu (w_g^{r1} - w_g^{r2}) + \mu (w_g^{r3} - w_g^{r4}), \quad (5.5)$$

$$v_{g+1}^i = w_g^{r1} + \mu (w_g^{r2} - w_g^{r3}) + \mu (w_g^{r4} - w_g^{r5}), \quad (5.6)$$

όπου $\mu > 0$ είναι μία πραγματική παράμετρος, που ονομάζεται *σταθερά μετάλλαξης* και ελέγχει την ένταση της διαφοράς μεταξύ του αρχικού και του μεταλλαγμένου διανύσματος, και $r_1, r_2, r_3, r_4, r_5 \in \{1, 2, \dots, i-1, i+1, \dots, NP\}$, είναι τυχαίοι ακέραιοι, μεταξύ τους διαφορετικοί και διαφορετικοί από τον τρέχοντα δείκτη i . Προφανώς, ο τελεστής μετάλλαξης παράγει ένα πραγματικό διάνυσμα, αφού η σταθερά μετάλλαξης $\mu \in \mathbb{R}$. Όμως ο σκοπός μας είναι να διατηρήσουμε ένα πληθυσμό από ακέραια διανύσματα. Έτσι κάθε συνιστώσα του μεταλλαγμένου διανύσματος στρογγυλοποιείται στον κοντινότερο ακέραιο.

Σχολιάζοντας τις παραπάνω σχέσεις, μπορούμε να πούμε ότι η σχέση (5.1) πρωτοχρησιμοποιήθηκε σαν τελεστής ανασυνδυασμού για Γενετικούς Αλγόριθμους [89] και μοιάζει με τις σχέσεις (5.2) και (5.3). Οι υπόλοιπες σχέσεις είναι παραλλαγές που μπορούν να προκύψουν από απλούς συνδυασμούς των (5.1), (5.2) και (5.3). Είναι φανερό ότι περισσότερες τέτοιες σχέσεις μπορούν να προκύψουν χρησιμοποιώντας σαν δομικά στοιχεία τις παραπάνω.

5.2.2 Ο τελεστής ανασυνδυασμού

Για να αυξήσουμε περαιτέρω την ποικιλία των στρογγυλοποιημένων μεταλλαγμένων ατόμων στον νέο πληθυσμό, εφαρμόζεται ο τελεστής ανασυνδυασμού. Πιο συγκεκριμένα, για κάθε ακέραια συνιστώσα j ($j = 1, 2, \dots, N$) του μεταλλαγμένου διανύσματος v_{g+1}^i , επιλέγουμε τυχαία ένα πραγματικό αριθμό r από το διάστημα $[0, 1]$. Στην συνέχεια, συγκρίνουμε αυτόν τον αριθμό με τη *σταθερά ανασυνδυασμού* ρ (crossover constant), και αν $r \leq \rho$, επιλέγουμε σαν την j συνιστώσα του δοκιμαστικού διανύσματος u_{g+1}^i την αντίστοιχη συνιστώσα j του μεταλλαγμένου διανύσματος· διαφορετικά επιλέγουμε την j συνιστώσα του ακέραιου διανύσματος στόχου w_{g+1}^i . Πρέπει να σημειώσουμε ότι το αποτέλεσμα του τελεστή ανασυνδυασμού είναι επίσης ένα ακέραιο διάνυσμα.

5.2.3 Αποτελέσματα εκπαίδευσης με ακέραια βάρη

Στην παράγραφο αυτή παρουσιάζουμε αποτελέσματα από την μελέτη της απόδοσης των ΔΕΑ αλγορίθμων σε κλασικά προβλήματα εκπαίδευσης ΤΝΔ, όπως το πρόβλημα του Αποκλειστικού-ΕΙΤΕ (βλ. Παράρτημα Α.1), της ισοτιμίας 3-bit (βλ. Παράρτημα Α.2), και το πρόβλημα του 4-2-4 Κωδικοποιητή/Αποκωδικοποιητή (βλ. Παράρτημα Α.3).

Για το υπόλοιπο του κεφαλαίου αυτού, θα ονομάζουμε DE_1 τον αλγόριθμο που χρησιμοποιεί τη σχέση (5.1) ως τελεστή μετάλλαξης, DE_2 τον αλγόριθμο που χρησιμοποιεί τη σχέση (5.2), και ούτω καθεξής. Για όλες τις προσομοιώσεις τα σύνολα εκπαίδευσης και ελέγχου χρησιμοποιούν διπολικά διανύσματα, δηλαδή διανύσματα με τιμές -1 ή 1 . Η συνάρτηση ενεργοποίησης που χρησιμοποιήσαμε, τόσο στο κρυφό όσο και στο στρώμα εξόδου, είναι η υπερβολική εφαιπτομένη.

Οι αναφερόμενες παράμετροι στους πίνακες που ακολουθούν, για τις προσομοιώσεις που έχουν φθάσει στη λύση (έχουν εκπαιδευσει το ΤΝΔ), είναι: Min ο ελάχιστος αριθμός, μ ο μέσος αριθμός, Max ο μέγιστος αριθμός, και σ η τυπική απόκλιση των υπολογισμών της τιμής της συνάρτησης σφάλματος E . Πρέπει να σημειώσουμε ότι όταν ένας αλγόριθμος

αποτυγχάνει να συγκλίνει, εντός του ορίου των συναρτησιακών υπολογισμών (*maxgen*), θεωρείται ότι αποτυγχάνει να εκπαιδεύσει το ΤΝΔ και οι υπολογισμοί της τιμής της συνάρτησης σφάλματος δεν συμπεριλαμβάνονται στους πίνακες. Πρέπει τέλος να τονισθεί ότι ένα κύριο χαρακτηριστικό των ΔΕΑ είναι πως απαιτούνται μόνο οι τιμές της συνάρτησης σφάλματος. Σχετικά με την κλίση της συνάρτησης σφάλματος δεν απαιτείται καμία πληροφορία, έτσι δεν υπάρχει ανάγκη για την οπισθοδρομική διάδοση του σφάλματος (back propagation).

Για τα προβλήματα που εξετάσαμε δεν καταβάλλαμε καμία προσπάθεια εύρεσης τιμών για τις σταθερές μετάλλαξης και ανασυνδυασμού, μ και ρ αντίστοιχα, για την επίτευξη βέλτιστης ή σχεδόν βέλτιστης ταχύτητας σύγκλισης. Αντιθέτως, χρησιμοποιήσαμε προεπιλεγμένες τιμές ($\mu = 0.5$ και $\rho = 0.7$) για όλα τα προβλήματα. Είναι προφανές ότι μπορούν να καθοριστούν με ακρίβεια οι παράμετροι αυτές για κάθε πρόβλημα, με σκοπό να επιτευχθούν τα καλύτερα δυνατά αποτελέσματα, δηλαδή λιγότεροι υπολογισμοί της τιμής της συνάρτησης σφάλματος ή/και υψηλότερο ποσοστό επιτυχίας· κάτι τέτοιο όμως προϋποθέτει πρόσθετο υπολογιστικό κόπο για κάθε πρόβλημα.

Όσον αφορά το μέγεθος του πληθυσμού NP , πειραματικά αποτελέσματα έχουν δείξει [117, 119, 120] ότι μια καλή επιλογή για την τιμή του NP είναι $2N \leq NP \leq 4N$. Είναι προφανές ότι η διερεύνηση του χώρου των βαρών είναι αποτελεσματικότερη για μεγάλες τιμές του NP , αλλά έτσι απαιτούνται περισσότεροι υπολογισμοί της συνάρτησης σφάλματος σε κάθε γενιά. Αφ' ετέρου, μικρές τιμές του NP καθιστούν τους ΔΕΑ ανεπαρκείς και συνήθως απαιτούνται περισσότερες γενιές για να επιτευχθεί η εκπαίδευση του ΤΝΔ. Συνεπώς, προσεκτική επιλογή του μεγέθους του πληθυσμού μπορεί να επιταχύνει την σύγκλιση του ΔΕΑ.

Ο πληθυσμός των διανυσμάτων των βαρών αρχικοποιήθηκε με τυχαίους ακέραιους αριθμούς από το διάστημα $[-\Delta, \Delta]$. Παρά το γεγονός ότι αν $\Delta > 1$ συχνά η σύγκλιση των ΔΕΑ είναι πιο γρήγορη, επιλέξαμε $\Delta = 1$ γιατί η πρόθεσή μας ήταν να εκπαιδεύσουμε με όσο το δυνατόν μικρότερα αρχικά βάρη.

Αποκλειστικό-ΕΙΤΕ

Το πρώτο πρόβλημα που θα αντιμετωπίσουμε είναι το πρόβλημα του Αποκλειστικού-ΕΙΤΕ, το οποίο ιστορικά πέρα των άλλων έχει θεωρηθεί ως καλό πρόβλημα για τον έλεγχο της αξιοπιστίας αλγορίθμων εκπαίδευσης ΤΝΔ. Ένα 2-2-1 ΤΝΔ (6 βάρη και 3 πολώσεις) χρησιμοποιήθηκε στις δοκιμές αυτές. Η εκπαίδευση σταμάτησε όταν βρέθηκε τιμή για τα βάρη, τέτοια ώστε η συνάρτηση σφάλματος E , να είναι $E \leq 0.1$ μέσα σε 100 γενιές (*maxgen* = 100). Ο πληθυσμός αποτελείται από 18 άτομα ($NP = 18$).

Τα αποτελέσματα φαίνονται στον Πίνακα 5.1. Ένα τυπικό διάνυσμα βαρών είναι το ακόλουθο: $w = (2, -3, -2, 2, 3, 3, -2, -2, 2)$, με $E(w) = 0.003$. Οι 6 πρώτες συντεταγμένες αντιστοιχούν στα βάρη του δικτύου, ενώ οι υπόλοιπες 3 στις πολώσεις. Αξίζει να σημειωθεί ότι ο μέσος αριθμός υπολογισμών της τιμής της συνάρτησης σφάλματος καθώς επίσης και ο αριθμός των επιτυχημένων προσομοιώσεων που έχουν οι προτεινόμενες στρατηγικές, είναι καλύτερος από τον αντίστοιχο αριθμό πολύ γνωστών μεθόδων (π.χ. των μεθόδων BP, BPM, και ABP), που χρησιμοποιούν συνεχή βάρη και πληροφορίες σχετικά με την παράγωγο της συνάρτησης σφάλματος.

Ισοτιμία των 3-bit

Το δεύτερο πρόβλημα που θα μελετήσουμε είναι το πρόβλημα της ισοτιμίας 3-bit, το οποίο μπορεί να θεωρηθεί ως ένα γενικευμένο πρόβλημα αποκλειστικού-ΕΙΤΕ, αλλά είναι δυσκολότερο. Ο στόχος είναι να εκπαιδευθεί ένα ΤΝΔ ώστε να παραγάγει το άθροισμα, *mod* 2 τριών δυαδικών εισόδων (ή αλλιώς να υπολογίσει τη συνάρτηση της περιττής ισοτιμίας).

Εδώ χρησιμοποιούμε ένα 3-3-1 ΤΝΔ (12 βάρη και 4 πολώσεις). Ο αρχικός πληθυσμός αποτελείται από 32 διανύσματα βαρών και η εκπαίδευση συνεχίστηκε μέχρι να βρεθεί τιμή για

Πίνακας 5.1: Αποτελέσματα από το πρόβλημα του αποκλειστικού ΕΙΤΕ

Αλγόριθμος	Min	μ	Max	σ	Επιτυχία (%)
DE_1	108	547.9	1332	268.1	91%
DE_2	60	180.5	500	91.0	80%
DE_3	126	551.7	1656	281.5	95%
DE_4	120	271.7	940	133.9	82%
DE_5	90	283.6	1530	256.2	81%
DE_6	126	720.9	1782	387.2	93%
$NP = 18, \mu = 0.5, \rho = 0.7, maxgen = 100$					

τα βάρη, τέτοια ώστε η τιμή της συνάρτησης σφάλματος να είναι $E \leq 0.1$ μέσα σε 100 γενιές ($maxgen = 100$). Ένα χαρακτηριστικό διάνυσμα βαρών μετά από το τέλος της εκπαίδευσης του ΤΝΔ είναι $w = (1, -3, 1, -2, 2, 2, -3, -4, 3, 3, 2, -3, -1, 0, -1, 0)$ και η αντίστοιχη τιμή της συνάρτησης σφάλματος είναι $E(w) = 0.009$. Στον Πίνακα 5.2 παρουσιάζονται τα αποτελέσματα για αυτό το πρόβλημα.

Πίνακας 5.2: Αποτελέσματα από το πρόβλημα της ισοτιμίας 3-bit

Αλγόριθμος	Min	μ	Max	σ	Επιτυχία (%)
DE_1	660	1721.2	2910	640.4	72%
DE_2	150	517.4	1980	298.2	97%
DE_3	900	2004.4	2910	561.6	81%
DE_4	300	768.2	2040	379.5	99%
DE_5	180	732.1	2430	472.9	89%
DE_6	600	2115.6	2970	645.4	50%
$NP = 32, \mu = 0.5, \rho = 0.7, maxgen = 100$					

4-2-4 Κωδικοποιητής/Αποκωδικοποιητής

Το τελευταίο πρόβλημα εκπαίδευσης που θα εξετάσουμε είναι ο 4-2-4 κωδικοποιητής/αποκωδικοποιητής. Τέσσερα διαφορετικά πρότυπα εισόδου παρουσιάζονται στο δίκτυο. Το κάθε πρότυπο έχει 4-bit· 3 από αυτά είναι 0 και μόνο ένα έχει την τιμή 1. Σκοπός στο πρόβλημα αυτό είναι να αναπαραχθεί στην έξοδο το ίδιο πρότυπο με αυτό που παρουσιάστηκε στην είσοδο. Για να συμβεί αυτό, το ΤΝΔ πρέπει να αναπτύξει μια μοναδική κωδικοποίηση για κάθε έναν από τα 4 πρότυπα και δεδομένου ότι όλες οι πληροφορίες περνούν από το κρυφό στρώμα που έχει μόνο 2 νευρώνες, μπορούμε να πούμε ότι το σύνολο των βαρών εκτελούν την κωδικοποίηση και αποκωδικοποίηση των δεδομένων της εισόδου.

Αυτή η κωδικοποίηση θεωρείται ιδιαίτερα δύσκολη (tight), δεδομένου ότι ο αριθμός των νευρώνων στο κρυφό στρώμα είναι ίσος με το λογάριθμο με βάση 2 των νευρώνων του στρώματος εξόδου ($\log_2 4 = 2$). Αυτό το πρόβλημα έχει επιλεγεί επειδή είναι αρκετά κοντά στα πραγματικά προβλήματα ταξινόμησης, όπου οι μικρές αλλαγές στα στοιχεία του συνόλου εισόδου προκαλούν τις μικρές αλλαγές στην έξοδο [35]. Στο πρόβλημα αυτό ο πληθυσμός αποτελείται από 64 άτομα ($NP = 64$). Ο Πίνακας 5.3 συνοψίζει τα αποτελέσματα.

Πίνακας 5.3: Αποτελέσματα από το πρόβλημα του 4-2-4 Κωδικοποιητή/Αποκωδικοποιητή

Αλγόριθμος	Min	μ	Max	σ	Επιτυχία (%)
DE_1	2640	2838.0	3036	280.1	2%
DE_2	448	912.7	4096	569.4	84%
DE_3	1984	4501.3	6272	1098.6	60%
DE_4	512	1026.6	3776	438.6	100%
DE_5	246	1192.5	4092	757.2	78%
DE_6	3136	4640.0	6272	956.7	18%
$NP = 64, \mu = 0.5, \rho = 0.7, maxgen = 100$					

5.3 Εκπαίδευση με Περιορισμένα Ακέραια Βάρη

Στην Ενότητα 5.2 εξετάσαμε στρατηγικές εκπαίδευσης ΤΝΔ με ακέραια βάρη. Εδώ θα προχωρήσουμε ακόμα περισσότερο περιορίζοντας τα βάρη αυτά στο διάστημα $[-2^{k-1} + 1, 2^{k-1} - 1]$, με $k = 3, 4, 5$, το οποίο αντιστοιχεί σε k -bit αναπαράσταση των ακέραιων βαρών [119, 120]. Η ιδιότητα αυτή ελαττώνει το ποσό της απαιτούμενης μνήμης για την αποθήκευση των βαρών σε ψηφιακές ηλεκτρονικές υλοποιήσεις. Επιπροσθέτως απλοποιεί την διαδικασία του πολλαπλασιασμού, αφού για να πολλαπλασιάσουμε οποιονδήποτε αριθμό με έναν ακέραιο k -bit, απαιτούνται μόνο: μια αλλαγή προσήμου, $(k-1)(k-2)/2$ αριστερές ολισθήσεις ενός βήματος (one-step left shift) και $(k-2)$ προσθέσεις. Τέλος, αν τα δεδομένα εισόδου είναι περιορισμένα στο σύνολο $\{-1, 1\}$ (διπολικά διανύσματα), τότε οι νευρώνες στο στρώμα εισόδου για τον πολλαπλασιασμό αρκεί να κάνουν μόνο μια αλλαγή προσήμου και ακέραιες προσθέσεις.

Για να μπορέσουμε να εφαρμόσουμε τους ΔΕΑ στην εκπαίδευση ΤΝΔ με k -bit ακέραια βάρη, θα πρέπει να αρχικοποιήσουμε τον πληθυσμό των διανυσμάτων των βαρών με τυχαίους ακέραιους από την ομοιόμορφη κατανομή στο διάστημα $[-2^{k-1} + 1, 2^{k-1} - 1]$, για $k = 3, 4, 5$.

Επίσης, θα πρέπει να τροποποιήσουμε ελαφρά τον τελεστή μετάλλαξης, ο οποίος έχει σαν αποτέλεσμα πραγματικά διανύσματα βαρών. Αφού ο στόχος μας είναι να διατηρήσουμε ένα πληθυσμό k -bit ακέραιων διανυσμάτων, το μεταλλαγμένο διάνυσμα, αρχικά, στρογγυλοποιείται στον πλησιέστερο ακέραιο. Στην συνέχεια, αν το νέο διάνυσμα δεν ανήκει στο διάστημα $[-2^{k-1} + 1, 2^{k-1} - 1]^N$, τότε υπολογίζουμε το μεταλλαγμένο διάνυσμα από την ακόλουθη σχέση:

$$v_{g+1}^i = \text{sign}(v_{g+1}^i) \cdot \left(|v_{g+1}^i| \bmod 2^{k-1} \right).$$

Τέλος, σημειώνουμε ότι το αποτέλεσμα του τελεστή ανασυνδυασμού είναι και αυτό ένα k -bit ακέραιο διάνυσμα, αφού απλά μεταθέτει συνιστώσες μεταξύ k -bit ακέραιων διανυσμάτων.

5.3.1 Αποτελέσματα εκπαίδευσης με περιορισμένα ακέραια βάρη

Για να ελέγξουμε την απόδοση των ΔΕΑ με περιορισμένα ακέραια βάρη, τους εφαρμόσαμε στα δυο κλασικά προβλήματα εκπαίδευσης που είδαμε και στην προηγούμενη ενότητα, δηλαδή το πρόβλημα του Αποκλειστικού-ΕΙΤΕ (βλ. Παράρτημα Α.1) και το πρόβλημα της ισοτιμίας 3-bit (βλ. Παράρτημα Α.2). Θα παρουσιάσουμε αποτελέσματα για τις διάφορες τιμές του k , ($k = 3, 4, 5$).

Το πλήθος, NP , των διανυσμάτων των βαρών που αποτελούν τον πληθυσμό των ΔΕΑ είναι και για τα δύο προβλήματα $NP = 2N$, όπου N είναι η διάσταση του προβλήματος. Τα διανύσματα των βαρών αρχικοποιήθηκαν με τυχαίους ακέραιους από το διάστημα $[-2^{k-1} + 1, 2^{k-1} - 1]$.

Αποκλειστικό-ΕΙΤΕ

Ένα 2-2-1 ΤΝΔ (6 βάρη και 3 πολώσεις) χρησιμοποιήθηκε στις δοκιμές αυτές. Η εκπαίδευση σταμάτησε όταν βρέθηκε τιμή για τα βάρη, τέτοια ώστε η συνάρτηση σφάλματος να είναι $E \leq 0.1$ μέσα σε 100 γενιές ($maxgen = 100$). Ο πληθυσμός αποτελείται από 18 άτομα ($NP = 18$). Τα αποτελέσματα φαίνονται στον Πίνακα 5.1. Ένα τυπικό ακέραιο 3-bit διάνυσμα βαρών είναι το ακόλουθο: $w = (3, 3, 2, 3, 2, -2, 1, -3, -2)$, με $E(w) = 0.0221$. Οι έξι πρώτες συντεταγμένες αντιστοιχούν στα βάρη του δικτύου, ενώ οι υπόλοιπες τρεις στις πολώσεις. Αξίζει να σημειωθεί ότι και σε αυτό το πρόβλημα ο αριθμός των επιτυχημένων προσομοιώσεων που έχουν οι προτεινόμενες στρατηγικές, είναι μεγαλύτερος από τον αντίστοιχο αριθμό άλλων μεθόδων, που χρησιμοποιούν συνεχή βάρη και πληροφορίες σχετικά με την κλίση της συνάρτησης σφάλματος.

Πίνακας 5.4: Αποτελέσματα εκπαίδευσης με περιορισμένα βάρη (Αποκλειστικό-ΕΙΤΕ)

k	Αλγόριθμος	Min	μ	Max	σ	Επιτυχία (%)
3	DE_1	54	191.9	810	89.7	63.0%
	DE_2	90	836.3	1782	371.7	93.5%
	DE_3	90	300.5	1584	171.2	82.5%
	DE_4	54	364.5	1676	222.6	93.4%
	DE_5	36	1047.8	1782	422.6	63.0%
	DE_6	54	931.5	1782	431.1	73.1%
4	DE_1	90	182.9	612	97.5	69.0%
	DE_2	72	266.7	1188	162.4	87.2%
	DE_3	198	647.8	1566	246.3	99.1%
	DE_4	126	764.1	1656	357.2	95.9%
	DE_5	72	316.9	1602	300.4	92.3%
	DE_6	108	660.0	1566	368.7	93.1%
5	DE_1	54	192.9	684	124.7	75.1%
	DE_2	72	284.9	1332	216.2	80.4%
	DE_3	144	583.9	1314	256.3	97.3%
	DE_4	180	706.1	1764	343.7	97.5%
	DE_5	72	300.5	1584	250.2	85.2%
	DE_6	90	482.9	1368	264.9	93.0%
$NP = 18, \mu = 0.5, \rho = 0.7, maxgen = 100$						

Ισοτιμία των 3-bit

Για το πρόβλημα της ισοτιμίας 3-bit χρησιμοποιήθηκε ένα 3-3-1 ΤΝΔ (12 βάρη και 4 πολώσεις). Ο πληθυσμός αποτελείται από 32 διανύσματα βαρών και τα αποτελέσματα φαίνονται στον Πίνακα 5.5. Ένα τυπικό ακέραιο διάνυσμα βαρών μετά το τέλος της εκπαίδευσης είναι το ακόλουθο: $w = (3, 3, 2, 3, -1, -1, 2, -2, -2, -3, 3, -3, 1, 0, 1, 1)$ και η αντίστοιχη τιμή της συνάρτησης σφάλματος είναι $E(w) = 0.0257$.

Στους παραπάνω πίνακες αποτελεσμάτων έχουμε δείξει ότι οι στρατηγικές DE_3 και DE_4 είναι οι πιο αποδοτικές για την εκπαίδευση ΤΝΔ με ακέραια βάρη. Το γεγονός αυτό επιβεβαιώνεται και από τον Πίνακα 5.5, όπου οι στρατηγικές DE_1, DE_3 και DE_4 έχουν την καλύτερη απόδοση για όλες τις τιμές του k που δοκιμάσαμε.

Πίνακας 5.5: Αποτελέσματα εκπαίδευσης με περιορισμένα βάρη (Ισοτιμία 3-bit)

k	Αλγόριθμος	Min	μ	Max	σ	Επιτυχία (%)
3	DE_1	96	809.2	2016	313.9	88.1%
	DE_2	704	1966.9	3072	769.2	1.5%
	DE_3	320	1123.4	3168	461.0	97.7%
	DE_4	160	2072.1	3168	631.9	75.5%
	DE_5	1344	2057.1	3072	703.9	0.7%
	DE_6	160	1890.0	3168	907.7	3.2%
4	DE_1	96	762.7	2624	515.1	90.3%
	DE_2	192	2056.0	3072	711.4	28.1%
	DE_3	320	978.3	3136	555.5	95.4%
	DE_4	288	1333.0	3104	652.6	96.5%
	DE_5	192	1959.1	2816	981.6	9.6%
	DE_6	1056	2153.4	3168	619.7	17.8%
5	DE_1	160	622.6	3072	522.1	90.8%
	DE_2	576	1994.1	3168	657.6	60.8%
	DE_3	224	896.3	2688	450.6	99.1%
	DE_4	256	1060.2	3168	716.6	97.5%
	DE_5	672	2112.0	3104	644.9	26.0%
	DE_6	352	2062.5	3168	794.8	44.1%
$NP = 32, \mu = 0.5, \rho = 0.7, maxgen = 100$						

5.4 Εκπαίδευση με Χρήση Συναρτήσεων Ενεργοποίησης με Κατώφλια

Σε αυτή την ενότητα, θα μελετήσουμε την εκπαίδευση ΤΝΔ με χρήση μόνο συναρτήσεων ενεργοποίησης με κατώφλια [121]. Αν και υπάρχουν αλγόριθμοι που είναι ικανοί να εκπαιδεύσουν με κατώφλια [26, 85], συνήθως απαιτούν το πρόβλημα να είναι στατικό, δηλαδή τα πρότυπα εκπαίδευσης να μην αλλάζουν με τον χρόνο. Μια ανασκόπηση των μεθόδων εκπαίδευσης με κατώφλια υπάρχει στην εργασία [105]. Η βασική τους ιδέα είναι να καταφέρουν να εκπαιδεύσουν το ΤΝΔ και να βρουν κάποια πραγματικά βάρη τα οποία μετά να υλοποιήσουν σε υλικό. Όμως πολλές πραγματικές εφαρμογές απαιτούν η εκπαίδευση να μπορεί να συνεχιστεί και στο υλικό όταν νέα πρότυπα εκπαίδευσης γίνουν διαθέσιμα. Ένα από τα πλεονεκτήματα των ΔΕΑ είναι ότι μπορούν να συνεχίσουν την εκπαίδευση σε υλικό με χρήση αποκλειστικά συναρτήσεων κατωφλιών.

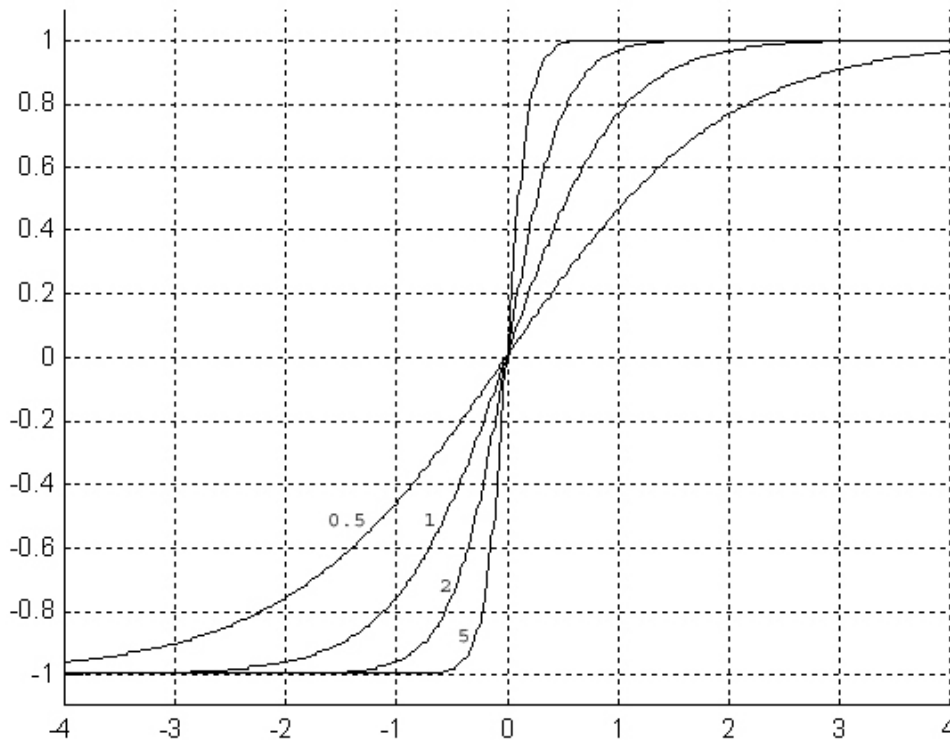
Όπως είδαμε και παραπάνω, οι ΔΕΑ δεν απαιτούν οι συναρτήσεις ενεργοποίησης να είναι παραγωγίσιμες και έτσι είναι κατάλληλοι για εκπαίδευση ΤΝΔ με ακέραια βάρη και με χρήση συναρτήσεων ενεργοποίησης με κατώφλια. Αυτό μπορεί να γίνει σε δύο φάσεις. Στην πρώτη φάση οι ΔΕΑ εκπαιδεύουν το ΤΝΔ, με ακέραια βάρη, χρησιμοποιώντας σιγμοειδείς συναρτήσεις ενεργοποίησης, όπως οι ακόλουθες:

$$f_1(x) = \frac{2}{1 + e^{-\lambda x}} - 1, \quad (5.7)$$

$$f_2(x) = \frac{1}{1 + e^{-\lambda x}}, \quad (5.8)$$

$$f_3(x) = \tanh\left(\frac{\lambda x}{2}\right), \quad (5.9)$$

όπου λ είναι μια παράμετρος που ρυθμίζει την μορφή της σιγμοειδούς συνάρτησης· όσο μεγαλύτερες τιμές παίρνει η παράμετρος λ τόσο πιο απότομη γίνεται η σιγμοειδής και προσεγγίζει τον κατακόρυφο άξονα. Αυτή η φάση επιτυγχάνει μια πρώτη γρήγορη εκπαίδευση. Στην δεύτερη φάση, σταδιακά αυξάνεται η τιμή του λ , με τέτοιο τρόπο ώστε η σιγμοειδής τελικά να προσεγγίζει την συνάρτηση κατώφλι. Συγκεκριμένα, αφού το ΤΝΔ έχει εκπαιδευτεί με σιγμοειδείς συναρτήσεις, η τιμή του λ αυξάνεται σύμφωνα με το ακόλουθο σχήμα: $(1, 10, 20, 30, 40, 50, \infty)$. Είναι πιθανό να απαιτηθούν επιπρόσθετες επαναλήψεις με την αλλαγή του λ από μια τιμή στην επόμενη ή τελικά με τη χρήση της συνάρτησης κατώφλι. Στο Σχήμα 5.1, βλέπουμε την επίδραση που έχει η αύξηση της παραμέτρου λ στην μορφή μιας σιγμοειδούς συνάρτησης ενεργοποίησης.



Σχήμα 5.1: Η επίδραση της παραμέτρου λ στην μορφή μιας σιγμοειδούς συνάρτησης ενεργοποίησης

Η διαδικασία αυτή παράγει μια συνάρτηση ενεργοποίησης που είναι παρόμοια με το όριο της αρχικής, όταν το λ τείνει στο άπειρο. Τελικά, το ΤΝΔ χρησιμοποιεί για ενεργοποιήσεις μόνο συναρτήσεις κατώφλια και έτσι η πολυπλοκότητά του έχει μειωθεί κατά πολύ, κάνοντάς το ιδιαίτερα εύκολο για υλοποίηση σε υλικό. Αξίζει να σημειωθεί ότι αν νέα πρότυπα εκπαίδευσης προκύψουν μετά την εκπαίδευση, η προτεινόμενη μεθοδολογία είναι ικανή να εκπαιδεύσει το ΤΝΔ και αφού έχει ήδη υλοποιηθεί σε υλικό (“on-chip” training), χωρίς τη χρήση σιγμοειδών συναρτήσεων.

5.4.1 Αποτελέσματα εκπαίδευσης με συναρτήσεις ενεργοποίησης με κατώφλια

Στους Πίνακες 5.6 και 5.7 φαίνονται τα αποτελέσματα των αλγορίθμων DE_3 και DE_4 , όταν η εκπαίδευση γίνεται σύμφωνα με την μεθοδολογία που αναπτύξαμε στην Ενότητα 5.4, δηλαδή όταν χρησιμοποιούνται μόνο συναρτήσεις ενεργοποίησης με κατώφλια. Τα προβλήματα που δοκιμάσαμε ήταν το Αποκλειστικό-ΕΓΓΕ και το πρόβλημα της ισοτιμίας 3-bit.

Από τα αποτελέσματα είναι σαφές ότι οι ΔΕΑ είναι ικανοί να εκπαιδεύσουν ΤΝΔ με χρήση

Πίνακας 5.6: Αποτελέσματα εκπαίδευσης με κατώφλια (Αποκλειστικό-ΕΙΤΕ)

k	Αλγόριθμος	Min	μ	Max	σ	Επιτυχία (%)
3	DE_3	270	442.4	756	87.1	100%
	DE_4	270	513.2	1386	204.4	100%
4	DE_3	270	449.6	4104	375.5	100%
	DE_4	252	385.0	918	91.7	100%
5	DE_3	270	432.4	1062	106.6	100%
	DE_4	252	394.0	1170	111.9	100%
$NP = 18, \mu = 0.5, \rho = 0.7, maxgen = 100$						

Πίνακας 5.7: Αποτελέσματα εκπαίδευσης με κατώφλια (Ισοτιμία 3-bit)

k	Αλγόριθμος	Min	μ	Max	σ	Επιτυχία (%)
3	DE_3	1024	2142.4	5376	760.0	100%
	DE_4	1056	3459.8	6240	1137.6	100%
4	DE_3	512	1348.2	3808	577.4	100%
	DE_4	608	1901.1	4960	903.2	100%
5	DE_3	672	1423.8	7488	825.6	100%
	DE_4	640	1732.8	9888	1195.8	100%
$NP = 32, \mu = 0.5, \rho = 0.7, maxgen = 100$						

ενεργοποιήσεων με κατώφλια. Συγκρίνοντας μάλιστα τα αποτελέσματα αυτά με τα υπόλοιπα αποτελέσματα των ΔΕΑ στο κεφάλαιο αυτό, βλέπουμε ότι αν και υπάρχει μια αύξηση στο υπολογιστικό κόστος (περισσότεροι υπολογισμοί των τιμών της συνάρτησης σφάλματος), λόγω της σταδιακής αύξησης της παραμέτρου λ , υπάρχει επίσης μια σημαντική αύξηση του ποσοστού επιτυχίας, που είναι 100% για όλες τις δοκιμές μας.

5.5 Εκπαίδευση με Παράλληλους ΔΕΑ

Σ' αυτή την ενότητα προτείνουμε και εξετάζουμε τους Παράλληλους Διαφοροεξελικτικούς Αλγόριθμους (ΠΔΕΑ) [122]. Οι ΠΔΕΑ είναι μια νέα κλάση αλγορίθμων για την εκπαίδευση ΤΝΔ με ακέραια βάρη. Το τελικό λογισμικό που υλοποιεί τους ΠΔΕΑ εκτελείται σε παράλληλα υπολογιστικά συστήματα που έχουν περισσότερους του ενός επεξεργαστές.

Γενικά, οι ΕΑ μπορούν εύκολα να υλοποιηθούν σε παράλληλο λογισμικό και να εκτελεστούν σε παράλληλα υπολογιστικά συστήματα. Αυτό είναι εφικτό γιατί επεξεργαζόμαστε κάθε άτομο του πληθυσμού ξεχωριστά. Το μοναδικό σημείο που πρέπει να υπάρξει συγχρονισμός και επικοινωνία μεταξύ των ατόμων είναι κατά την εφαρμογή των τελεστών μετάλλαξης και ανασυνδυασμού. Η επικοινωνία αυτή μπορεί να γίνει και παράλληλα αφού δεν απαιτεί συνεργασία του συνόλου του πληθυσμού.

Στη συνέχεια θα περιγράψουμε τα δυο μοντέλα για την παράλληλη επεξεργασία ΕΑ. Σύμφωνα με το πρώτο (fine grain parallelism), κάθε άτομο αντιπροσωπεύεται από ένα επεξεργαστή. Αυτό το μοντέλο προφανώς δημιουργεί την ανάγκη ύπαρξης μεγάλου αριθμού διαθέσιμων επεξεργαστών και υποδομής για την ταχύτατη επικοινωνία μεταξύ τους, με αποτέλεσμα να έχει αυξημένο κόστος κατά την υλοποίησή του.

Το δεύτερο μοντέλο αντιστοιχεί σε κάθε επεξεργαστή ένα τμήμα του αρχικού πληθυσμού

(υποπληθυσμός). Οι υποπληθυσμοί σε κάθε επεξεργαστή εξελίσσονται παράλληλα και ανεξάρτητα, και καθένας εξελίσσεται προς κάποια λύση του προβλήματος. Περιστασιακά γίνεται «μετανάστευση» των καλύτερων ατόμων μεταξύ των υποπληθυσμών με σκοπό την συνεργασία των υποπληθυσμών και των ατόμων για την ταχύτερη διερεύνηση του χώρου των βαρών και σύγκλιση του αλγόριθμου. Στις δοκιμές που θα παρουσιάσουμε στην συνέχεια, έχει χρησιμοποιηθεί το δεύτερο μοντέλο.

Η μετανάστευση των καλύτερων ατόμων ακολουθεί μια προεπιλεγμένη τοπολογία των υποπληθυσμών. Η τοπολογία που χρησιμοποιήθηκε εδώ είναι αυτή του δακτύλιου (ring), δηλαδή το καλύτερο άτομο από κάθε υποπληθυσμό μετακινείται στον επόμενο υποπληθυσμό του δακτύλιου. Αυτή η τοπολογία μειώνει τις μεταναστεύσεις και συνεπώς και τα μηνύματα που ανταλλάσσονται μεταξύ των επεξεργαστών. Άλλες πιθανές τοπολογίες είναι η σύνδεση όλων των υποπληθυσμών με όλους (full mesh), η τυχαία σύνδεση των υποπληθυσμών (random) και η σύνδεση με βάση ένα πλέγμα (grid), δηλαδή κάθε επεξεργαστής επικοινωνεί με άλλους 4 σε ένα διδιάστατο πλέγμα, με άλλους 6 σε ένα τριδιάστατο πλέγμα κτλ.

Η συχνότητα της μετανάστευσης των καλύτερων ατόμων ελέγχεται από την *σταθερά μεταναστεύσης* (migration constant), που ανήκει στο διάστημα $\phi \in (0, 1)$. Σε κάθε γενιά, ένας τυχαίος πραγματικός αριθμός από το διάστημα $(0, 1)$ επιλέγεται τυχαία από την ομοιόμορφη κατανομή και συγκρίνεται με τη σταθερά μετανάστευσης. Εάν η σταθερά μετανάστευσης είναι μεγαλύτερη, τότε το καλύτερο άτομο κάθε υποπληθυσμού μετακινείται σύμφωνα με την τοπολογία σε κάποιο άλλο υποπληθυσμό και παίρνει την θέση ενός τυχαία επιλεγμένου ατόμου (διαφορετικό από το καλύτερο του υποπληθυσμού)· αλλιώς καμία μετανάστευση δεν γίνεται. Πειραματικά αποτελέσματα δείχνουν ότι μια καλή τιμή για τη σταθερά μετανάστευσης είναι $\phi = 0.1$, αφού επιτρέπει στους υποπληθυσμούς να εξελιχθούν αυτόνομα για μερικές γενιές πριν την μετακίνηση των καλύτερων ατόμων τους.

Πρέπει εδώ να επισημάνουμε ότι είναι δυνατή η σειριακή προσομοίωση των ΠΔΕΑ σε υπολογιστικά συστήματα με ένα επεξεργαστή. Στην περίπτωση αυτή δεν έχουμε βέβαια το κέρδος της μείωσης του χρόνου εκτέλεσης, αλλά παρόλα αυτά η εφαρμογή ΠΔΕΑ συστήνεται αφού η ιδέα των υποπληθυσμών (όπως θα φανεί και από τα επόμενα αποτελέσματα) βοηθά σημαντικά στην αύξηση του ποσοστού επιτυχίας των αλγόριθμων.

5.5.1 Η γήρανση του πληθυσμού

Για να αποφύγουμε την παραμονή στον πληθυσμό των ατόμων για πολύ μεγάλο χρονικό διάστημα, χρησιμοποιούμε την ιδέα της *γήρανσης* [142, 152]. Για να το επιτύχουμε αυτό, σε κάθε διάνυσμα του πληθυσμού αναθέτουμε τυχαία ένα ακέραιο αριθμό, ο οποίος προσδιορίζει την μέγιστη ηλικία του ατόμου. Η μέγιστη ηλικία ανήκει στο διάστημα $[\alpha, \beta]$, όπου α και β είναι η μικρότερη και η μεγαλύτερη δυνατή ηλικία των ατόμων, αντίστοιχα.

Σε κάθε επανάληψη του αλγόριθμου η ηλικία κάθε ατόμου αυξάνεται κατά μία μονάδα και όταν ξεπεράσει την μέγιστη ηλικία του, τότε το άτομο «πεθαίνει». Αυτό έχει σαν αποτέλεσμα το συγκεκριμένο άτομο να αντικατασταθεί από ένα άλλο, τυχαία επιλεγμένο, άτομο του πληθυσμού. Πρέπει να σημειώσουμε ότι γενικά δεν είναι επιθυμητό να αντικαταστήσουμε το καλύτερο άτομο του πληθυσμού· για το λόγο αυτό όταν το καλύτερο άτομο «πεθαίνει», απλά αρχικοποιούμε και πάλι τυχαία την μέγιστη ηλικία του. Η ιδέα της γήρανσης γενικά βοηθά την σύγκλιση των ΔΕΑ [152] και σε συνδυασμό με την ιδέα των υποπληθυσμών ο αριθμός των επιτυχιών των ΔΕΑ αυξάνεται σημαντικά.

Εναλλακτικά, και για να αυξήσουμε την ποικιλία του πληθυσμού μπορούμε το νέο άτομο να μην ανήκει στον ήδη υπάρχοντα πληθυσμό, αλλά να αρχικοποιείται τυχαία.

5.5.2 Αποτελέσματα εκπαίδευσης με παράλληλους ΔΕΑ

Στη συνέχεια παραθέτουμε αποτελέσματα εφαρμογής των ΠΔΕΑ στην εκπαίδευση ΤΝΔ με ακέραια βάρη, με και χωρίς σιγμοειδείς συναρτήσεις ενεργοποίησης, στα προβλήματα

του Αποκλειστικού-ΕΙΤΕ (βλ. Παράρτημα Α.1), της ισοτιμίας 3-bit (βλ. Παράρτημα Α.2), και του 4-2-4 Κωδικοποιητή/Αποκωδικοποιητή (βλ. Παράρτημα Α.3). Ονομάζουμε PDE_1 την παράλληλη εκδοχή του αλγόριθμου DE_1 με χρήση υποπληθυσμών και γήρανσης, PDE_2 την παράλληλη εκδοχή του αλγόριθμου DE_2 κτλ.

Δεν επιχειρήσαμε καμία προσπάθεια ρύθμισης των σταθερών μετάλλαξης, ανασυνδυασμού και μετανάστευσης ώστε να αυξήσουμε την ταχύτητα των ΠΔΕΑ. Αντίθετα, χρησιμοποιήσαμε τυπικές προκαθορισμένες τιμές για όλα τα προβλήματα $\mu = 0.5$, $\rho = 0.7$ και $\phi = 0.1$, αντίστοιχα. Μικρότερες τιμές της σταθεράς μετανάστευσης ϕ οδηγούν σε ακόμα λιγότερα μηνύματα μεταξύ των επεξεργαστών, αλλά μπορεί να κάνουν τις μεταναστεύσεις σπάνιες και συνεπώς ανεπαρκείς. Για καθένα από τα παρακάτω πειράματα χρησιμοποιήσαμε 3 υποπληθυσμούς σε τοπολογία δακτυλίου.

Αποκλειστικό-ΕΙΤΕ

Για το πρόβλημα του Αποκλειστικού-ΕΙΤΕ η συνθήκη τερματισμού ήταν να βρεθούν βάρη τέτοια ώστε η συνάρτηση σφάλματος E , να έχει τιμή $E \leq 0.1$. Το μέγεθος κάθε υποπληθυσμού ήταν $NP = 10$. Το κάτω και άνω φράγμα ηλικίας κάθε ατόμου, ήταν $\alpha = 20$ και $\beta = 30$, αντίστοιχα. Στους Πίνακες 5.8 και 5.9 παραθέτουμε τα αποτελέσματα με χρήση σιγμοειδών συναρτήσεων ενεργοποίησης και με τη χρήση κατωφλίων.

Πίνακας 5.8: Αποτελέσματα ΠΔΕΑ με σιγμοειδείς συναρτήσεις ενεργοποίησης (Αποκλειστικό-ΕΙΤΕ)

Αλγόριθμος	Min	μ	Max	σ	Επιτυχία (%)
PDE_1	40	238.4	750	136.3	100%
PDE_2	130	720.1	1840	352.6	100%
PDE_3	80	342.2	1090	186.1	100%
PDE_4	50	395.6	1080	218.5	100%
PDE_5	140	1209.7	3360	661.5	100%
PDE_6	60	402.4	1220	248.4	100%

Πίνακας 5.9: Αποτελέσματα ΠΔΕΑ με κατώφλια (Αποκλειστικό-ΕΙΤΕ)

Αλγόριθμος	Min	μ	Max	σ	Επιτυχία (%)
PDE_1	360	590.4	1150	159.5	100%
PDE_2	420	1197.6	5340	614.0	100%
PDE_3	390	651.3	1250	177.5	100%
PDE_4	380	746.9	1480	225.6	100%
PDE_5	490	1473.8	3790	586.2	100%
PDE_6	410	832.7	2130	294.2	100%

Ισοτιμία των 3-bit

Το δεύτερο πρόβλημα που εξετάσαμε ήταν το πρόβλημα της ισοτιμίας 3-bit. Εδώ κάθε υποπληθυσμός αποτελείται από 11 άτομα και η μέγιστη ηλικία κάθε ατόμου επιλέγεται τυχαία από το διάστημα $[\alpha, \beta]$, όπου $\alpha = 50$ και $\beta = 100$. Η συνθήκη τερματισμού ήταν η συνάρτηση σφάλματος E , να έχει τιμή $E \leq 0.1$. Τα αποτελέσματα φαίνονται στους Πίνακες 5.10 και 5.11

Πίνακας 5.10: Αποτελέσματα ΠΔΕΑ με σιγμοειδείς συναρτήσεις ενεργοποίησης (Ισοτιμία 3-bit)

Αλγόριθμος	<i>Min</i>	μ	<i>Max</i>	σ	Επιτυχία (%)
PDE_1	275	1272.9	3949	619.1	82%
PDE_2	1353	3562.7	8525	1367.8	86%
PDE_3	198	1473.0	6457	873.3	91%
PDE_4	264	2227.3	5104	903.5	99%
PDE_5	1430	4829.6	9306	1598.2	91%
PDE_6	341	2465.8	5412	1011.3	100%

Πίνακας 5.11: Αποτελέσματα ΠΔΕΑ με κατώφλια (Ισοτιμία 3-bit)

Αλγόριθμος	<i>Min</i>	μ	<i>Max</i>	σ	Επιτυχία (%)
PDE_1	561	3022.3	12078	3523.2	100%
PDE_2	1562	5222.3	25377	3757.7	100%
PDE_3	660	2238.8	11847	2100.7	100%
PDE_4	1419	3147.7	11517	1742.4	100%
PDE_5	2387	6868.2	35310	5473.2	100%
PDE_6	1672	2965.8	11088	1863.4	100%

4-2-4 Κωδικοποιητής/Αποκωδικοποιητής

Το τελευταίο πρόβλημα που δοκιμάσαμε τους ΠΔΕΑ, ήταν το πρόβλημα του 4-2-4 κωδικοποιητή/αποκωδικοποιητή. Στις δοκιμές αυτές, το μέγεθος κάθε υποπληθυσμού ήταν $NP = 20$ και το κάτω και άνω φράγμα της ηλικίας των ατόμων ήταν $\alpha = 50$ και $\beta = 200$ αντίστοιχα. Η συνθήκη τερματισμού ήταν η συνάρτηση σφάλματος E , να έχει τιμή $E \leq 0.1$. Ένα τυπικό διάνυσμα βαρών 3-bit είναι το ακόλουθο $w = (0, 2, -2, 3, -3, -3, 2, 3, -3, -3, 2, -3, -2, 2, 3, 2, 1, 0, -3, -3, -2, -2)$, με τιμή της συνάρτησης σφάλματος $E(w) = 0.0459$. Οι Πίνακες 5.12 και 5.13 δείχνουν τα αποτελέσματα από το πρόβλημα του 4-2-4 κωδικοποιητή/αποκωδικοποιητή

Πίνακας 5.12: Αποτελέσματα ΠΔΕΑ με σιγμοειδείς συναρτήσεις ενεργοποίησης (4-2-4 Κωδικοποιητής/Αποκωδικοποιητής)

Αλγόριθμος	<i>Min</i>	μ	<i>Max</i>	σ	Επιτυχία (%)
PDE_1	330	1614.8	4680	868.5	100%
PDE_2	3960	8160.6	13360	2160.5	100%
PDE_3	300	1428.2	4020	660.9	100%
PDE_4	660	4540.5	8520	1505.4	100%
PDE_5	7260	13110.9	20640	3092.4	100%
PDE_6	720	4832.5	9140	1734.3	100%

5.6 Μελέτη της Γενίκευσης

Εκτός από τις δοκιμές για τον έλεγχο της αξιοπιστίας και τις ταχύτητας σύγκλισης των ΔΕΑ, στην ενότητα αυτή εξετάζουμε και την ικανότητα γενίκευσης των δικτύων που προέρ-

Πίνακας 5.13: Αποτελέσματα ΠΔΕΑ με κατώφλια (4-2-4 Κωδικοποιητής/Αποκωδικοποιητής)

Αλγόριθμος	<i>Min</i>	μ	<i>Max</i>	σ	Επιτυχία (%)
PDE_1	980	2520.8	23260	2326.4	100%
PDE_2	4780	8724.9	16580	2264.6	100%
PDE_3	1040	2104.5	4660	680.0	100%
PDE_4	1860	4778.1	9720	1278.0	100%
PDE_5	6080	14070.3	20740	2795.4	100%
PDE_6	1920	5132.3	1820	1321.6	100%

χονται από αυτούς. Για τον σκοπό αυτό δοκιμάσαμε τις καλύτερες από τις ΔΕΑ στρατηγικές (δηλαδή τις DE_3 και DE_4) στο πρόβλημα γενίκευσης MONK (βλ. Παράρτημα Α.4 και [156]).

Δοκιμάσαμε λοιπόν τις DE_3 και DE_4 έναντι των πολύ γνωστών μεθόδων της οπισθοδρομικής διάδοσης του σφάλματος (BP) [133], της μεθόδου της οπισθοδρομικής διάδοσης του σφάλματος με εξασθένηση των βαρών (BP with Weight Decay – BPWD) [124], και της μεθόδου των διαδοχικών συσχετίσεων (Cascade Correlation – CC) [37]. Τα αποτελέσματα από τα προβλήματα MONK συνοψίζονται στον Πίνακα 5.14.

Πίνακας 5.14: Σύγκριση της γενίκευσης στα προβλήματα MONK

Αλγόριθμος	MONK-1	MONK-2	MONK-3
BP	100%	100%	93.1%
BPWD	100%	100%	97.2%
CC	100%	100%	97.2%
DE_3	100%	100%	100%
DE_4	100%	100%	100%

Είναι σαφές από τον Πίνακα 5.14, ότι οι στρατηγικές DE εκπαιδεύουν ΤΝΔ που είναι τουλάχιστον το ίδιο ικανά με τα καλύτερα ΤΝΔ που χρησιμοποιούν πραγματικά βάρη και πραγματικές πολώσεις. Τα ΤΝΔ των στρατηγικών DE έχουν εντοπίσει την ιδέα που βρίσκεται κρυμμένη στα πρότυπα εισόδου. Αυτό γίνεται περισσότερο εμφανές στο πρόβλημα MONK-3, όπου υπάρχει 5% εσκεμμένα λαθεμένη ταξινόμηση και τα ΤΝΔ που εκπαιδεύτηκαν με τις BP, BPWD, και CC δεν έχουν 100% επιτυχία ταξινόμησης.

Η τοπολογία των ΤΝΔ που εξετάσαμε φαίνεται στον Πίνακα 5.15. Όπως είναι γνωστό τα ΤΝΔ που παρουσιάζουν υψηλή γενίκευση δεν είναι ούτε πολύ απλά, αλλά ούτε πολύ σύνθετα· μοιάζουν να ακολουθούν την πολυπλοκότητα του προβλήματος που καλούνται να επιλύσουν.

Για τον λόγο αυτό, για να μπορούν να συγκλίνουν και να έχουν καλή γενίκευση τα ΤΝΔ που χρησιμοποιούν ακέραια βάρη είναι μεγαλύτερα από ΤΝΔ με πραγματικά βάρη, αφού περισσότεροι ακέραιοι αριθμοί χρειάζονται για να εναρμονιστούν με την πολυπλοκότητα του προβλήματος.

Στο τέλος αυτού του κεφαλαίου, στους Πίνακες 5.16, 5.17 και 5.18 παρουσιάζουμε τα ακέραια βάρη από ΤΝΔ που εκπαιδεύτηκαν με τη στρατηγική DE_3 , για τα 3 προβλήματα MONK. Ανάλογα βάρη προκύπτουν και με την εφαρμογή της στρατηγικής DE_4 .

Πίνακας 5.15: Η τοπολογία των δικτύων για τα προβλήματα MONK

Αλγόριθμος	MONK-1	MONK-2	MONK-3
BP	17:3:1	17:2:1	17:4:1
BPWD	17:2:1	17:2:1	17:2:1
CC	17:1:1	17:1:1	17:3:1
DE_3	17:4:1	17:4:1	17:3:1
DE_4	17:4:1	17:4:1	17:3:1

5.7 Συμπεράσματα – Συνεισφορά

Σ' αυτό το κεφάλαιο προτείναμε και μελετήσαμε νέους Διαφοροεξελικτικούς Αλγορίθμους για την εκπαίδευση ΤΝΔ, με σιγμοειδείς συναρτήσεις ενεργοποίησης αλλά και με κατώφλια, που έχουν περιορισμένα ακέραια βάρη. Οι προσαρμοσμένοι αυτοί ΔΕΑ εφαρμόζονται σε ένα πληθυσμό ακέραιων διανυσμάτων βαρών με σκοπό να τον εξελίσσουν κατά τη διάρκεια του χρόνου και να εξερευνήσουν όσο το δυνατόν ευρύτερα το χώρο των ακεραίων βαρών. Αυτό είναι ένα ενδιαφέρον είδος ΤΝΔ, επειδή το ποσό μνήμης που απαιτείται για την αποθήκευση των βαρών είναι σημαντικά μειωμένο έναντι των δικτύων με πραγματικά βάρη. Επιπλέον, οι ψηφιακές διαδικασίες πρόσθεσης και πολλαπλασιασμού που απαιτούνται απλοποιούνται σημαντικά.

Περιορίζοντας τα βάρη και τις πολώσεις στο διάστημα $[-2^{k-1}+1, 2^{k-1}-1]$, για $k = 3, 4, 5$, μπορούν να αντιπροσωπευθούν από ακριβώς k δυαδικά ψηφία. Προφανώς, όσο μικρότερο το k τόσο λιγότερη μνήμη απαιτείται για την αποθήκευση των βαρών. Αφ' ετέρου όμως έχουμε παρατηρήσει ότι η διαδικασία εκπαίδευσης είναι αποτελεσματικότερη και αποδοτικότερη όταν χρησιμοποιούνται περισσότερα δυαδικά ψηφία για την αποθήκευση των βαρών. Κατά συνέπεια, για μια δεδομένη εφαρμογή πρέπει να εξεταστεί η σωστή σχέση μεταξύ της αποτελεσματικότητας και της κατανάλωσης μνήμης. Επίσης είναι γνωστό ότι είναι δύσκολη η υλοποίηση σε υλικό της οπισθοδρομικής διάδοσης των σφαλμάτων, που υπολογίζει το διάνυσμα των μερικών παραγώγων της συνάρτησης σφάλματος. Για το λόγο αυτό όλοι οι προτεινόμενοι αλγόριθμοι απαιτούν μόνο εμπρόσθια διάδοση των προτύπων εκπαίδευσης, για τον υπολογισμό μόνο της τιμής της συνάρτησης σφάλματος.

Η απόδοση αυτών των αλγορίθμων έχει εξεταστεί και έχουν παρουσιαστεί αποτελέσματά τους από κλασικά προβλήματα εκπαίδευσης ΤΝΔ. Συνοψίζοντας τις προσομοιώσεις, καταλήγουμε στο συμπέρασμα ότι οι αλγόριθμοι DE_3 και DE_4 είναι οι καλύτερες επιλογές. Αφ' ετέρου, ακόμη και ο αλγόριθμος DE_1 , βασισμένος στην απλή στρατηγική της Σχέσης (5.1) απέδωσε πολύ καλά. Μια ενδιαφέρουσα παρατήρηση είναι ότι γενικά οι αλγόριθμοι αυξάνουν την απόδοσή τους όταν αυξάνεται η τιμή του k . Τα αποτελέσματα δείχνουν ότι αυτή η νέα κλάση των αλγορίθμων είναι αποτελεσματική, ακόμα και όταν συγκρίνεται με γνωστούς αλγόριθμους που απαιτούν πληροφορίες για την κλίση της συνάρτησης σφάλματος και εκπαιδεύουν ΤΝΔ με πραγματικά βάρη. Επιπλέον, έχουμε εξετάσει την ικανότητα γενίκευσης των δικτύων που παράγονται από τους αλγόριθμους DE_3 και DE_4 . Και οι δύο αλγόριθμοι έχουν άριστη απόδοση και ξεπερνούν άλλους γνωστούς αλγόριθμους εκπαίδευσης με πραγματικά βάρη.

Η χρήση κατωφλίων για όλους τους νευρώνες μειώνει κατά πολύ την πολυπλοκότητα της υλοποίησης σε υλικό, επειδή δεν υπάρχει ανάγκη να σχεδιαστούν και να εφαρμοστούν περίπλοκες μη-γραμμικές συναρτήσεις ενεργοποίησης. Τα ΤΝΔ που είναι βασισμένα σε νευρώνες των οποίων η έξοδος μπορεί να βρίσκεται σε μια ιδιαίτερη κατάσταση ($\{-1, 1\}$ ή $\{0, 1\}$), είναι σημαντικά δεδομένου ότι μπορούν να χειριστούν πολλές από τις εγγενώς δυαδικές στοιχειώδεις εργασίες για τις οποίες τα ΤΝΔ χρησιμοποιούνται. Οι εσωτερικές αναπαραστάσεις τους μπορούν να ερμηνευτούν και είναι υπολογιστικά απλούστερα, σε σχέση με τα ΤΝΔ που χρη-

οιμοποιούν σιγμοειδείς συναρτήσεις ενεργοποίησης [85]. Συνεπώς μπορούν να αποτελέσουν μια βάση για την καλύτερη κατανόηση των ιδιοτήτων και της διεργασίας εκπαίδευσης ΤΝΔ. Τέλος, ένα ακόμα βασικό χαρακτηριστικό των ΤΝΔ που χρησιμοποιούν κατώφλια είναι ότι η εκπαίδευση μπορεί να συνεχιστεί στο υλικό εάν το σύνολο των προτύπων έχει αλλάξει.

Ένα άλλο πλεονέκτημα των ΤΝΔ με τα ακέραια βάρη και πολώσεις, καθώς και κατώφλια είναι ότι το εκπαιδευμένο ΤΝΔ μπορεί να είναι σε πολλές περιπτώσεις ανθεκτικό στο θόρυβο που περιέχεται στα πρότυπα εκπαίδευσης. Τέτοια δίκτυα είναι ικανά να συλλάβουν το βασικό χαρακτηριστικό γνώρισμα των προτύπων εκπαίδευσης (όπως φαίνεται και από τα αποτελέσματα γενίκευσης) και γενικά θόρυβος χαμηλής έντασης δεν μπορεί να διαταράξει τα ακέραια βάρη, αφού απαιτούνται σχετικά μεγάλες διακυμάνσεις, έτσι ώστε τα βάρη και οι πολώσεις να μετακινηθούν από μια ακέραια τιμή στην επόμενη ή στην προηγούμενη. Τέλος, οι προτεινόμενοι αλγόριθμοι μπορούν να εκτελεστούν αποδοτικά από υπολογιστικές μηχανές με έναν ή περισσότερους επεξεργαστές. Για την περίπτωση της παράλληλης εκτέλεσης, προτείναμε και μελετήσαμε νέους αλγόριθμους ΠΔΕΑ.

Πίνακας 5.16: Ακέραια βάρη και πολώσεις για το πρόβλημα MONK-1

Από τον νευρώνα	Προς τον νευρώνα				Έξοδος
	Κρυφό ₁	Κρυφό ₂	Κρυφό ₃	Κρυφό ₄	
Είσοδος ₁	3	-3	3	-1	
Είσοδος ₂	2	-1	2	3	
Είσοδος ₃	3	-2	-3	-2	
Είσοδος ₄	2	-3	3	2	
Είσοδος ₅	1	1	2	3	
Είσοδος ₆	1	-1	-2	1	
Είσοδος ₇	-3	3	-3	3	
Είσοδος ₈	-2	3	-3	3	
Είσοδος ₉	2	2	-3	3	
Είσοδος ₁₀	0	2	-2	3	
Είσοδος ₁₁	1	2	-3	3	
Είσοδος ₁₂	1	-3	-3	3	
Είσοδος ₁₃	2	2	-3	-3	
Είσοδος ₁₄	0	2	-3	-3	
Είσοδος ₁₅	2	2	-3	-3	
Είσοδος ₁₆	-3	-3	-3	2	
Είσοδος ₁₇	-1	-3	-2	2	
Πόλωση	-2	3	3	-3	
Κρυφό ₁					-3
Κρυφό ₂					-1
Κρυφό ₃					-3
Κρυφό ₄					2
Πόλωση					3

Πίνακας 5.17: Ακέραια βάρη και πολώσεις για το πρόβλημα MONK-2

Από τον νευρώνα	Προς τον νευρώνα				Έξοδος
	Κρυφό ₁	Κρυφό ₂	Κρυφό ₃	Κρυφό ₄	
Είσοδος ₁	−2	−1	−3	−2	
Είσοδος ₂	3	2	−2	3	
Είσοδος ₃	2	2	−2	3	
Είσοδος ₄	1	1	−3	−2	
Είσοδος ₅	2	2	−2	3	
Είσοδος ₆	3	2	−2	3	
Είσοδος ₇	2	2	2	−3	
Είσοδος ₈	1	−3	3	2	
Είσοδος ₉	3	1	2	−2	
Είσοδος ₁₀	2	2	3	3	
Είσοδος ₁₁	1	1	3	3	
Είσοδος ₁₂	−2	−3	2	−2	
Είσοδος ₁₃	−1	3	3	3	
Είσοδος ₁₄	−2	0	3	3	
Είσοδος ₁₅	−1	2	3	3	
Είσοδος ₁₆	2	−1	2	−3	
Είσοδος ₁₇	2	2	3	2	
Πόλωση	−1	0	3	−3	
Κρυφό ₁					0
Κρυφό ₂					0
Κρυφό ₃					2
Κρυφό ₄					−2
Πόλωση					0

Πίνακας 5.18: Ακέραια βάρη και πολώσεις για το πρόβλημα MONK-3

Από τον νευρώνα	Προς τον νευρώνα			Έξοδος
	Κρυφό ₁	Κρυφό ₂	Κρυφό ₃	
Είσοδος ₁	0	0	1	
Είσοδος ₂	-1	0	1	
Είσοδος ₃	-1	0	-1	
Είσοδος ₄	0	-2	0	
Είσοδος ₅	-2	-2	2	
Είσοδος ₆	-1	2	0	
Είσοδος ₇	1	1	-2	
Είσοδος ₈	-2	1	1	
Είσοδος ₉	-2	0	0	
Είσοδος ₁₀	-1	2	2	
Είσοδος ₁₁	-2	2	1	
Είσοδος ₁₂	1	-1	1	
Είσοδος ₁₃	-2	-1	-3	
Είσοδος ₁₄	1	-3	0	
Είσοδος ₁₅	0	3	1	
Είσοδος ₁₆	2	0	0	
Είσοδος ₁₇	0	0	1	
Πόλωση	0	1	-2	
Κρυφό ₁				0
Κρυφό ₂				-2
Κρυφό ₃				0
Πόλωση				2

Εκπαίδευση με Μεθόδους Αποφυγής Τοπικών Ελαχίστων

Αγωνιζόμενοι για το καλύτερο,
συχνά καταστρέφουμε το καλό.

—*Shakespeare (1564–1616)*

Το αντικείμενο του κεφαλαίου αυτού είναι η παρουσίαση διαφόρων μεθόδων ολικής βελτιστοποίησης (Global Optimization – GO) καθώς και τεχνικών με τις οποίες αυτές μπορούν να χρησιμοποιηθούν για την αποτελεσματική εκπαίδευση ΤΝΔ [109, 111, 114]. Συνήθως η εκπαίδευση των ΤΝΔ είναι βασισμένη σε μεθόδους τοπικής σύγκλισης, που δεν παρέχουν κανένα μηχανισμό που να τους επιτρέπει να αποφύγουν την περιοχή ενός ανεπιθύμητου τοπικού ελαχίστου. Ένα τοπικό ελάχιστο είναι ανεπιθύμητο όταν το σημείο αυτό έχει σχετικά μεγάλη συναρτησιακή τιμή με αποτέλεσμα το εκπαιδευμένο ΤΝΔ να μην έχει ικανοποιητική γενίκευση ή όταν έχει συναρτησιακή τιμή μεγαλύτερη από την συνθήκη τερματισμού που έχουμε θέσει. Θα παρουσιάσουμε εδώ στρατηγικές που τροποποιούν γνωστές μεθόδους τοπικής σύγκλισης σε νέες μεθόδους ολικής βελτιστοποίησης και θα ερευνήσουμε περαιτέρω τη χρήση τέτοιων μεθόδων για την εκπαίδευση ΤΝΔ.

Οι προτεινόμενες μέθοδοι τείνουν να οδηγούν σε επιθυμητά σύνολα βαρών και επιτρέπουν στο δίκτυο να μάθει ολόκληρο το σύνολο εκπαίδευσης, και, υπό αυτή την έννοια, βελτιώνουν την αποδοτικότητα της διαδικασίας εκπαίδευσης. Παρακάτω, παρουσιάζουμε προσομοιώσεις από προβλήματα γνωστά για τα πολλά και ανεπιθύμητα τοπικά ελάχιστα που παρουσιάζουν, και παρέχουμε μια εκτενή σύγκριση διάφορων αλγορίθμων εκπαίδευσης.

6.1 Εισαγωγή

Όπως έχουμε δει και στα πρώτα κεφάλαια, η εκπαίδευση των ΤΝΔ συνήθως πραγματοποιείται από την ελαχιστοποίηση μιας συνάρτησης κόστους που είναι η συνάρτηση (πολλών μεταβλητών) του σφάλματος του δικτύου. Αυτή η θεώρηση δίνει κάποιο πλεονέκτημα στην ανάπτυξη αποτελεσματικών αλγορίθμων εκπαίδευσης, επειδή το πρόβλημα της ελαχιστοποίησης μιας συνεχούς συνάρτησης πολλών μεταβλητών είναι ευρέως γνωστό στον τομέα της Βελτιστοποίησης και αποτελεί εδώ και χρόνια αντικείμενο έρευνας. Εντούτοις, λόγω των ειδικών χαρακτηριστικών των ΤΝΔ, οι αλγόριθμοι εκπαίδευσης μπορούν να «παγιδευτούν» σε ένα ανεπιθύμητο τοπικό ελάχιστο της συνάρτησης σφάλματος, καθώς είναι βασισμένοι σε μεθόδους τοπικής ανίχνευσης και δεν έχουν κανέναν μηχανισμό που να τους επιτρέπει να «ξεφύγουν» από την περιοχή ενός ανεπιθύμητου τοπικού ελαχίστου.

Σε πρακτικές εφαρμογές, οι μέθοδοι ολικής βελτιστοποίησης μπορεί να ανιχνεύσουν και να εντοπίσουν με ακρίβεια «σχεδόν βέλτιστες» λύσεις της αντικειμενικής συνάρτησης. Σε πολλές περιπτώσεις αυτές οι λύσεις είναι αποδεκτές, αλλά υπάρχουν και εφαρμογές όπου η

βέλτιστη λύση είναι όχι μόνο επιθυμητή αλλά και απαραίτητη. Επομένως, η ανάπτυξη σθεναρών και αποδοτικών μεθόδων ολικής βελτιστοποίησης αποτελεί αντικείμενο της τρέχουσας έρευνας.

Σε αυτό το κεφάλαιο θα μελετήσουμε τη χρήση μεθόδων ολικής βελτιστοποίησης για την εκπαίδευση ΤΝΔ και θα παρουσιάσουμε στρατηγικές ολικής ανίχνευσης (global search), που στοχεύουν να αποτρέψουν το πρόβλημα της περιστασιακής σύγκλισης σε ανεπιθύμητα τοπικά ελάχιστα. Οι μέθοδοι ολικής ανίχνευσης αναμένεται να οδηγήσουν σε «βέλτιστες» ή «σχεδόν βέλτιστες» λύσεις στο χώρο των βαρών και παράλληλα θα παρέχουν μηχανισμούς αποφυγής των τοπικών ελαχίστων κατά τη διάρκεια της εκπαίδευσης.

Αξίζει να σημειωθεί ότι, γενικά οι αλγόριθμοι εκπαίδευσης που είναι βασισμένοι σε μεθόδους ολικής βελτιστοποίησης κατέχουν τις θεωρητικές ιδιότητες σύγκλισης των μεθόδων αυτών, και, τουλάχιστον σε γενικές γραμμές, είναι απλοί στην εφαρμογή και την υλοποίησή τους. Για να αυξήσουμε την αριθμητική αποδοτικότητά τους εξοπλίζουμε τους αλγόριθμους ολικής βελτιστοποίησης με ένα αλγόριθμο τοπικής βελτιστοποίησης, που έχει γρήγορη σύγκλιση σε τοπικά ελάχιστα. Βέβαια, ακόμη και τώρα ο αλγόριθμος ολικής βελτιστοποίησης είναι υπεύθυνος για την ανεύρεση ολικών ελαχίστων, που θεωρητικά απαιτεί «εξαντλητική» ανίχνευση. Αυτές οι παρατηρήσεις δείχνουν την έμφυτη απαίτηση σε υπολογιστικούς πόρους των αλγορίθμων ολικής βελτιστοποίησης, η οποία αυξάνεται μη-πολυωνυμικά, σαν συνάρτηση του μεγέθους του προβλήματος, ακόμη και στις απλούστερες περιπτώσεις εκπαίδευσης ΤΝΔ.

Το πρόβλημα που παρατηρείται στην εκπαίδευση ΤΝΔ είναι ότι τόσο οι αλγόριθμοι που βασίζονται στην οπισθοδρομική διάδοση του σφάλματος όσο και οι αλγόριθμοι δεύτερης τάξης, συχνά συγκλίνουν σε ανεπιθύμητα τοπικά ελάχιστα, με αποτέλεσμα να επηρεάζεται αρνητικά η διαδικασία της εκπαίδευσης του ΤΝΔ, αφού έχει σαν αποτέλεσμα κάποιο διάνυσμα βαρών που δεν είναι ικανοποιητικό. Διαισθητικά, η συνάρτηση σφάλματος παρουσιάζει πολλαπλά τοπικά ελάχιστα διότι είναι η σύνθεση των μη γραμμικών συναρτήσεων ενεργοποίησης (που έχουν ελάχιστα σε διαφορετικά σημεία), με αποτέλεσμα πολλές φορές η τελική συνάρτηση να μην είναι κυρτή [42]. Ο ανεπαρκής αριθμός νευρώνων στο κρυφό στρώμα, η ακατάλληλη αρχικοποίηση των βαρών και η λανθασμένη ρύθμιση των ευρετικών παραμέτρων επιδεινώνουν την κατάσταση με αποτέλεσμα την σύγκλιση σε τοπικά ελάχιστα με μη επιθυμητή συναρτησιακή τιμή. Τελικά το ΤΝΔ δεν καταφέρνει να εκπαιδευτεί σε όλα τα πρότυπα εισόδου και η απόδοσή του δεν είναι η αναμενόμενη.

Αρκετοί ερευνητές έχουν ασχοληθεί με το πρόβλημα της πρόωρης σύγκλισης και έχουν παρουσιάσει αποτελέσματα που βελτιώνουν την απόδοση των μεθόδων, αλλά απαιτούν να ισχύουν ειδικές συνθήκες σχετικά με την τοπολογία του ΤΝΔ, του συνόλου προτύπων εκπαίδευσης, και των αρχικών βαρών [42, 68, 178]. Όμως, συνθήκες όπως η γραμμική διαχωριστικότητα των προτύπων εισόδου και η μορφή «πυραμίδας» για τοπολογία του ΤΝΔ [42], καθώς επίσης και η ανάγκη για μεγαλύτερο αριθμό νευρώνων στο κρυφό επίπεδο (π.χ. τόσοι νευρώνες όσα και τα πρότυπα εισόδου) κάνουν τα ενδιαφέροντα αυτά αποτελέσματα δύσκολα υλοποιήσιμα σε πρακτικές εφαρμογές ακόμα και για τα πιο απλά προβλήματα.

Αξίζει να σημειωθεί ότι μια μέθοδος τοπικής ανίχνευσης μπορεί να τροποποιηθεί έτσι ώστε να έχει ευρεία σύγκλιση χρησιμοποιώντας για παράδειγμα τις συνθήκες του Wolfe που μελετήθηκαν στο Κεφάλαιο 2 ή τις μη μονότονες στρατηγικές που θα αναλύσουμε στο Κεφάλαιο 8. Ένας αλγόριθμος εκπαίδευσης τοπικής ανίχνευσης μπορεί να αποκτήσει την ιδιότητα της ευρείας σύγκλισης με τον καθορισμό ενός τέτοιου βήματος ώστε σε κάθε επανάληψη το σφάλμα να ελαχιστοποιείται ακριβώς κατά μήκος της τρέχουσας κατεύθυνσης ανίχνευσης, δηλαδή να ισχύει σε κάθε επανάληψη $E(w^{k+1}) < E(w^k)$. Για αυτό τον λόγο απαιτείται μια επαναληπτική ευθύγραμμη (μονοδιάστατη) ανίχνευση ανά κατεύθυνση, που είναι συχνά ακριβή από την άποψη των υπολογισμών της τιμής της συνάρτησης σφάλματος. Πρέπει να σημειώσουμε εδώ ότι η ανωτέρω απλή συνθήκη δεν εγγυάται ευρεία σύγκλιση για γενικές συναρτήσεις, δηλαδή την σύγκλιση σε ένα τοπικό ελάχιστο από οποιοδήποτε αρχικό σημείο

(βλ. την εργασία [30] για μια γενική συζήτηση σχετικά με τις μεθόδους ευρείας σύγκλισης). Στη συνέχεια του κεφαλαίου θα εστιάσουμε σε δυο διαφορετικές προσεγγίσεις:

- σε μεθόδους ολικής βελτιστοποίησης που εφαρμόζονται στην εκπαίδευση ΤΝΔ, και
- σε μετασχηματισμούς της αντικειμενικής συνάρτησης (συνάρτηση σφάλματος) με σκοπό την σταδιακή εξάλειψη των ανεπιθύμητων τοπικών ελαχίστων.

Αυτές οι προσεγγίσεις μπορούν γενικά να χρησιμοποιηθούν για την τροποποίηση υπάρχοντων αλγορίθμων για την επίλυση του προβλήματος της περιστασιακής σύγκλισης σε ανεπιθύμητα τοπικά ελάχιστα.

6.2 Μέθοδοι Ολικής Βελτιστοποίησης για την Εκπαίδευση ΤΝΔ

Σ' αυτή την ενότητα θα περιγράψουμε τις βασικές αρχές τριών μεθόδων ολικής βελτιστοποίησης. Συγκεκριμένα θα μελετήσουμε την μέθοδο της προσομοιωμένης ανόπτησης, τους γενετικούς αλγόριθμους, και τη μέθοδο βελτιστοποίησης με σμήνος σωματιδίων. Σ' αυτή την κατηγορία μεθόδων ανήκουν και οι Διαφοροεξελικτικοί αλγόριθμοι που έχουμε εκτενώς αναλύσει στο Κεφάλαιο 5.

6.2.1 Η μέθοδος της προσομοιωμένης ανόπτησης

Η μέθοδος της Προσομοιωμένης Ανόπτησης (ΠΑ) (Simulated Annealing - SA) [62, 90] αναφέρεται στη διαδικασία κατά την οποία εισάγεται στην αντικειμενική συνάρτηση τυχαίος θόρυβος και στη συνέχεια μειώνεται συστηματικά με ένα σταθερό ρυθμό. Έτσι αρχικά υπάρχει «εξερεύνηση» του χώρου των λύσεων, ενώ στη συνέχεια ενισχύεται η ορθή απόκριση του συστήματος κάνοντας δυνατή τη σύγκλιση σε κάποια «βέλτιστη» ή «σχεδόν βέλτιστη» λύση.

Στο πλαίσιο της αριθμητικής βελτιστοποίησης, η ΠΑ είναι μια διαδικασία που έχει την ικανότητα να κινηθεί έξω από τις περιοχές των τοπικών ελαχίστων. Η ΠΑ βασίζεται σε τυχαίους υπολογισμούς της αντικειμενικής συνάρτησης, κατά τέτοιο τρόπο ώστε να είναι δυνατές οι μεταβάσεις έξω από την περιοχή ενός τοπικού ελαχίστου. Δεν εγγυάται, φυσικά, ότι θα συγκλίνει στο ολικό ελάχιστο, αλλά εάν η αντικειμενική συνάρτηση έχει πολλές σχεδόν βέλτιστες λύσεις, τότε πιθανότατα θα εντοπίσει μια από αυτές.

Ειδικότερα, η ΠΑ είναι σε θέση να κάνει διακρίσεις μεταξύ περιοχών με «ακαθόριστη συμπεριφορά» της συνάρτησης σφάλματος και περιοχών με «μικρότερες διακυμάνσεις». Η μέθοδος είναι ικανή μετά από μια φαινομενικά τυχαία αρχική συμπεριφορά, λόγω της μεγάλης έντασης του θορύβου, να κινηθεί σε μια περιοχή της αντικειμενικής συνάρτησης όπου είναι παρόντα επιθυμητά τοπικά ελάχιστα ή κάποιο ολικό ελάχιστο, ασχέτως από τα πιθανά γειτονικά τοπικά ελάχιστα που βρίσκονται στην περιοχή. Όσο η ένταση του θορύβου είναι σχετικά μεγάλη, η μέθοδος κινείται σε διάφορες περιοχές του χώρου λύσεων βρίσκοντας μια περιοχή με ικανοποιητική συναρτησιακή τιμή. Έπειτα, αφού η τιμή του θορύβου έχει γίνει σχετικά μικρή, ασχολείται με τοπικές λεπτομέρειες της περιοχής του χώρου των λύσεων που κατάληξε, βρίσκοντας ένα «σχεδόν βέλτιστο» τοπικό ελάχιστο, εάν όχι το ίδιο το ολικό ελάχιστο.

Στα πλαίσια όμως της εκπαίδευσης ΤΝΔ η απόδοση της κλασικής ΠΑ δεν είναι η αναμενόμενη: η μέθοδος απαιτεί μεγαλύτερο αριθμό υπολογισμών της συνάρτησης σφάλματος από αυτόν που απαιτούν συνήθως αλγόριθμοι εκπαίδευσης πρώτης τάξης και δεν αξιοποιεί πληροφορίες σχετικές με το διάνυσμα των παραγώγων της συνάρτησης σφάλματος. Επιπροσθέτως, η συνάρτηση σφάλματος δεν έχει τοπικά ελάχιστα καλά καθορισμένα σε μια μικρή περιοχή του χώρου των βαρών, αλλά αντιθέτως ευρείες περιοχές που είναι σχεδόν επίπεδες (flat regions). Σε αυτήν την περίπτωση, η ελαχιστοποίηση είναι πολύ δύσκολη και η αποκαλούμενη κίνηση Metropolis (Metropolis move) [88] δεν είναι αρκετά ισχυρή για να απομακρύνει τον αλγόριθμο από αυτές τις περιοχές [170].

Για τους παραπάνω λόγους, οι Burton και Mpitsos [21] προτείνουν την παρακάτω τροποποίηση της κλασικής ΠΑ, η οποία αξιοποιεί πληροφορίες σχετικές με την παράγωγο της συνάρτησης σφάλματος, και μπορεί να χρησιμοποιηθεί για να βελτιώσει τον αλγόριθμο οπισθοδρομικής διάδοσης του σφάλματος:

$$w^{k+1} = w^k - \mu \nabla E(w^k) + nc2^{-dk}, \quad (6.1)$$

όπου n είναι μια παράμετρος που ελέγχει την αρχική ένταση του θορύβου που εισάγουμε στο σύστημα, $c \in (-0.5, +0.5)$ είναι ένας τυχαίος αριθμός από την ομοιόμορφη κατανομή και d είναι η σταθερά μείωσης του θορύβου (noise decay constant).

Στα αποτελέσματα από τα πειράματα που αναφέρουμε στην Ενότητα 6.4, παρουσιάζουμε αποτελέσματα από δύο παραλλαγές της μεθόδου που δίνεται από τη Σχέση (6.1). Στην πρώτη παραλλαγή, έχουμε εφαρμόσει αποκλειστικά την Σχέση (6.1) όπως ακριβώς προτείνεται στο [21]. Ονομάζουμε αυτή τη μέθοδο SA. Εναλλακτικά, μπορούμε να εκπαιδεύουμε το ΤΝΔ με την γνωστή μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος (BP), αλλά όταν αυτή συγκλίνει σε ένα ανεπιθύμητο τοπικό ελάχιστο, χρησιμοποιούμε την μέθοδο SA για να το αποφύγουμε. Ονομάζουμε την συνδυαστική αυτή μέθοδο BPSA.

6.2.2 Γενετικοί αλγόριθμοι

Οι Γενετικοί Αλγόριθμοι (ΓΑ) (Genetic Algorithms – GA) είναι απλοί αλλά ισχυροί αλγόριθμοι βελτιστοποίησης βασισμένοι στους μηχανισμούς της φυσικής επιλογής και της Γενετικής. Το μαθηματικό πλαίσιο των ΓΑ αναπτύχθηκε στη δεκαετία του 1960 και παρουσιάζεται στο πρωτοποριακό βιβλίο του Holland [51]. Οι ΓΑ έχουν χρησιμοποιηθεί πρώτιστα σε δύσκολα προβλήματα βελτιστοποίησης και οι βασικές αρχές της λειτουργίας τους περιγράφονται σύντομα ως εξής: σε κάθε γενιά¹ (generation) ενός ΓΑ, δημιουργείται ένα νέο σύνολο προσεγγίσεων της λύσης χρησιμοποιώντας τους τελεστές των ΓΑ (που είναι δανεισμένοι από τη Γενετική) και επιλέγοντας τα άτομα για την επόμενη γενιά σύμφωνα με τον τελεστή επιλογής. Αυτή η διαδικασία οδηγεί στην εξέλιξη του πληθυσμού των ατόμων, που έτσι προσαρμόζονται καλύτερα στο περιβάλλον από τους προγόνους τους, ακριβώς όπως στην φυσική επιλογή.

Πιο συγκεκριμένα, ένας απλός ΓΑ επεξεργάζεται έναν πεπερασμένο πληθυσμό δυαδικών διανυσμάτων σταθερού μήκους, που αποκαλούνται *χρωμοσώματα* (chromosomes). Οι συνιστώσες του χρωμοσώματος αποκαλούνται γονίδια (genes). Οι ΓΑ έχουν δύο βασικούς τελεστές: (α) τον *τελεστή ανασυνδυασμού* (crossover operator) των χρωμοσωμάτων, και (β) τον *τελεστή μετάλλαξης* (mutation operator) των χρωμοσωμάτων για την τυχαία μεταβολή της τιμής των γονιδίων τους.

Ο τελεστής ανασυνδυασμού ερευνά τις διαφορετικές δομές και σχήματα των χρωμοσωμάτων, με την ανταλλαγή μια ομάδας γονιδίων μεταξύ δύο προεπιλεγμένων χρωμοσωμάτων, ενώ ο τελεστής μετάλλαξης χρησιμεύει πρώτιστα στην αποφυγή τοπικών ελαχίστων του χώρου των βαρών με την αλλαγή της τιμής ενός γονιδίου, εισάγοντας έτσι ποικιλομορφία στον πληθυσμό. Ένα μεγάλο μέρος της αποτελεσματικότητας της αναζήτησης των ΓΑ οφείλεται στην συνδυασμένη δράση του ανασυνδυασμού και της μετάλλαξης. Ένας άλλος τελεστής που συνδέεται με τους παραπάνω και δρα μετά από αυτούς, είναι ο *τελεστής επιλογής* (selection operator), ο οποίος είναι υπεύθυνος για την επιλογή των χρωμοσωμάτων για την επόμενη γενιά και την επιβίωση των καταλληλότερων. Για μια περιγραφή της λειτουργίας ενός απλού ΓΑ βλέπε το Σχήμα 6.1.

Η δυνατότητα της παράλληλης εκτέλεσης των ΓΑ, η αντοχή τους στον θόρυβο, καθώς επίσης και η ικανότητά τους να ανιχνεύουν και περιοχές με μεγαλύτερες συναρτησιακές τιμές από το τρέχων καλύτερο σημείο, καθιστούν τους ΓΑ κατεξοχήν κατάλληλους για την

¹Ο όρος γενιά στους ΓΑ ταυτίζεται με τον όρο επανάληψη σε οποιονδήποτε άλλο κλασικό αλγόριθμο βελτιστοποίησης.

εκπαίδευση ΤΝΔ, καθώς φαίνονται ικανοί να εξερευνήσουν και να ανιχνεύσουν αποτελεσματικά το χώρο των βαρών. Για τα πειράματά μας χρησιμοποιήσαμε το πρόγραμμα Genetic Algorithm for Optimization Toolbox (GAOT) [55]. Χρησιμοποιήσαμε τους προεπιλεγμένους τελεστές ανασυνδυασμού και μετάλλαξης του GAOT και πραγματικούς αριθμούς για την κωδικοποίηση των βαρών των ΤΝΔ.

ΜΟΝΤΕΛΟ ΑΠΛΟΥ ΓΕΝΕΤΙΚΟΥ ΑΛΓΟΡΙΘΜΟΥ

```
{
  //Αρχικοποίησε την μεταβλητή μέτρησης του χρόνου
  t := 0;
  //Αρχικοποίησε τον πληθυσμό
  InitPopulation(P(t));
  //Υπολόγισε την αντικειμενική συνάρτηση για όλα τα άτομα
  Evaluate(P(t));
  //Έλεγχε το κριτήριο τερματισμού
  while not Termination Criterion do
    t := t + 1;
    //Επέλεξε τον υποπληθυσμό για την δημιουργία των απογόνων
    Q(t) := SelectParents(P(t));
    //Ανασυνδύασε τα «γονίδια» των γονέων
    Recombine(Q(t));
    //Διατάραξε στοχαστικά τον νέο πληθυσμό
    Mutate(Q(t));
    //Υπολόγισε την αντικειμενική συνάρτηση για όλα τα άτομα
    Evaluate(Q(t));
    //Επέλεξε τα άτομα για την επόμενη γενιά
    P(t + 1) := Survive(P(t), Q(t));
  end
}
```

Σχήμα 6.1: Περιγραφή ενός απλού Γενετικού Αλγόριθμου

6.2.3 Η μέθοδος βελτιστοποίησης με σμήνος σωματιδίων

Η μέθοδος Βελτιστοποίησης με Σμήνος Σωματιδίων (ΒΣΣ) (Particle Swarm Optimization – PSO) χρησιμοποιεί ένα πληθυσμό από πιθανές λύσεις, που η δυναμική του μπορεί να παρομοιασθεί με την συμπεριφορά ενός σμήνους σωματιδίων ή καλύτερα πουλιών. Έτσι πραγματοποιείται διανομή των πληροφοριών μεταξύ των μελών του σμήνους και τα άτομα μπορούν να ωφεληθούν από τις ανακαλύψεις και την προηγούμενη εμπειρία όλων των άλλων συντρόφων τους κατά τη διάρκεια της αναζήτησης τροφής! Κατά συνέπεια, κάθε ένας σύντροφος, που ονομάζεται *σωματίδιο* (particle), του πληθυσμού, που καλείται τώρα *σμήνος* (swarm), υποτίθεται ότι «πετά» πάνω από τον χώρο των βαρών προκειμένου να βρεθούν οι πιο «ελπιδοφόρες» περιοχές του τοπίου. Παραδείγματος χάριν, στην περίπτωση ελαχιστοποίησης, τέτοιες περιοχές είναι αυτές που έχουν τις μικρότερες συναρτησιακές τιμές. Σε αυτό το πλαίσιο, κάθε σωματίδιο αντιμετωπίζεται ως ένα σημείο στον N -διάστατο χώρο των βαρών, που ρυθμίζει την «πτήση» του σύμφωνα με την εμπειρία του καθώς επίσης και την εμπειρία των άλλων σωματιδίων (σύντροφοι).

Μέχρι σήμερα έχουν προταθεί πολλές παραλλαγές της ΒΣΣ, μετά από την εισαγωγή της από τους Eberhart και Kennedy [33, 61]. Στα πειράματά μας έχουμε χρησιμοποιήσει μια πρόσφατη τροποποίηση αυτού του αλγορίθμου, η οποία παράγεται με την προσθήκη μια νέας

παραμέτρου (βάρος αδράνειας) στην αρχική δυναμική της ΒΣΣ [32]. Αυτή η τροποποίηση περιγράφεται στη συνέχεια.

Πρώτα ας καθορίσουμε τον συμβολισμό που θα χρησιμοποιήσουμε σε αυτή την παράγραφο: το i -στο σωματίδιο του σμήνους αναπαρίσταται από ένα N -διάστατο διάνυσμα $X_i = (x_{i1}, x_{i2}, \dots, x_{iN})$ και το καλύτερο σωματίδιο του σμήνους, δηλαδή αυτό που έχει τη μικρότερη συνάρτησιακή τιμή δηλώνεται από τον δείκτη g . Η προηγούμενη καλύτερη θέση του i -στού σωματιδίου καταγράφεται και συμβολίζεται σαν $P_i = (p_{i1}, p_{i2}, \dots, p_{iN})$, και τέλος η μεταβολή της θέσης του κάθε σωματιδίου, δηλαδή η *ταχύτητα* (velocity) του i -στού σωματιδίου συμβολίζεται $V_i = (v_{i1}, v_{i2}, \dots, v_{iN})$. Το μέγεθος του σμήνους (NP) παραμένει σταθερό καθ' όλη την διάρκεια της εκπαίδευσης.

Η ταχύτητα και η θέση των σωματιδίων του σμήνους υπολογίζονται από τις ακόλουθες εξισώσεις:

$$v_{in} = w v_{in} + c_1 r_1 (p_{in} - x_{in}) + c_2 r_2 (p_{gn} - x_{in}), \quad (6.2)$$

$$x_{in} = x_{in} + v_{in}, \quad (6.3)$$

όπου $n = 1, 2, \dots, N$, $i = 1, 2, \dots, NP$, w είναι το βάρος αδράνειας, c_1 και c_2 είναι θετικές σταθερές, και r_1 και r_2 είναι τυχαίοι πραγματικοί αριθμοί από το διάστημα $[0, 1]$.

Η Σχέση (6.2) χρησιμοποιείται για τον υπολογισμό της νέας ταχύτητας του i -στού σωματιδίου, λαμβάνοντας υπόψη τρεις όρους: την προηγούμενη ταχύτητα του σωματιδίου, την απόσταση μεταξύ της καλύτερης προηγούμενης και της τρέχουσας θέσης του σωματιδίου, και την απόσταση μεταξύ της καλύτερης εμπειρίας του σμήνους (τη θέση του καλύτερου σωματιδίου του σμήνους) και της τρέχουσας θέσης του σωματιδίου. Κατόπιν, το i -στο σωματίδιο κινείται σε μια νέα θέση σύμφωνα με τη Σχέση (6.3).

Ο ρόλος του βάρους αδράνειας w , θεωρείται πολύ σημαντικός στη συμπεριφορά σύγκλισης της ΒΣΣ. Το βάρος αδράνειας χρησιμοποιείται για να ελέγξει τον αντίκτυπο της προηγούμενης ιστορίας των ταχυτήτων στην τρέχουσα ταχύτητα. Κατ' αυτό τον τρόπο, η παράμετρος w παρέχει εξισορρόπηση μεταξύ της ολικής και τοπικής δυνατότητας αναζήτησης του σμήνους.

Ένα σχετικά μεγάλο βάρος αδράνειας διευκολύνει την ολική εξερεύνηση (συνεπώς και την ανίχνευση νέων περιοχών), ενώ ένα σχετικά μικρό βάρος αδράνειας τείνει να διευκολύνει την τοπική εξερεύνηση, δηλαδή την σύγκλιση σε κάποιο ελάχιστο. Η κατάλληλη τιμή για το βάρος αδράνειας w παρέχει συνήθως ισορροπία μεταξύ της ολικής και τοπικής ανίχνευσης και συνεπώς έχει σαν αποτέλεσμα τη μείωση του αριθμού των επαναλήψεων που απαιτούνται για να εντοπιστεί μια βέλτιστη λύση. Ένας εμπειρικός κανόνας προτείνει ότι είναι καλύτερα το βάρος αδράνειας να έχει αρχικά μια σχετικά μεγάλη τιμή, προκειμένου να γίνει η καλύτερη δυνατή ανίχνευση του χώρου των βαρών, και βαθμιαία να μειώνεται ώστε να βοηθηθεί η σύγκλιση της μεθόδου. Για το λόγο αυτό η τιμή του βάρους αδράνειας μεταβάλλεται αντιστρόφως ανάλογα του αριθμού των επαναλήψεων.

Από την ανωτέρω συζήτηση είναι προφανές ότι ως ένα ορισμένο βαθμό η ΒΣΣ μοιάζει με τους Εξελικτικούς Αλγορίθμους που μελετήσαμε στο Κεφάλαιο 5. Εντούτοις, η ΒΣΣ αντί να χρησιμοποιούν γενετικούς τελεστές, κάθε σωματίδιο ενημερώνει τη θέση του βασισμένο στην εμπειρία του και στην εμπειρία και τις «ανακαλύψεις» των συντρόφων του. Η προσθήκη του όρου της ταχύτητας στην τρέχουσα θέση του σωματιδίου, προκειμένου να υπολογιστεί η επόμενη θέση, μοιάζει με τον τελεστή μετάλλαξης. Σημειώνουμε ότι στην ΒΣΣ, εντούτοις, ο τελεστής μετάλλαξης καθοδηγείται από την εμπειρία του σωματιδίου αλλά και του σμήνους. Με άλλα λόγια, η ΒΣΣ θεωρείται να χρησιμοποιεί ένα τελεστή μετάλλαξης με «συνείδηση», όπως επισημαίνεται στο [32].

6.3 Μετασχηματισμοί της Συνάρτησης Σφάλματος

Σ' αυτή την ενότητα θα εξετάσουμε μετασχηματισμούς της συνάρτησης σφάλματος με σκοπό της σταδιακή εξάλειψη ανεπιθύμητων τοπικών ελαχίστων. Έστω το σημείο $\bar{w}$ τότε ώστε υπάρχει μια γειτονιά $\mathcal{B}$ του $\bar{w}$ με:

$$E(\bar{w}) \leq E(w), \quad \forall w \in \mathcal{B}. \quad (6.4)$$

Τότε αυτό το σημείο είναι ένα τοπικό ελάχιστο της συνάρτησης σφάλματος και, όπως έχει ήδη αναφερθεί ανωτέρω, πολλές μέθοδοι παγιδεύονται σε τέτοια ανεπιθύμητα τοπικά ελάχιστα. Η κύρια ιδέα της εφαρμογής ενός μετασχηματισμού στη συνάρτηση σφάλματος είναι μερικά από τα ανεπιθύμητα τοπικά ελάχιστα να εξαφανιστούν, ενώ το ολικό ελάχιστο καθώς και άλλα επιθυμητά τοπικά ελάχιστα να παραμείνουν αμετάβλητα. Οι τεχνικές που θα περιγράψουμε κατωτέρω στοχεύουν στο μετασχηματισμό της συνάρτησης σφάλματος κατά τέτοιο τρόπο ώστε η σύγκλιση σε ένα ολικό ή γενικότερα επιθυμητό τοπικό ελάχιστο ενισχύεται για *οποιοδήποτε* αλγόριθμο εκπαίδευσης που είναι εξοπλισμένος με αυτές. Δύο τεχνικές περιγράφονται παρακάτω: (α) η τεχνική της παρεκκλίνουσας τροχιάς, και (β) η τεχνική του «εφελκυσμού» της αντικειμενικής συνάρτησης.

6.3.1 Η τεχνική της παρεκκλίνουσας τροχιάς

Η τεχνική της Παρεκκλίνουσας Τροχιάς (deflection procedure) προτάθηκε στο [83], και σύμφωνα με αυτή όταν η ακολουθία των διανυσμάτων των βαρών $\{w^k\}_0^\infty$ συγκλίνει σε ένα ανεπιθύμητο τοπικό ελάχιστο $\bar{w} \in \mathbb{R}^N$, τότε η συνάρτηση σφάλματος $E(w)$ μετασχηματίζεται σύμφωνα με την ακόλουθη σχέση:

$$F(w) = S(w; \bar{w}, \lambda)^{-1} E(w),$$

όπου $S(w; \bar{w}, \lambda)$ είναι μια συνάρτηση που εξαρτάται από το διάνυσμα w και το τοπικό ελάχιστο $\bar{w}$ της E , και λ είναι μια παράμετρος χαλάρωσης (relaxation parameter). Στην περίπτωση που θέλουμε να εξαλείψουμε m τοπικά ελάχιστα της συνάρτησης σφάλματος $\bar{w}_1, \dots, \bar{w}_m \in \mathbb{R}^N$, η παραπάνω σχέση γίνεται:

$$F(w) = S(w; \bar{w}_1, \lambda_1)^{-1} \dots S(w; \bar{w}_m, \lambda_m)^{-1} E(w).$$

Η τεχνική της παρεκκλίνουσας τροχιάς παρέχει μια «κατάλληλη» συνάρτηση $S(\cdot)$ τέτοια ώστε η νέα συνάρτηση $F(w)$ να μην παρουσιάζει ελάχιστο στα σημεία $\bar{w}_i, i = 1, \dots, m$, ενώ όλα τα υπόλοιπα ελάχιστα της αρχικής συνάρτησης σφάλματος E να παραμένουν αμετάβλητα. Δηλαδή, κατασκευάζουμε συναρτήσεις S που παρέχουν στην συνάρτηση F την ακόλουθη ιδιότητα: για κάθε τοπικό ελάχιστο $\bar{w}_i$ της συνάρτησης E (το οποίο βρίσκει η μέθοδος εκπαίδευσης), η F δεν θα παρουσιάζει ελάχιστο στο $w = \bar{w}_i$. Επιπρόσθετα, η συνάρτηση F διατηρεί όλα τα άλλα ελάχιστα της E . Αυτή είναι η τεχνική της παρεκκλίνουσας τροχιάς [83]. Για παράδειγμα έστω η συνάρτηση:

$$S(w; \bar{w}_i, \lambda_i) = \tanh(\lambda_i \|w - \bar{w}_i\|),$$

η οποία δίνει στην νέα συνάρτηση F αυτή την ιδιότητα, όπως θα δείξουμε παρακάτω.

Ας υποθέσουμε ότι έχουμε βρει ένα ανεπιθύμητο τοπικό ελάχιστο $\bar{w}_i$. Τότε:

$$\lim_{w \rightarrow \bar{w}_i} \frac{E(w)}{\tanh(\lambda \|w - \bar{w}_i\|)} = +\infty,$$

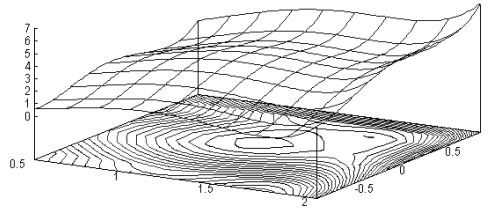
που σημαίνει ότι το $\bar{w}_i$ δεν είναι ελάχιστο της F . Επιπροσθέτως, είναι εύκολο να δειχθεί ότι

για $\|w - \bar{w}_i\| \geq \varepsilon$, όπου ε είναι μια μικρή θετική σταθερά, ισχύει

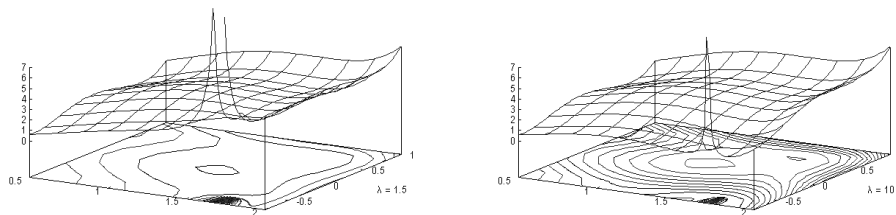
$$\lim_{\lambda \rightarrow +\infty} F(w) = \lim_{\lambda \rightarrow +\infty} \frac{E(w)}{\tanh(\lambda \|w - \bar{w}_i\|)} = E(w), \quad (6.5)$$

καθώς ο παρονομαστής τείνει στην μονάδα. Συνεπώς, η συνάρτηση σφάλματος παραμένει αμετάβλητη στον υπόλοιπο χώρο των βαρών.

Αξίζει να σημειωθεί ότι το αποτέλεσμα της τεχνικής της παρεκκλίνουσας τροχιάς εξαρτάται από το πρόβλημα και την τιμή της παραμέτρου λ . Για τυχαίες τιμές του λ υπάρχει μια μικρή γειτονιά $\mathcal{R}(\bar{w}, \rho)$ με κέντρο $\bar{w}$ και ακτίνα ρ , με $\rho \propto \lambda^{-1}$, όπου για κάθε $x \in \mathcal{R}(\bar{w}, \rho)$ ισχύει ότι $F(w) > E(w)$. Πιο συγκεκριμένα, όταν η τιμή του λ είναι σχετικά μικρή (έστω $\lambda < 1$), ο παρονομαστής της Σχέσης (6.5) γίνεται ίσος με την μονάδα για w «μακριά» από το $\bar{w}$. Έτσι, η τεχνική της παρεκκλίνουσας τροχιάς επηρεάζει μια «μεγάλη» γειτονιά του χώρου των βαρών γύρω από το $\bar{w}$. Όταν η τιμή της παραμέτρου λ είναι σχετικά μεγάλη, νέα τοπικά ελάχιστα είναι πιθανό να δημιουργηθούν στη γειτονιά του $\bar{w}$, σαν ένα «Μεξικάνικο καπέλο». Αυτά τα τοπικά ελάχιστα έχουν συναρτησιακές τιμές μεγαλύτερες από $F(\bar{w})$ και μπορούν εύκολα να αποφευχθούν με την επιλογή ενός κατάλληλου ρυθμού εκπαίδευσης ή αλλάζοντας την τιμή του λ . Μπορούμε να δούμε το αποτέλεσμα της τεχνικής της παρεκκλίνουσας τροχιάς στην



Σχήμα 6.2: Γραφική παράσταση της συνάρτησης Six Hump Camel Back



Σχήμα 6.3: Εφαρμόζοντας την τεχνική της παρεκκλίνουσας τροχιάς στη συνάρτηση Six Hump Camel Back, για $\lambda = 1.5$ και $\lambda = 10$

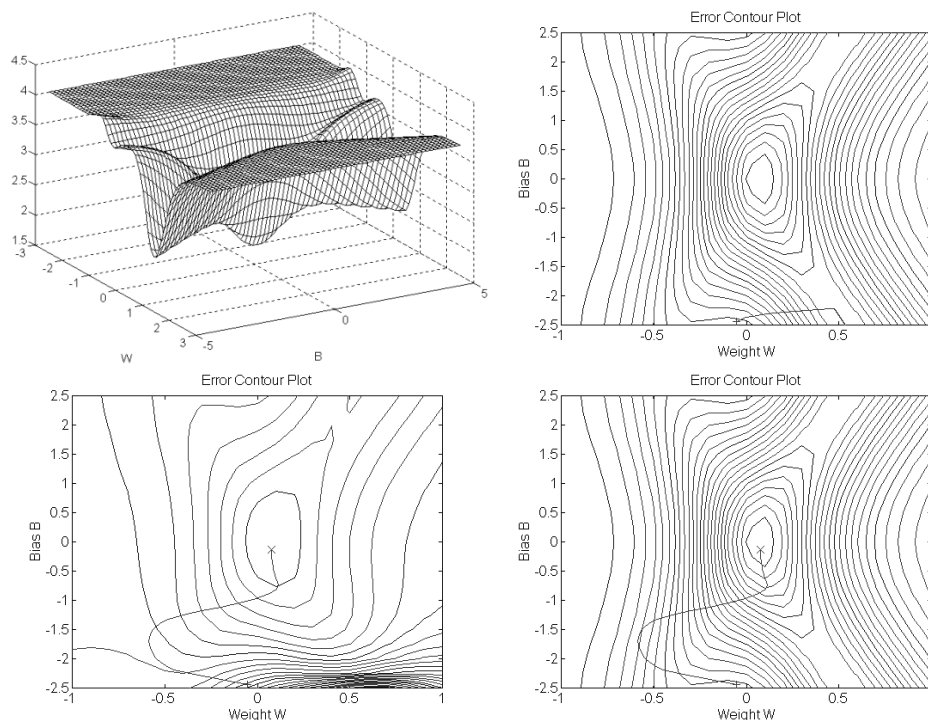
γνωστή συνάρτηση Six Hump Camel Back, που δίνεται από την έκφραση:

$$f(x_1, x_2) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4.$$

Η γραφική παράσταση της συνάρτησης φαίνεται στο Σχήμα 6.2. Η εφαρμογή της τεχνικής της παρεκκλίνουσας τροχιάς για $\lambda = 1.5$ απεικονίζεται στο Σχήμα 6.3 (αριστερά) και για

$\lambda = 10$ στο Σχήμα 6.3 (δεξιά).

Για να δώσουμε μια εποπτική εικόνα της τεχνικής της παρεκκλίνουσας τροχιάς δίνουμε ακόμα ένα παράδειγμα, το οποίο αφορά την εκπαίδευση ενός νευρώνα (single neuron) με την μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος. Το πρόβλημα αφορά την εκμάθηση 8 ζευγών τιμών εισόδου-εξόδου από το ελάχιστο αυτό ΤΝΔ. Η επιφάνεια σφάλματος απεικονίζεται στο Σχήμα. 6.4 (επάνω αριστερά). Το επιθυμητό ελάχιστο βρίσκεται στο κέντρο του σχήματος, και επίσης υπάρχουν δυο «κοιλιάδες» που οδηγούν σε ανεπιθύμητα τοπικά ελάχιστα. Στο Σχήμα 6.4 (επάνω δεξιά) απεικονίζεται η τροχιά των διανυσμάτων των βαρών κατά την διάρκεια της εκπαίδευσης, όπου οι αρχική συνθήκη οδήγησε την μέθοδο σε ένα από τα ανεπιθύμητα τοπικά ελάχιστα. Στο Σχήμα 6.4 (κάτω αριστερά) και στο Σχήμα 6.4 (κάτω δεξιά) παρουσιάζουμε την παρεκκλίνουσα τροχιά των διανυσμάτων των βαρών, η οποία απεικονίζεται μαζί με τις ισοϋψείς καμπύλες του μετασχηματισμού F , καθώς και της αρχικής συνάρτησης σφάλματος E .



Σχήμα 6.4: Εφαρμόζοντας την μέθοδο της παρεκκλίνουσας τροχιάς στην εκπαίδευση ΤΝΔ (με x σημειώνουμε το ελάχιστο της συνάρτησης E)

Αξίζει να παρατηρήσουμε ότι η τεχνική της παρεκκλίνουσας τροχιάς μπορεί να ενσωματωθεί σε *οποιοδήποτε* αλγόριθμο εκπαίδευσης ΤΝΔ και βοηθά στην αποφυγή ανεπιθύμητων τοπικών ελαχίστων. Στα αποτελέσματα που παρουσιάζουμε παρακάτω, έχουμε δοκιμάσει την κλασική μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος σε συνδυασμό με την τεχνική της παρεκκλίνουσας τροχιάς. Ονομάζουμε την νέα μέθοδο (BPD).

6.3.2 Η τεχνική του «εφελκυσμού» της αντικειμενικής συνάρτησης

Η τεχνική του «εφελκυσμού» της αντικειμενικής συνάρτησης (function “stretching”) αποτελεί ένα μετασχηματισμό της συνάρτησης σφάλματος $E(w)$ και μπορεί να εφαρμοστεί αφού ένα τοπικό ελάχιστο $\bar{w}$ της συνάρτησης E έχει βρεθεί [100–102]. Ο μετασχηματισμός

αποτελείται από δύο φάσεις, που υλοποιούνται σύμφωνα με τις ακόλουθες σχέσεις:

$$G(w) = E(w) + \frac{\gamma_1}{2} \|w - \bar{w}\| (\text{sign}(E(w) - E(\bar{w})) + 1), \quad (6.6)$$

$$H(w) = G(w) + \gamma_2 \frac{\text{sign}(E(w) - E(\bar{w})) + 1}{2 \tanh(\mu(G(w) - G(\bar{w})))}, \quad (6.7)$$

όπου γ_1, γ_2 και μ είναι τυχαίες θετικές σταθερές, και $\text{sign}(\cdot)$ είναι η γνωστή συνάρτηση του προσήμου. Η συνάρτηση του προσήμου μπορεί να προσεγγιστεί από μια σιγμοειδή συνάρτηση, όπως π.χ. η λογιστική συνάρτηση ή η υπερβολική εφαπτομένη:

$$\text{sign}(w) \approx \text{logsig}(w) = \frac{2}{1 + \exp(-\lambda w)} - 1 \equiv \tanh\left(\frac{\lambda}{2} w\right),$$

για μεγάλες τιμές του λ . Αυτές οι σιγμοειδείς συναρτήσεις είναι συνεχώς παραγωγίσιμες και χρησιμοποιούνται σαν συναρτήσεις ενεργοποίησης ΤΝΔ.

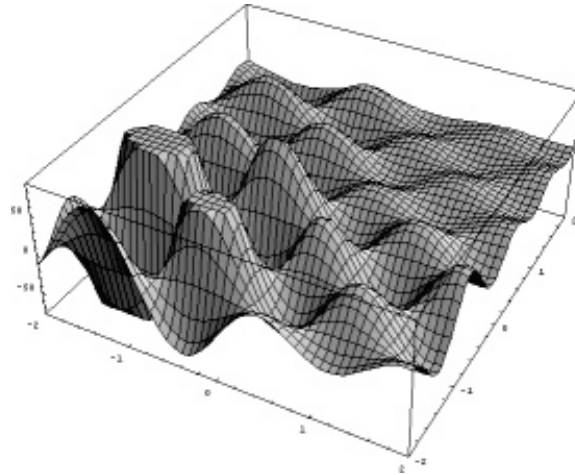
Πρέπει να σημειωθεί ότι η πρώτη φάση του μετασχηματισμού, που δίνεται από τη Σχέση (6.6), «ανασηκώνει» την συνάρτηση σφάλματος $E(w)$ με σκοπό να απαλείψει όλα τα ελάχιστα που έχουν μεγαλύτερες τιμές από $E(\bar{w})$. Η δεύτερη φάση του μετασχηματισμού, που δίνεται από τη Σχέση (6.7), «τεντώνει» την γειτονιά του σημείου $\bar{w}$ προς τα επάνω, αφού δίνει σε όλα αυτά τα σημεία μεγαλύτερες συναρτησιακές τιμές. Και οι δύο φάσεις του μετασχηματισμού δεν τροποποιούν περιοχές της συνάρτησης σφάλματος που έχουν μικρότερη συναρτησιακή τιμή από το τοπικό ελάχιστο $\bar{w}$. Έτσι το ολικό ελάχιστο και όλα τα ελάχιστα με μικρότερη συναρτησιακή τιμή παραμένουν αμετάβλητα.

Σε αυτό το σημείο παρέχουμε ένα παράδειγμα της τεχνικής αυτής για να δείξουμε την λειτουργία της. Το πρόβλημα που θα δοκιμάσουμε είναι η συνάρτηση δύο μεταβλητών που ονομάζεται Levy No. 5 [70], είναι γνωστή για την πληθώρα τοπικών ελαχίστων που έχει και δίνεται από τον ακόλουθο τύπο:

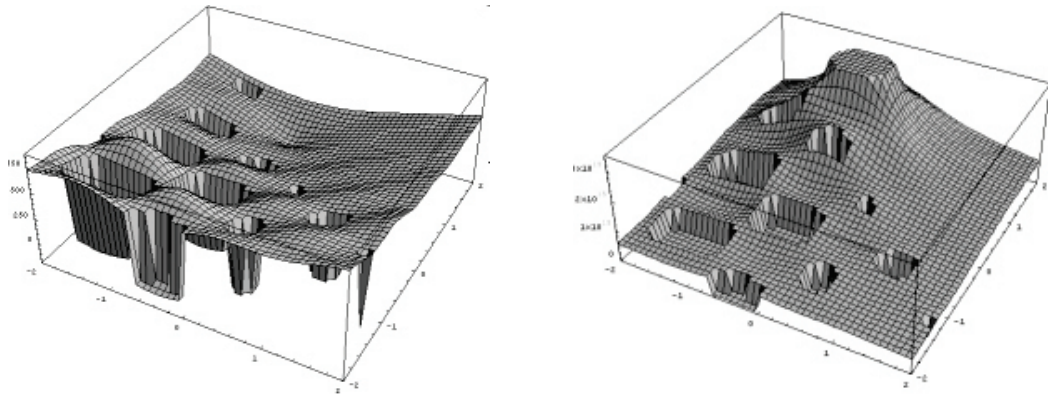
$$\begin{aligned} f(x) = & \sum_{i=1}^5 i \cos[(i+1)x_1 + i] \times \sum_{j=1}^5 j \cos[(j+1)x_2 + j] + \\ & + (x_1 + 1.42513)^2 + (x_2 + 0.80032)^2, \end{aligned} \quad (6.8)$$

όπου $-10 \leq x_i \leq 10, i = 1, 2$. Η συνάρτηση αυτή παρουσιάζει περίπου 760 τοπικά ελάχιστα και ένα ολικό ελάχιστο στο σημείο $x^* = (-1.3068, -1.4248)$ με συναρτησιακή τιμή $f^* = -176.1375$. Ο μεγάλος αριθμός τοπικών ακροτάτων κάνει την προσπάθεια για τον εντοπισμό του ολικού ελαχίστου εξαιρετικά δύσκολη. Στο Σχήμα 6.5, παρουσιάζουμε την γραφική παράσταση της Levy No. 5 στο τετράγωνο $[-2, 2]^2$.

Αφού εφαρμόσουμε τον μετασχηματισμό της Σχέσης (6.6) (πρώτη φάση του μετασχηματισμού) στη συνάρτηση Levy No. 5, η νέα μορφή της συνάρτησης φαίνεται στο Σχήμα 6.6 (αριστερά). Είναι φανερό ότι τα ελάχιστα με συναρτησιακές τιμές μεγαλύτερες από το ελάχιστο στο οποίο εφαρμόσαμε τον μετασχηματισμό απαλείφονται, ενώ η περιοχή «κάτω» από το τοπικό ελάχιστο $\bar{w}$ παραμένει αναλλοίωτη. Στο Σχήμα 6.6 (δεξιά), παρουσιάζουμε την τελική συνάρτηση μετά την εφαρμογή και της Σχέσης (6.7) (δεύτερη φάση του μετασχηματισμού) στην συνάρτηση Levy No. 5. Είναι σαφές από το σχήμα ότι ολόκληρη η γειτονιά του τοπικού ελαχίστου έχει «ανυψωθεί», με αποτέλεσμα το ελάχιστο να έχει μετατραπεί σε μέγιστο. Στη συνέχεια δίνουμε λεπτομέρειες σχετικά με την απόδοση της τεχνικής αυτής σε συνδυασμό με τον αλγόριθμο ΒΣΣ. Ονομάζουμε (SPSO) την νέα σύνθετη μέθοδο.



Σχήμα 6.5: Γραφική παράσταση της συνάρτησης Levy No. 5



Σχήμα 6.6: Γραφική παράσταση της συνάρτησης Levy No. 5 μετά από το πρώτο στάδιο (αριστερά) και μετά από το δεύτερο στάδιο (δεξιά) του προτεινόμενου μετασχηματισμού

6.4 Αποτελέσματα

Στην ενότητα αυτή δίνουμε συγκριτικά αποτελέσματα και μια αξιολόγηση της απόδοσης των παραπάνω μεθόδων, τις οποίες έχουμε δοκιμάσει σε δύο προβλήματα εκπαίδευσης γνωστά για τα πολλά τοπικά ελάχιστα που παρουσιάζουν, και τις συγκρίνουμε με την μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος (BP) [133] και την μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος με ορμή (BPM) [57, 133]. Οι αλγόριθμοι αρχικοποιήθηκαν με τα ίδια αρχικά βάρη, που επιλέχθηκαν τυχαία από την ομοιόμορφη κατανομή στο διάστημα $(-1, 1)$. Οι μέθοδοι BPSA και BPD ενημερώνουν το διάνυσμα των βαρών χρησιμοποιώντας την μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος, έως ότου επιτευχθεί σύγκλιση σε ένα ολικό ή επιθυμητό τοπικό ελάχιστο. Το διάνυσμα των βαρών w^k θεωρείται σημείο επιθυμητού τοπικού ελαχίστου αν $E(w^k) \leq 0.04$. Πρέπει να σημειωθεί ότι η σύγκλιση σε ένα τοπικό ελάχιστο σχετίζεται με το μέγεθος της παραγώγου της συνάρτησης σφάλματος. Έτσι, όταν ικανοποιείται το κριτήριο τερματισμού $\|\nabla E(w^k)\| \leq 10^{-3}$, το w^k θεωρείται ως ένα τοπικό ελάχιστο $\bar{w}_i$ της συνάρτησης σφάλματος E .

Στις δοκιμές μας δεν κάναμε καμία προσπάθεια για την ρύθμιση των παραμέτρων των μεθόδων. Αντιθέτως, χρησιμοποιήσαμε σταθερές τιμές. Στα πειράματα της PSO που πα-

ραθέτουμε παρακάτω, οι τιμές των παραμέτρων γ_1, γ_2 και μ ήταν για όλες τις δοκιμές $\gamma_1 = 10000, \gamma_2 = 1$ και $\mu = 10^{-10}$. Ο αριθμός των χρωμοσωμάτων που αποτελούν τον πληθυσμό για τον Γενετικό Αλγόριθμο (GA), επιλέχθηκε να ισούται με δύο φορές τη διάσταση του προβλήματος (ισχύουν και εδώ τα σχόλια της Υποενότητας 5.2.3 του Κεφαλαίου 5 για το μέγεθος του πληθυσμού).

Η εξισορρόπηση μεταξύ της ολικής και τοπικής ανίχνευσης του χώρου των βαρών από τις μεθόδους PSO και SPSO βασικά ελέγχεται από το βάρος αδράνειας. Ακολουθώντας τον συλλογισμό που αναπτύχθηκε στην Υποενότητα 6.2.3, επιλέξαμε να μεταβάλλουμε (μειώνουμε) την τιμή του βάρους αδράνειας με το χρόνο. Έτσι δώσαμε την αρχική τιμή 1 και σταδιακά που σταδιακά μειωνόταν έως την τιμή 0.4. Η στρατηγική αυτή αποδείχθηκε στην πράξη να έχει πολύ καλύτερα αποτελέσματα από την χρήση μια σταθερή τιμή για το βάρος αδράνειας, αφού οι σχετικά μεγάλη αρχική τιμή επιτρέπει την ευρεία εξερεύνηση του χώρου των βαρών, ενώ η σταδιακή μείωσή της βοηθά στην καλύτερη εξερεύνηση μικρότερων περιοχών.

Κάθε επανάληψη των μεθόδων BP, BPM, BBP, SA, BPSA και BPD αντιστοιχεί σε ένα υπολογισμό της τιμής και ένα υπολογισμό του διανύσματος των μερικών παραγώγων της συνάρτησης σφάλματος, σε αντίθεση με τις μεθόδους BPVS, NMBPM, NMBBP και NMBPVS όπου, γενικά, ο αριθμός των υπολογισμών της τιμής της συνάρτησης σφάλματος (Function Evaluations – FE) είναι μεγαλύτερος από τον αριθμό των υπολογισμών του διανυσμάτων των μερικών παραγώγων της (Gradient Evaluations – GE), λόγω της ευθύγραμμης (μονοδιάστατης) ανίχνευσης που χρησιμοποιούν. Στον Πίνακα 6.1 παρουσιάζουμε δυο γραμμές με αποτελέσματα για τις μεθόδους αυτές: η πρώτη αναφέρεται στους υπολογισμούς της τιμής της συνάρτησης σφάλματος (FE), ενώ η δεύτερη στους υπολογισμούς του διανύσματος των μερικών παραγώγων (GE). Τέλος, να τονίσουμε ότι ένα από τα βασικά γνωρίσματα των αλγορίθμων GA, PDE, PSO και SPSO είναι ότι απαιτούν *μόνο* τον υπολογισμό των τιμών της συνάρτησης σφάλματος.

Για την ευκολότερη σύγκριση των αλγορίθμων, δοκιμάσαμε και τις μεθόδους οπισθοδρομικής διάδοσης του σφάλματος (BP) [133], οπισθοδρομικής διάδοσης του σφάλματος με ορμή (momentum) [57, 133] (BPM), και οπισθοδρομικής διάδοσης του σφάλματος με προσαρμοστικό ρυθμό εκπαίδευσης και ορμή [159] (ABP). Τέλος, στον Πίνακα των αποτελεσμάτων έχουμε συμπεριλάβει και αποτελέσματα από τις μη μονότονες μεθόδους (NMBPM, NMBBP και NMBPVS) που παρουσιάζονται στο Κεφάλαιο 8 και από τους Παράλληλους Διαφοροεξελικτικούς αλγόριθμους ($PDE_1, PDE_2, PDE_3, PDE_4, PDE_5$ και PDE_6) που παρουσιάζονται στο Κεφάλαιο 5.

1) *Το πρόβλημα του αποκλειστικού-ΕΙΤΕ*. Το κλασικό πρόβλημα του αποκλειστικού-ΕΙΤΕ [133, 146] (βλ. Παράρτημα Α.1) αποτελεί την πρώτη δοκιμή μας, γιατί παρουσιάζει ευαισθησία στις αρχικές τιμές των βαρών και μεγάλο αριθμό τοπικών ελαχίστων [17]. Ο ρυθμός εκπαίδευσης που χρησιμοποιήσαμε ήταν ίσος με 1.5 και οι παράμετροι των μεθόδων SA, BPSA και PSO είχαν τις τιμές $n = 0.3, d = 0.002$ και $c_1 = c_2 = 0.5$. Για όλες τις μεθόδους έγιναν 100 πειράματα και τα αποτελέσματα συνοψίζονται στον Πίνακα 6.1.

2) *Το πρόβλημα της ισοτιμίας 3-bit*. Το δεύτερο πρόβλημα που δοκιμάσαμε ήταν το πρόβλημα της ισοτιμίας 3-bit [133] (βλ. Παράρτημα Α.2). Είναι γνωστό ότι στο πρόβλημα αυτό ο χώρος των βαρών παρουσιάζει πολλά τοπικά ελάχιστα καθώς και μεγάλες επίπεδες περιοχές. Ο ρυθμός εκπαίδευσης επιλέχθηκε να είναι ίσος με 0.5 και οι παράμετροι των μεθόδων SA, BPSA και PSO είχαν τις τιμές $n = 0.1, d = 0.00025, c_1 = 0.1$ και $c_2 = 1$. Τα αποτελέσματα από 100 ανεξάρτητες δοκιμές για κάθε αλγόριθμο παρουσιάζονται στον Πίνακα 6.1.

6.5 Συμπεράσματα – Συνεισφορά

Από τη μελέτη των αποτελεσμάτων του Πίνακα 6.1 είναι φανερό ότι ο συνδυασμός μεθόδων ολικής και τοπικής ανίχνευσης, όπως οι μέθοδοι BPSA και BPD, παρέχει μεγαλύτερο

Πίνακας 6.1: Συγκριτικά αποτελέσματα

Αλγόριθμος Εκπαίδευσης	Πρόβλημα Αποκλειστικού-ΕΙΤΕ			Πρόβλημα Ισοτιμίας 3-bit		
	μ	σ	Επιτυχία	μ	σ	Επιτυχία
BP	144.1	112.6	42%	932.0	1320.8	91%
BPM	249.7	322.1	49%	219.9	198.9	93%
BBP	93.3	201.5	71%	150.3	137.3	94%
NMBPM	(FE) 260.4	287.8	68%	244.3	205.9	99%
	(GE) 254.4	287.3		235.1	204.4	
NMBBP	(FE) 191.6	328.9	80%	106.6	123.1	99%
	(GE) 102.1	173.4		99.2	164.5	
BPVS	(FE) 199.1	373.1	78%	105.8	186.9	98%
	(GE) 185.2	343.3		100.4	171.6	
NMBPVS	(FE) 208.4	395.2	80%	102.1	109.9	99%
	(GE) 201.3	378.8		95.3	183.5	
SA	424.2	420.8	43%	805.4	2103.1	22%
BPSA	1661.9	2775.7	65%	2634.0	6866.8	66%
GA	422.3	397.5	95%	1091.5	766.2	73%
PDE ₁	238.4	136.3	100%	1272.9	619.1	82%
PDE ₂	720.1	352.6	100%	3562.7	1367.8	86%
PDE ₃	342.2	186.1	100%	1473.0	873.3	91%
PDE ₄	395.6	218.5	100%	2227.3	903.5	99%
PDE ₅	1209.7	661.5	100%	4829.6	1598.2	91%
PDE ₆	402.4	248.4	100%	2465.8	1011.3	100%
PSO	1459.7	1143.1	77%	6422.4	2992.1	42%
SPSO	7869.6	13905.4	100%	9803.6	5436.6	95%
BPD	575.1	387.3	100%	760.0	696.4	100%

ποσοστό επιτυχίας σε σύγκριση με την κλασική μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος (BP). Η απόδοση της BPSA είναι σαφώς καλύτερη από την απόδοση της απλής SA, αλλά δεν είναι τόσο καλή όσο αναμενόταν αν και η μέθοδος χρησιμοποιεί και πληροφορίες σχετικά με την παράγωγο της συνάρτησης σφάλματος. Αντίθετα, η μέθοδος BPD κατάφερε να αποφύγει τα ανεπιθύμητα τοπικά ελάχιστα σε όλες τις δοκιμές μας και συνεπώς είχε ποσοστό επιτυχίας 100%. Τα αποτελέσματα δείχνουν ότι οι ΓΑ (GA) και οι ΕΑ (PDE) είναι αποτελεσματικοί και αποδοτικοί, ακόμα και όταν συγκριθούν με μεθόδους που εκτός από την τιμή της συνάρτησης σφάλματος απαιτούν και πληροφορίες για την κλίση της. Έτσι βλέπουμε ότι η μέθοδος GA και οι μέθοδοι PDE επέδειξαν πολύ καλή απόδοση στα προβλήματα που δοκιμάσαμε. Τέλος, η μέθοδος PSO συνεπικουρούμενη από την τεχνική του «εφελκυσμού» της συνάρτησης σφάλματος (SPSO) είχε αυξημένο ποσοστό επιτυχίας συγκρινόμενη με την κλασική μέθοδο PSO, αν και απαιτήθηκαν περισσότερες επαναλήψεις για να συγκλίνει.

Εν κατακλείδι, σε αυτό το κεφάλαιο παρουσιάστηκαν και συγκρίθηκαν νέες μέθοδοι ολικής ανίχνευσης, οι οποίες συχνά καταφέρνουν να αποφύγουν την σύγκλιση σε ανεπιθύμητα τοπικά ελάχιστα, κατά την διαδικασία εκπαίδευσης ΤΝΔ. Η αποφυγή τέτοιων ελαχίστων, στην πράξη, δεν είναι πάντα εφικτή, αλλά οι μέθοδοι αυτές έχουν μεγαλύτερη πιθανότητα να καταλήξουν σε μια αποδεκτή λύση και κατά αυτή την έννοια βελτιώνουν την διαδικασία εκπαίδευσης. Επίσης αναλύσαμε και δύο μετασχηματισμούς της συνάρτησης σφάλματος, που έχουν σαν σκοπό την απαλοιφή τοπικών ελαχίστων της. Η τεχνική της παρεκκλίνουσας τροχιάς και η τεχνική του «εφελκυσμού» της αντικειμενικής συνάρτησης παρέχουν σταθερή σύγκλιση και πολύ μεγάλη πιθανότητα επιτυχίας. Τα πειράματά μας δείχνουν ότι αλγό-

ριθμοί εκπαίδευσης που βοηθούνται από τους μετασχηματισμούς αυτούς είναι δυνατόν να ανακαλύπτουν επιθυμητά ελάχιστα με πολύ μεγαλύτερη πιθανότητα. Γενικά, τα αποτελέσματα από δύο προβλήματα γνωστά για τα τοπικά τους ελάχιστα είναι ενθαρρυντικά και δείχνουν τη χρησιμότητα και εφαρμοσιμότητα όλων των τεχνικών που παρουσιάσαμε.

Μέρος IV

Μέθοδοι μη Μονότονης Εκπαίδευσης ΤΝΔ

Εκπαίδευση ανά Πρότυπο Εισόδου

Είναι η φύση του προβλήματος τέτοια που η κάθε μέθοδος γίνεται σχοινοτενής καθώς οι αριθμοί γίνονται μεγαλύτεροι.

—*Karl Friedrich Gauss (1777–1855)*

Σε αυτό το κεφάλαιο θα ασχοληθούμε με την εκπαίδευση ΤΝΔ ανά πρότυπο εισόδου (online training). Αυτή η προσέγγιση στην εκπαίδευση ΤΝΔ θεωρείται κατεξοχήν κατάλληλη για μεγάλα σύνολα προτύπων ή/και δίκτυα, όπου η εκπαίδευση ανά ομάδα προτύπων εισόδου θα απαιτούσε περισσότερο χρόνο και μεγαλύτερο αποθηκευτικό χώρο. Επιπρόσθετα, βοηθά στην αποφυγή των τοπικών ελάχιστων και παρέχει μια πιο φυσική προσέγγιση για την εκπαίδευση μη στατικών προβλημάτων, δηλαδή προβλημάτων που το σύνολο των προτύπων μεταβάλλεται με τον χρόνο.

Στη συνέχεια θα παρουσιάσουμε μια μέθοδο για την αυτόματη προσαρμογή ενός κοινού ρυθμού εκπαίδευσης για όλα τα βάρη, που χρησιμοποιείται στην εκπαίδευση ανά πρότυπο εισόδου. Η προτεινόμενη τεχνική αποτελεί τροποποίηση της στοχαστικής μεθόδου της πιο απότομης καθόδου (stochastic gradient descent), σε συνδυασμό με την μέθοδο που προτείνεται στην εργασία [3]. Η νέα πτυχή της προσέγγισής μας συνίσταται στο γεγονός ότι λαμβάνει υπόψη ήδη υπολογισμένες πληροφορίες για να επιτύχει την καλύτερη προσαρμογή του κοινού ρυθμού εκπαίδευσης.

Ο προτεινόμενος αλγόριθμος έχει εφαρμοστεί, εξεταστεί και συγκριθεί με άλλες μεθόδους εκπαίδευσης ανά πρότυπο εισόδου, αλλά και μεθόδους εκπαίδευσης ανά ομάδα προτύπων εισόδου (batch training). Τα αποτελέσματα που παρουσιάζουμε δείχνουν ότι συμπεριφέρεται προβλέψιμα, είναι αποδοτικός και ότι έχει αρκετά ικανοποιητική μέση απόδοση [77, 78, 108, 110].

7.1 Εισαγωγή

Στο ερευνητικό πεδίο των ΤΝΔ, τα δίκτυα που η εκπαίδευσή τους βασίζεται στην μέθοδο της οπισθοδρομικής διάδοσης του σφάλματος (Backpropagation) [133] είναι τα πιο δημοφιλή. Αποδοτικοί αλγόριθμοι εκπαίδευσης τέτοιων ΤΝΔ είναι ένα βασικό θέμα της επίκαιρης έρευνας και για αυτόν τον σκοπό έχουν προταθεί πολυάριθμοι αλγόριθμοι.

Μια κοινή προσέγγιση εκπαίδευσης είναι η ελαχιστοποίηση του σφάλματος των δικτύων, το οποίο αποτελεί ένα μέτρο της απόδοσής τους. Το σφάλμα αυτό υπολογίζεται συνήθως ως η διαφορά μεταξύ του πραγματικού διανύσματος εξόδου του δικτύου και του επιθυμητού διανύσματος εξόδου (εκπαίδευση με επίβλεψη). Ο γρήγορος υπολογισμός ενός συνόλου βαρών που ελαχιστοποιεί αυτό το σφάλμα είναι μια μάλλον δύσκολη εργασία λόγω του γεγονότος ότι, γενικά, ο αριθμός των βαρών του δικτύων είναι ιδιαίτερα μεγάλος και η συνάρτηση σφάλματος παράγει μια περίπλοκη επιφάνεια. Η συνάρτηση σφάλματος συνήθως έχει πολλά

τοπικά ελάχιστα και μεγάλες επιπέδες περιοχές (flat regions), που ακολουθούνται από πολύ στενές περιοχές με μεγάλη κλίση.

Οι αλγόριθμοι εκπαίδευσης με επίβλεψη μπορούν να διαιρεθούν σε δύο βασικές κατηγορίες:

- αλγόριθμοι εκπαίδευσης ανά πρότυπο εισόδου (on-line ή stochastic training), και
- αλγόριθμοι εκπαίδευσης ανά ομάδα προτύπων εισόδου (batch ή off-line training).

Η εκπαίδευση ανά ομάδα προτύπων εισόδου είναι η κλασική προσέγγιση, όπου ένα σύνολο προτύπων λαμβάνεται και χρησιμοποιείται προκειμένου να εκπαιδευτεί το ΤΝΔ, προτού αυτό χρησιμοποιηθεί σε κάποια εφαρμογή. Αντίθετα, στην εκπαίδευση ανά πρότυπο εισόδου τα στοιχεία που συγκεντρώνονται κατά τη διάρκεια της κανονικής λειτουργίας του συστήματος χρησιμοποιούνται για την συνεχή εκπαίδευση και προσαρμογή του ΤΝΔ.

Η εκπαίδευση ανά ομάδα προτύπων εισόδου αντιμετωπίζεται από τη θεωρία της βελτιστοποίησης χωρίς περιορισμούς, αφού η πραγματική κλίση της συνάρτησης σφάλματος και πληροφορίες από ολόκληρο το σύνολο χρησιμοποιούνται. Όπως είδαμε και στο Κεφάλαιο 2, η εκπαίδευση ανά ομάδα προτύπων εισόδου αντιμετωπίζεται με την ελαχιστοποίηση της συνάρτησης σφάλματος E που ορίζεται ως το άθροισμα, για όλα τα πρότυπα εισόδου, των τετραγώνων των διαφορών της πραγματικής εξόδου του ΤΝΔ και την επιθυμητή έξοδο:

$$E(w) = \frac{1}{2} \sum_{p=1}^P \sum_{j=1}^{N_L} (y_{j,p}^L - t_{j,p})^2 = \sum_{p=1}^P E_p, \quad (7.1)$$

όπου P είναι ο συνολικός αριθμός προτύπων, $y_{j,p}^L$ η έξοδος του j νευρώνα που ανήκει στο L στρώμα, N_L ο αριθμός των νευρώνων του στρώματος εξόδου, $t_{j,p}$ η επιθυμητή έξοδος του j νευρώνα εξόδου στο πρότυπο p , και E_p το σφάλμα του δικτύου ανά πρότυπο εισόδου. Στην εκπαίδευση ανά πρότυπο εισόδου τα βάρη του δικτύου ενημερώνονται μετά από την παρουσίαση κάθε πρότυπο εισόδου, το οποίο μπορεί να επιλεγεί από το σύνολο εκπαίδευσης με ή χωρίς επανατοποθέτηση. Αυτό αντιστοιχεί στην ελαχιστοποίηση του στιγμιαίου σφάλματος E_p του ΤΝΔ, όπως φαίνεται στη Σχέση (7.1).

Η εκπαίδευση ανά πρότυπο εισόδου μπορεί να επιλεγεί για προβλήματα που έχουν είτε πολύ μεγάλο αριθμό προτύπων (πιθανά και κάποιο αριθμό περιττών ή λανθασμένων προτύπων), είτε όταν προσπαθούμε να προσεγγίσουμε ένα αργά μεταβαλλόμενο σύστημα. Αν και η εκπαίδευση ανά ομάδα προτύπων εισόδου φαίνεται να είναι ταχύτερη για μικρά σύνολα προτύπων και δίκτυα, η εκπαίδευση ανά πρότυπο εισόδου είναι σαφώς γρηγορότερη για μεγάλα σύνολα προτύπων και ΤΝΔ με πολλά βάρη (πολλές χιλιάδες βάρη και πολώσεις), βοηθά σημαντικά στην αποφυγή τοπικών ελαχίστων και παρέχει μια φυσική προσέγγιση για την εκπαίδευση ΤΝΔ σε μη στατικά προβλήματα, δηλαδή προβλήματα που το σύνολο προτύπων μεταβάλλεται αργά με τον χρόνο.

Λαμβάνοντας υπόψη την έμφυτη αποδοτικότητα της στοχαστικής μεθόδου της πιο απότομης καθόδου, έχουν προταθεί πρόσφατα διάφοροι αλγόριθμοι [3, 136, 140, 141, 154]. Δυστυχώς, η εκπαίδευση ανά πρότυπο εισόδου πάσχει από κάποια μειονεκτήματα όπως για παράδειγμα η ευαισθησία στις παραμέτρων εκπαίδευσης [136]. Ένα άλλο μειονέκτημα είναι ότι πιο προηγμένες μέθοδοι βελτιστοποίησης, όπως οι συζυγείς μέθοδοι κλίσης (conjugate gradient), οι μέθοδοι μεταβλητής μετρικής (variable metric), η μέθοδος της προσομοιωμένης ανόπτησης (simulated annealing) κ.α., μπορούν να εφαρμοστούν μόνο σε σταθερές επιφάνειες σφάλματος, και έτσι υπάρχει δυσκολία στη χρησιμοποίησή τους στην εκπαίδευση ανά πρότυπο εισόδου [136].

Εντούτοις, η εκπαίδευση ανά πρότυπο εισόδου έχει πολλά πλεονεκτήματα σε σχέση με την εκπαίδευση ανά ομάδα προτύπων εισόδου. Οι μέθοδοι εκπαίδευσης ανά πρότυπο εισόδου παρουσιάζουν σθεναρότητα ως προς λανθασμένα και περιττά πρότυπα, και έτσι παραλείψεις στο σύνολο προτύπων εκπαίδευσης μπορούν να διορθωθούν κατά τη διάρκεια της

εκπαίδευσης του ΤΝΔ. Επιπλέον, τα πρότυπα εκπαίδευσης μπορούν συχνά να παραχθούν εύκολα και σε μεγάλες ποσότητες όταν το σύστημα είναι σε λειτουργία, ενώ είναι συνήθως λιγυστά πριν την αρχή της εκπαίδευσης. Γενικά, η εκπαίδευση ανά πρότυπο εισόδου είναι απαραίτητη εάν χρειαζόμαστε συστήματα με δυνατότητα εκμάθησης, σε αντιδιαστολή με τα απλά εκπαιδευμένα συστήματα [155].

7.2 Αλγόριθμοι Εκπαίδευσης ανά Πρότυπο Εισόδου

Παρά την αφθονία μεθόδων για εκπαίδευση ΤΝΔ, υπάρχουν μόνο λίγες μέθοδοι που μπορούν να χρησιμοποιηθούν αποδοτικά για εκπαίδευση ανά πρότυπο εισόδου. Για παράδειγμα, οι κλασικοί αλγόριθμοι εκπαίδευσης ανά ομάδα προτύπων εισόδου δεν μπορούν απευθείας να χειριστούν μη στατικά σύνολα προτύπων. Ακόμα και όταν μερικοί από αυτούς χρησιμοποιούνται στην εκπαίδευση ανά πρότυπο εισόδου, συχνά παρουσιάζεται το πρόβλημα της «καταστροφικής παρέμβασης», δηλαδή η εκπαίδευση του ΤΝΔ στα νέα πρότυπα παρεμποδίζεται υπερβολικά από τα προηγούμενα, οδηγώντας σε κορεσμό και την αργή σύγκλιση [155].

Οι μέθοδοι που μπορούν να χρησιμοποιηθούν για εκπαίδευση ανά πρότυπο εισόδου, είναι εκείνες που μπορούν με επιτυχία να χειριστούν μεταβλητά με το χρόνο σύνολα προτύπων (time-varying training sets), ενώ συγχρόνως απαιτούν σχετικά λίγους πρόσθετους υπολογιστικούς πόρους (μνήμη και επεξεργαστική ισχύ) προκειμένου να υποβληθεί σε επεξεργασία κάθε πρόσθετο πρότυπο. Παραδείγματα τέτοιων μεθόδων είναι οι παραλλαγές της στοχαστικής μεθόδου της πιο απότομης καθόδου [3].

Η πρώτη μέθοδος που θα μελετήσουμε, ονομάζεται ALAP₁, και χρησιμοποιεί σε κάθε επανάληψη ένα κοινό ρυθμό εκπαίδευσης για όλα τα βάρη:

$$\eta_i^k = \eta_i^{k-1} + \gamma \langle \nabla E_{p-1}(w^{k-1}), \nabla E_p(w^k) \rangle, \quad (7.2)$$

όπου $\eta_i^0 = c$ για όλα τα βάρη του δικτύου (το c είναι μια μικρή θετική σταθερά), και $\langle \cdot, \cdot \rangle$ συμβολίζουμε το εσωτερικό γινόμενο στον $\mathbb{R}^n$.

Οι άλλες δύο μέθοδοι που περιγράφονται στην εργασία [3] (ALAP₂ και ALAP₃), χρησιμοποιούν διαφορετικό ρυθμό εκπαίδευσης για κάθε βάρος. Αυτό το χαρακτηριστικό γνώρισμα καθιστά αυτούς τους αλγόριθμους εκπαίδευσης ανά πρότυπο εισόδου ικανούς να χρησιμοποιούν παραλλαγές της κατεύθυνσης της πιο απότομης καθόδου και να κινούνται κατά μήκος μιας κατεύθυνσης που δεν συμπίπτει απαραίτητως με αυτή, με αποτέλεσμα συχνά να επιταχύνεται η διαδικασία ελαχιστοποίησης.

Ο αλγόριθμος ALAP₂ χρησιμοποιεί τον ακόλουθο τύπο για την προσαρμογή του ρυθμού εκπαίδευσης:

$$\eta_i^k = \eta_i^{k-1} \left[1 + \gamma \partial_i E_{p-1}(w^{k-1}) \partial_i E_p(w^k) \right]. \quad (7.3)$$

Ο αλγόριθμος ALAP₃ χρησιμοποιεί μια κανονικοποιημένη έκδοση του κανόνα προσαρμογής του ALAP₂, που δίνεται από τον τύπο:

$$\eta_i^k = \eta_i^{k-1} \left[1 + \gamma \frac{\partial_i E_{p-1}(w^{k-1}) \partial_i E_p(w^k)}{u_i^k} \right], \quad (7.4)$$

όπου u_i^k αποτελεί τον εκθετικό μέσο όρο του τετραγώνου της μερικής παραγώγου του στιγμιαίου σφάλματος, $\partial_i E_p(w^k)$, που δίνεται από τον τύπο:

$$u_i^k = \mu u_i^{k-1} + (1 - \mu) \left[\partial_i E_p(w^k) \right]^2,$$

όπου μ και γ είναι θετικές σταθερές. Ενδεικτικά, οι τιμές $\mu = 0.9$ και $\gamma = 0.01$ είναι

κατάλληλες για τις περισσότερες των περιπτώσεων [3].

Στη συνέχεια προτείνουμε τον παρακάτω νέο αλγόριθμο για την προσαρμογή ενός κοινού ρυθμού εκπαίδευσης για όλα τα βάρη, στα πλαίσια της στοχαστικής μεθόδου της πιο απότομης καθόδου:

$$\begin{aligned}\eta^{k+1} = \eta^k + \gamma_1 \langle \nabla E_{p-1}(w^{k-1}), \nabla E_p(w^k) \rangle \\ + \gamma_2 \langle \nabla E_{p-2}(w^{k-2}), \nabla E_{p-1}(w^{k-1}) \rangle.\end{aligned}\quad (7.5)$$

Το βασικό γνώρισμα του προτεινόμενου αλγόριθμου είναι ότι αξιοποιεί πληροφορίες σχετικές με την κλίση, από την τρέχουσα καθώς επίσης και τις δύο προηγούμενες παρουσιάσεις προτύπων [77, 78, 110].

Αυτό φαίνεται να παρέχει κάποια σταθεροποίηση στην προσαρμογή των τιμών του ρυθμού εκπαίδευσης και βοηθά την στοχαστική μέθοδο της πιο απότομης καθόδου να επιτύχει γρήγορη σύγκλιση και υψηλό ποσοστό επιτυχίας. Μια αλγοριθμική περιγραφή του προτεινόμενου αλγορίθμου δίνεται στον Αλγόριθμο 7.1.

Αλγόριθμος 7.1: Ο προτεινόμενος αλγόριθμος σε ψευδοκώδικα.

ΣΤΟΧΑΣΤΙΚΗ ΜΕΘΟΔΟΣ ΤΗΣ ΠΙΟ ΑΠΟΤΟΜΗΣ ΚΑΘΟΔΟΥ ΜΕ ΜΕΤΑΒΛΗΤΟ ΡΥΘΜΟ ΕΚΠΑΙΔΕΥΣΗΣ

- 0: Αρχικοποίηση w^0 , η^0 , γ_1 , και γ_2 .
 - 1: **Repeat**
 - 2: Θέσε $k = k + 1$.
 - 3: Επέλεξε τυχαία το p πρότυπο από το σύνολο προτύπων.
 - 4: Υπολόγισε $E_p(w^k)$ και μετά $\nabla E_p(w^k)$.
 - 5: Υπολόγισε τα νέα βάρη:
 $w^{k+1} = w^k - \eta^k \nabla E_p(w^k)$.
 - 6: Υπολόγισε το νέο ρυθμό εκπαίδευσης:
 $\eta^{k+1} = \eta^k + \gamma_1 \langle \nabla E_{p-1}(w^{k-1}), \nabla E_p(w^k) \rangle$
 $+ \gamma_2 \langle \nabla E_{p-2}(w^{k-2}), \nabla E_{p-1}(w^{k-1}) \rangle$.
 - 7: **Until** Συνθήκη Τερματισμού.
 - 8: **Return** τα τελικά βάρη w^{k+1} .
-

Σε αυτό το αλγοριθμικό μοντέλο, η είναι ο ρυθμός εκπαίδευσης, και γ_1 και γ_2 είναι μετα-ρυθμοί εκπαίδευσης (meta-learning rates). Σαν συνθήκη τερματισμού μπορεί να χρησιμοποιηθεί το σφάλμα ταξινόμησης ή ένα άνω όριο στους υπολογισμούς της συνάρτησης σφάλματος.

7.3 Προσομοιώσεις και Αποτελέσματα

Σε αυτή την υποενότητα παρουσιάζουμε τα αποτελέσματα του προτεινόμενου αλγόριθμου εκπαίδευσης ανά πρότυπο εισόδου σε γνωστά προβλήματα εκπαίδευση ΤΝΔ. Για τον λόγο αυτό, ο Αλγόριθμος 7.1 έχει αξιολογηθεί και έχει συγκριθεί με στοχαστικές μεθόδους καθώς επίσης και με μεθόδους εκπαίδευσης ανά ομάδα προτύπων εισόδου. Πιο συγκεκριμένα, στις προσομοιώσεις έχουμε συγκρίνει τον Αλγόριθμο 7.1 με τις 3 στοχαστικές μεθόδους προσαρμογής του ρυθμού εκπαίδευσης που έχουν προταθεί από τους Almeida, Langlois, Amaral και Plankhron στην εργασία [3] (ALAP₁, ALAP₂ και ALAP₃), και την μέθοδο οπισθοδρομικής διάδοσης του σφάλματος για εκπαίδευση ανά πρότυπο εισόδου (On-line BP) [133]. Επιπλέον, για τις συγκρίσεις έχουμε εξετάσει και τις μεθόδους εκπαίδευσης ανά ομάδα προτύπων εισόδου: την μέθοδο οπισθοδρομικής διάδοσης του σφάλματος για εκπαίδευση ανά ομάδα προτύπων εισόδου [133] (Batch BP) και την μέθοδο οπισθοδρομικής διάδοσης του σφάλματος με προσαρμοστικό ρυθμό εκπαίδευσης και ορμή (momentum) για εκπαίδευση ανά ομάδα προτύπων εισόδου [159] (Batch ABP).

Οι αλγόριθμοι εξετάστηκαν χρησιμοποιώντας τα ίδια αρχικά βάρη, που αρχικοποιήθηκαν με τη μέθοδο των Nguyen-Widrow [94], και έλαβαν την ίδια ακολουθία προτύπων εισόδου. Για κάθε πρόβλημα που περιγράφεται κατωτέρω, παρουσιάζουμε έναν πίνακα που συνοψίζει την απόδοση των αλγορίθμων για τις προσομοιώσεις που έφθασαν σε σύγκλιση. Η εκπαίδευση θεωρείται επιτυχής όταν το ΤΝΔ δεν παρουσιάζει λάθη ταξινόμησης στο σύνολο εκπαίδευσης.

Οι αναφερόμενες στους παρακάτω πίνακες παράμετροι είναι: Min ο ελάχιστος αριθμός παρουσιάσεων προτύπων, μ η μέση τιμή των παρουσιάσεων προτύπων, Max η ανώτατη τιμή παρουσιάσεων προτύπων, και Επιτυχία (%) ο αριθμός των επιτυχημένων προσομοιώσεων σε σύνολο 100 δοκιμών. Εάν ένας αλγόριθμος αποτύχει να συγκλίνει μέσα σε ένα προκαθορισμένο όριο υπολογισμών της συνάρτησης σφάλματος, θεωρείται ότι αποτυγχάνει να εκπαιδεύσει το ΤΝΔ, και οι παρουσιάσεις προτύπων κατά την δοκιμή αυτή δεν συμπεριλαμβάνονται στη στατιστική ανάλυση των αποτελεσμάτων.

Οι τιμές των παραμέτρων γ_1 και γ_2 επιλέχθηκαν να είναι $\gamma_1 \ll \gamma_2 = 1$. Φαίνεται ότι η επιλογή τιμών για τους μετά-ρυθμούς εκπαίδευσης γ_1 και γ_2 δεν είναι κρίσιμη για την επιτυχή εκπαίδευση. Εντούτοις, μπορεί να επιτευχθεί γρηγορότερη σύγκλιση, εάν γίνει μια ακριβής προσαρμογή τους ανάλογα με το πρόβλημα. Τέτοια προσαρμογή δεν κρίθηκε αναγκαία στα πειράματά μας, γιατί βασικός σκοπός μας ήταν η συγκριτική μελέτη των μεθόδων και όχι η βέλτιστη απόδοσή τους. Αφ' ετέρου, μεγάλη προσπάθεια έχει γίνει για να προσδιοριστούν κατάλληλες τιμές για τις ευρετικές παραμέτρους των μεθόδων εκπαίδευσης ανά ομάδα προτύπων BP και ABP. Η εμπειρία μας με προσομοιώσεις των μεθόδων αυτών δείχνει ότι είναι χαρακτηριστική η συμπεριφορά των αλγορίθμων που περιγράφεται στα παραδείγματα που ακολουθούν.

7.3.1 Αποκλειστικό-ΕΙΤΕ

Το πρόβλημα του Αποκλειστικού-ΕΙΤΕ (βλ. και Παράρτημα Α.1) θεωρείται ως ένα κλασικό πρόβλημα εκπαίδευσης ΤΝΔ. Στην περίπτωσή μας έχει χρησιμοποιηθεί ένα 2-2-1 ΤΝΔ (6 βάρη και 3 πολώσεις). Το δίκτυο είναι βασισμένο σε νευρώνες με λογιστικές συναρτήσεις ενεργοποιήσεις. Το όριο υπολογισμών της συνάρτησης σφάλματος ήταν 4000, δηλαδή επιτράπηκαν μόνο 4000 παρουσιάσεις προτύπων. Τα συγκριτικά αποτελέσματα εμφανίζονται στον Πίνακα 7.1.

Από τα αποτελέσματα είναι εμφανές ότι ο προτεινόμενος αλγόριθμος ξεπερνά σαφώς τους αλγόριθμους $ALAP_1$, $ALAP_2$ και $ALAP_3$, και τις δύο εκδοχές της μεθόδου οπισθοδρομικής διάδοσης του σφάλματος (Batch BP και On-line BP), αλλά η μέθοδος ABP έχει ελαφρώς υψηλότερο ποσοστό επιτυχίας. Αυτό αναμενόταν δεδομένου ότι, γενικά, οι αλγόριθμοι εκπαίδευσης ανά ομάδα προτύπων εισόδου είναι πολύ καλοί σε προβλήματα που έχουν μικρά σύνολα προτύπων ή/και μικρές τοπολογίες δικτύων, αλλά είναι πιο αργοί από τις μεθόδους εκπαίδευσης ανά πρότυπο εισόδου.

Πίνακας 7.1: Αποτελέσματα από το πρόβλημα του αποκλειστικού ΕΙΤΕ

Αλγόριθμος	Min	μ	Max	Επιτυχία (%)
Batch BP	176	1693.9	3840	17%
Batch ABP	144	1430.4	3708	49%
On-line BP	72	724.2	2972	43%
$ALAP_1$	56	736.1	3900	38%
$ALAP_2$	40	816.9	3960	37%
$ALAP_3$	52	1000.5	3636	43%
Αλγόριθμος 7.1	44	680.2	3388	48%

7.3.2 Αναγνώριση αριθμών

Στο δεύτερο πείραμα, ένα δίκτυο με 64 νευρώνες εισόδου, 6 κρυφούς νευρώνες και 10 νευρώνες εξόδου (444 βάρη και 16 πολώσεις) εκπαιδεύεται για να αναγνωρίσει τους αριθμούς από το 0 έως το 9, εκτυπωμένους με πλάγιους χαρακτήρες (*italics*) και μέγεθος 8×8 εικονοκύτταρα (pixels) [82] (βλ. και Παράρτημα Α.7). Το δίκτυο είναι βασισμένο σε νευρώνες με την λογιστική συνάρτηση ενεργοποίησης. Η συνθήκη τερματισμού για όλους τους αλγόριθμους είναι να έχουν μηδέν λάθος ταξινόμησης στο σύνολο εκπαίδευσης, μέσα σε 1000 υπολογισμούς της συνάρτησης σφάλματος. Λεπτομερή αποτελέσματα σχετικά με την απόδοση των αλγορίθμων εκπαίδευσης παρουσιάζονται στον Πίνακα 7.2.

Πίνακας 7.2: Αποτελέσματα από το πρόβλημα αναγνώρισης αριθμών

Αλγόριθμος	<i>Min</i>	μ	<i>Max</i>	Επιτυχία (%)
Batch BP	210	500.8	980	90%
Batch ABP	420	789.2	990	51%
On-line BP	230	507.7	950	99%
ALAP ₁	130	475.5	990	90%
ALAP ₂	190	433.6	860	90%
ALAP ₃	210	486.3	990	96%
Αλγόριθμος 7.1	170	436.3	870	99%

Σχολιάζοντας τα αποτελέσματα βλέπουμε ότι η μέθοδος οπισθοδρομικής διάδοσης του σφάλματος για εκπαίδευση ανά πρότυπο εισόδου (On-line BP) είχε πολύ υψηλό ποσοστό επιτυχίας, αλλά οι μέθοδοι ALAP₁, ALAP₂ και ALAP₃ είχαν γρηγορότερη σύγκλιση. Παρόλα αυτά, η προτεινόμενη μέθοδος και η μέθοδος οπισθοδρομικής διάδοσης του σφάλματος για εκπαίδευση ανά πρότυπο εισόδου είχαν σχεδόν τέλει ποσοστό επιτυχίας (99%). Επιπλέον, ο Αλγόριθμος 7.1 επιδεικνύει πολύ γρήγορη σύγκλιση, δεδομένου ότι χρειάστηκε κατά μέσον όρο μόνο 436 παρουσιάσεις προτύπων εισόδου προκειμένου να εκπαιδεύσει το ΤΝΔ. Συνεπώς, η γενική απόδοση του Αλγορίθμου 7.1 κρίνεται πολύ ικανοποιητική.

7.3.3 Αναγνώριση των κεφαλαίων γραμμάτων

Για το πρόβλημα πρόβλημα αναγνώρισης των κεφαλαίων γραμμάτων (βλ. και Παράρτημα Α.6), 26 πίνακες με τα κεφαλαία γράμματα του αγγλικού αλφάβητου παρουσιάζονται σε ένα 35-30-26 ΤΝΔ (1830 βάρη, 56 πολώσεις). Κάθε γράμμα προσδιορίζεται από δυαδικές τιμές (0 ή 1) σε ένα πλέγμα (grid) του μεγέθους 5×7 . Το ΤΝΔ βασίστηκε σε νευρώνες με λογιστικές συναρτήσεις ενεργοποίησης. Τα αποτελέσματα δίνονται στον Πίνακα 7.3.

Για ακόμα μια φορά, η προτεινόμενη μέθοδος (Αλγόριθμος 7.1) είχε πολύ υψηλό ποσοστό επιτυχίας (96%) και ήταν γρηγορότερη από όλες τις άλλες μεθόδους που εξετάσαμε. Κατά μέσον όρο χρειάστηκε μόνο 749 παρουσιάσεις προτύπων προκειμένου να ολοκληρωθεί επιτυχώς η εκπαίδευση.

7.4 Υβριδικές Μέθοδοι για την Επανεκπαίδευση ΤΝΔ

Σε αυτή την υποενότητα θα παρουσιάσουμε μια νέα υβριδική μέθοδο που βελτιώνει την γενίκευση των ΤΝΔ, σε προβλήματα που το σύνολο προτύπων μεταβάλλεται αργά με τον χρόνο. Αυτή η προσέγγιση συμπίπτει με τον τρόπο που οι Γενετικοί και Εξελικτικοί Αλγόριθμοι επεκτείνονται και εξερευνούν χώρους που μεταβάλλονται αργά [6, 139, 158]. Η νέα μέθοδος αποτελείται από δύο Φάσεις [79, 112]. Στην Φάση 1, ο Αλγόριθμος 7.1 εφαρμόζεται για να

Πίνακας 7.3: Αποτελέσματα από το πρόβλημα αναγνώρισης κεφαλαίων γραμμάτων

Αλγόριθμος	<i>Min</i>	μ	<i>Max</i>	Επιτυχία (%)
Batch BP	4498	21375.9	41860	79%
Batch ABP	3588	3815.7	4212	98%
On-line BP	1404	1861.1	2418	87%
ALAP ₁	494	1519.4	2548	72%
ALAP ₂	338	756.6	1846	94%
ALAP ₃	338	754.5	2418	79%
Αλγόριθμος 7.1	364	749.6	1872	96%

εκπαιδεύσει το ΤΝΔ ανά πρότυπο εισόδου. Μετά το τέλος της εκπαίδευσης, το ΤΝΔ επανεκπαιδεύεται ανά πρότυπο εισόδου χρησιμοποιώντας τους Διαφοροεξελικτικούς Αλγόριθμους (ΔΕΑ), που περιγράφονται στο Κεφάλαιο 5.

Με τον τρόπο αυτό συνδυάζουμε την ταχύτητα της στοχαστικής μεθόδου της πιο απότομης καθόδου, που εφαρμόζεται για την εκπαίδευση ανά πρότυπο εισόδου, με ένα Εξελικτικό Αλγόριθμο, που αποτελεί μέθοδο ολικής βελτιστοποίησης. Με την ιδέα της επανεκπαίδευσης είμαστε σε θέση να συνεχίσουμε την εκπαίδευση ενός ήδη εκπαιδευμένου ΤΝΔ. Ο λόγος για να γίνει αυτό είναι ότι πιθανά νέα πρότυπα είναι διαθέσιμα και επιθυμούμε να εκπαιδεύσουμε το ΤΝΔ και με αυτά, χωρίς όμως να πρέπει να αρχίσουμε την διαδικασία της εκπαίδευσης από την αρχή.

Επισημαίνουμε τέλος ότι η διαδικασία της εκπαίδευσης και επανεκπαίδευσης ανά πρότυπο εισόδου είναι πολύ σημαντική για την επίλυση πολλών πραγματικών προβλημάτων, όπως για παράδειγμα ο έλεγχος της κατεύθυνσης ενός κινούμενου οχήματος ανάλογα με τις συνθήκες που επικρατούν στην πορεία του [10] και η αναγνώριση και εξαγωγή προτύπων από εικόνες που προέρχονται από συσκευές λήψης κάτω από μεταβαλλόμενες συνθήκες αντίληψης (φωτεινότητα, σκιές, συνθήκες φωτισμού και αντανάκλασης) [9, 31, 97, 98, 173].

Ο Αλγόριθμος 7.2 περιγράφει την παραπάνω διαδικασία. Ο νέος αυτός αλγόριθμος δοκιμάστηκε σε δύο δύσκολα προβλήματα ταξινόμησης: το πρόβλημα ταξινόμησης υφής και το πρόβλημα της αναγνώρισης ανωμαλιών σε κολονοσκοπήσεις.

7.4.1 Το πρόβλημα ταξινόμησης υφής

Στο πρόβλημα ταξινόμησης υφής (βλ. και Παράρτημα Α.8), ένα 16–8–12 ΤΝΔ εκπαιδεύεται να αναγνωρίζει 12 εικόνες διαφορετικής υφής. Το δίκτυο βασίζεται σε νευρώνες με λογιστικές συναρτήσεις ενεργοποίησης και τα βάρη αρχικοποιήθηκαν με τυχαίους αριθμούς από το διάστημα $(-1, 1)$. Η συνθήκη τερματισμού της εκπαίδευσης ανά πρότυπο εισόδου ήταν να επιτύχουμε σφάλμα ταξινόμησης $CE \leq 3\%$. Τότε, ξεκινά η δεύτερη Φάση, για την επανεκπαίδευση ανά πρότυπο εισόδου με χρήση ΔΕΑ. Τελικά, η ικανότητα γενίκευσης του ΤΝΔ δοκιμάστηκε σε ένα σύνολο από 320 πρότυπα, που δεν συμμετείχαν στο σύνολο εκπαίδευσης. Το ΤΝΔ αναγνώρισε σωστά 304 από τα 320, δηλαδή είχε γενίκευση 95%. Στο ίδιο πρόβλημα, ο Αλγόριθμος 7.1 χωρίς την χρήση της επανεκπαίδευσης με ΔΕΑ, είχε ποσοστό 93%, ενώ άλλες μέθοδοι εκπαίδευσης ανά ομάδα προτύπων (π.χ. η μέθοδος BPVS [82]) έχουν ποσοστό περίπου 90%.

7.4.2 Το πρόβλημα αναγνώρισης ανωμαλιών σε κολονοσκοπήσεις

Η κολονοσκόπηση (ή κολοσκόπηση) είναι η εξέταση του παχέος εντέρου με εύκαμπο ενδοσκόπιο και χαρακτηρίζεται ως εξέταση ελάχιστης εισβολής (minimal invasive). Το ενδοσκόπιο είναι ένας εύκαμπος λεπτός σωλήνας που στο μπροστινό άκρο του έχει προσαρ-

Αλγόριθμος 7.2: Γενικό μοντέλο του υβριδικού αλγόριθμου**Φάση 1 - «Εκπαίδευση»**

- Βήμα 0α: Αρχικοποίησε w^0, η^0, γ_1 και γ_2 .
 Βήμα 1α: **Repeat** για κάθε πρότυπο p .
 Βήμα 2α: Υπολόγισε $E_p(w^k)$ και μετά $\nabla E_p(w^k)$.
 Βήμα 3α: Υπολόγισε τα νέα βάρη:

$$w^{k+1} = w^k - \eta^k \nabla E_p(w^k).$$

 Βήμα 4α: Υπολόγισε το ρυθμό εκπαίδευσης για το επόμενο πρότυπο $(p + 1)$:

$$\eta^{k+1} = \eta^k + \gamma_1 \langle \nabla E_{p-1}(w^{k-1}), \nabla E_p(w^k) \rangle + \gamma_2 \langle \nabla E_{p-2}(w^{k-2}), \nabla E_{p-1}(w^{k-1}) \rangle.$$

 Βήμα 5α: **Until** Συνθήκη Τερματισμού.
 Βήμα 6α: **Return** τα τελικά βάρη w^{k+1} στη Φάση 2.

Φάση 2 - «Εξέλιξη»

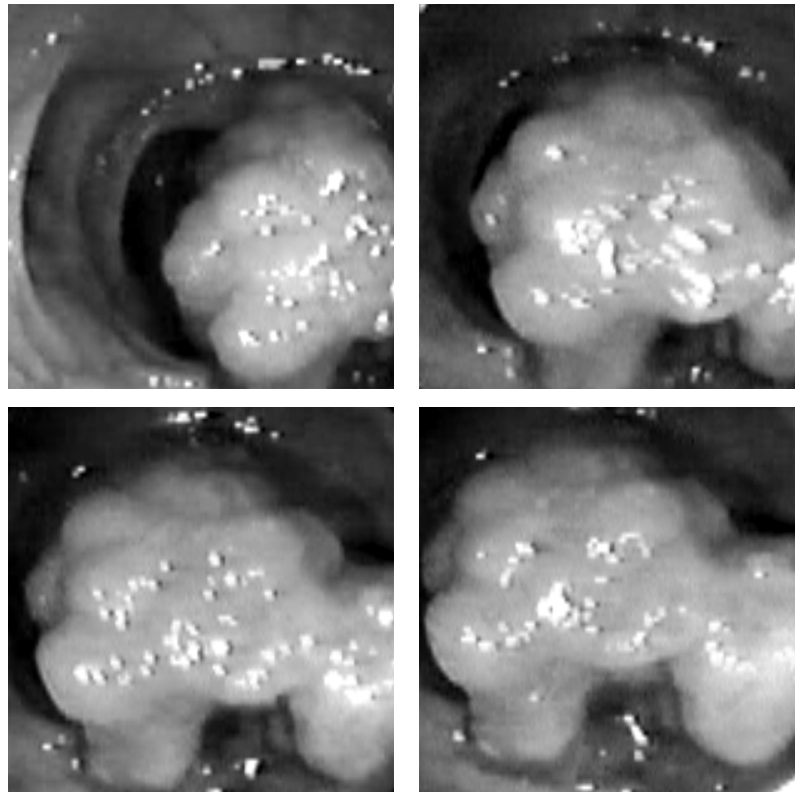
- Βήμα 0β: Αρχικοποίησε τον πληθυσμό στην γειτονιά του w^{k+1} .
 Βήμα 1β: **Repeat** για κάθε πρότυπο p .
 Βήμα 2β: **For** $i = 1$ to NP
 Βήμα 3β: $MUTATION(w_i^k) \rightarrow \text{Mutant_Vector}$.
 Βήμα 4β: $CROSSOVER(\text{Mutant_Vector}) \rightarrow \text{Trial_Vector}$.
 Βήμα 5β: **If** $E_p(\text{Trial_Vector}) \leq E_p(w_i^k)$, αποδοχή του Trial_Vector.
 Βήμα 6β: **EndFor**
 Βήμα 7β: **Until** Συνθήκη Τερματισμού.

μοσμένη μια συσκευή λήψης και μαγνητικής εγγραφής εικόνων (video camera), η οποία στέλνει το σήμα σε μία μικρή τηλεόραση. Το ενδοσκόπιο εισέρχεται στο σώμα του ασθενούς και εικόνες από το παχύ έντερο εμφανίζονται στην οθόνη. Ο γιατρός ελέγχει την κατεύθυνση της συσκευής χρησιμοποιώντας διακόπτες και τροχούς.

Η χρήση ΤΝΔ για την αυτόματη ανακάλυψη κακοηθών όγκων σε εικόνες από ενδοσκόπια παρουσιάζει αρκετές δυσκολίες, καθώς οι εικόνες αυτές έχουν διαφορές στη φωτεινότητα, είναι από διαφορετική οπτική γωνία ανάλογα με τον γιατρό, και υπάρχουν διαφοροποιήσεις στη διάχυση του φωτός. Τέλος, από τη φύση του το πρόβλημα αυτό είναι ένα πρόβλημα που τα δεδομένα αλλάζουν κατά την διάρκεια της εξέτασης. Για να αντιμετωπίσουμε τα παραπάνω προβλήματα, προτείνουμε την εκπαίδευση ανά πρότυπο εισόδου και στην συνέχεια επανεκπαίδευση ανά πρότυπο εισόδου με χρήση ΔΕΑ. Η υβριδική μέθοδος φαίνεται να καταφέρνει να τροποποιεί κατάλληλα τα βάρη του ΤΝΔ, και κατά τη διάρκεια της εξέτασης, λαμβάνοντας υπόψη και τα δεδομένα του συνόλου εκπαίδευσης αλλά και τη γνώση από τα νέα πρότυπα.

Για το πρόβλημα αυτό χρησιμοποιήσαμε ένα 16-30-2 ΤΝΔ (540 βάρη και 32 πολώσεις), όπου οι νευρώνες του είχαν λογιστικές συναρτήσεις ενεργοποίησης. Το ΤΝΔ πρέπει να διακρίνει τα φυσιολογικά από τα «ύποπτα» τμήματα της εικόνας, δηλαδή να ξεχωρίσει εικόνες που περιέχουν τμήματα εντέρου από εικόνες που περιέχουν τμήματα όγκων. Χρησιμοποιήσαμε 4 σχεδόν διαδοχικές εικόνες από το ενδοσκόπιο (βλ. Σχήμα 7.1) και κάθε εικόνα χωρίστηκε σε περιοχές 16×16 εικονοκυττάρων. Συνολικά κάθε εικόνα χωρίστηκε σε περίπου 4000 τμήματα, από τα οποία δημιουργήσαμε τα σύνολα εκπαίδευσης και ελέγχου. Μια πολύ σημαντική φάση της αναγνώρισης ανωμαλιών σε κολονοσκοπήσεις είναι η επιλογή της διαδικασίας εξαγωγής προτύπων. Στα πειράματά μας χρησιμοποιήσαμε την μέθοδο εξαγωγής προτύπων που πρότεινε ο Haralick στην εργασία [46]. Περισσότερες τεχνικές λεπτομέρειες σχετικά με την εξαγωγή των προτύπων εκπαίδευσης βρίσκονται στην εργασία [58].

Έτσι δημιουργήσαμε 4 σύνολα εκπαίδευσης (ένα για κάθε εικόνα), που το καθένα αποτελείται από 300 τυχαία επιλεγμένα πρότυπα (από τα 4000 συνολικά πρότυπα για κάθε εικόνα). Στη συνέχεια εκπαιδεύσαμε το παραπάνω ΤΝΔ ανά πρότυπο εισόδου με τον Αλ-



Σχήμα 7.1: Διαδοχικές εικόνες από ενδοσκόπιο

γόριθμο 7.1, χρησιμοποιώντας μόνο τα πρότυπα από την πρώτη εικόνα. Η ιδέα ήταν να δοκιμάσουμε το εκπαιδευμένο ΤΝΔ σε πρότυπα από τις επόμενες εικόνες, ώστε να διαπιστώσουμε την ικανότητα γενίκευσής του ειδικότερα όταν το σύνολο εκπαίδευσης μεταβάλλεται με τον χρόνο.

Η εκπαίδευση τερματίστηκε όταν το ΤΝΔ είχε σφάλμα ταξινόμησης στο σύνολο εκπαίδευσης, $CE \leq 3\%$. Πρέπει εδώ να σημειωθεί ότι η φάση αυτή της εκπαίδευσης ήταν πολύ γρήγορη· χρειάστηκαν μόνο 40 επαναλήψεις. Το εκπαιδευμένο ΤΝΔ επανεκπαιδεύεται ανά πρότυπο εισόδου με χρήση ΔΕΑ. Στη φάση αυτή χρησιμοποιούμε σαν σύνολο εκπαίδευσης και τα 4 αρχικά σύνολα εκπαίδευσης, δηλαδή το νέο σύνολο εκπαίδευσης αποτελείται από 1200 πρότυπα. Έτσι μπορούμε να δούμε την συμπεριφορά του ΤΝΔ και των αλγόριθμων εκπαίδευσης όταν τα δεδομένα μεταβάλλονται με τον χρόνο, αφού οι άλλες τρεις εικόνες είναι αυτές που λάβαμε από το ενδοσκόπιο μετά την πρώτη. Επιτρέπουμε στον ΔΕΑ να κάνει μόνο δύο επαναλήψεις με κάθε πρότυπο, έτσι ώστε να αποφύγουμε την «καταστροφική παρέμβαση» των προτύπων από τις 3 νέες εικόνες με τα ήδη γνωστά πρότυπα της πρώτης εικόνας. Για τον έλεγχο του τελικού ΤΝΔ χρησιμοποιήσαμε όλα τα διαθέσιμα πρότυπα, δηλαδή 4000 πρότυπα από κάθε μια από τις 4 εικόνες. Τα αποτελέσματα γενίκευσης παρουσιάζονται στον Πίνακα 7.4.

Στον Πίνακα 7.4 βλέπουμε π.χ. ότι το ΤΝΔ που εκπαιδεύσαμε με πρότυπα από την πρώτη εικόνα, πριν την επανεκπαίδευση, είχε 83.77% επιτυχία αναγνώρισης, ενώ μετά την επανεκπαίδευση το ποσοστό αναγνώρισης για την πρώτη εικόνα ανέβηκε στο 91.91%. Είναι προφανές ότι το επανεκπαιδευμένο ΤΝΔ έχει μεγαλύτερο ποσοστό επιτυχίας και στις 4 εικόνες που εξετάσαμε. Μετά το τέλος της εκπαίδευσης, το ΤΝΔ είναι σε θέση να αναγνωρίζει αυτόματα «ύποπτα» τμήματα στις εικόνες που παρέχει το ενδοσκόπιο. Τέλος, αξίζει να σημειώσουμε ότι τα ποσοστά γενίκευσης που επιτύχαμε με την μέθοδο που περιγράφηκε είναι

ανάλογα με τα καλύτερα ποσοστά που έχουν βρεθεί σε ΤΝΔ που εκπαιδεύτηκαν με μεθόδους εκπαίδευσης ανά ομάδα προτύπων εισόδου [58].

Πίνακας 7.4: Αποτελέσματα από το πρόβλημα αναγνώρισης ανωμαλιών σε κολονοσκοπήσεις

	Χωρίς Εξέλιξη (Φάση 1 μόνο)	Με Εξέλιξη (Νέος Υβριδικός Αλγόριθμος)
Εικόνα 1	83.77%	91.91%
Εικόνα 2	77.18%	83.57%
Εικόνα 3	82.84%	93.09%
Εικόνα 4	87.60%	89.24%

7.5 Συμπεράσματα – Συνεισφορά

Σε αυτό το κεφάλαιο, προτάθηκε ένας νέος αλγόριθμος εκπαίδευσης ανά πρότυπο εισόδου για ΤΝΔ. Οι αλγόριθμοι εκπαίδευσης ανά πρότυπο εισόδου είναι σε θέση να εκπαιδεύσουν αποδοτικά μεγάλα ΤΝΔ με βάρη πλήθους της τάξης των χιλιάδων και ταιριάζουν καλύτερα για την εκπαίδευση μεγάλων συνόλων προτύπων, ή συνόλων που περιέχουν επαναλαμβανόμενα ή αργά μεταβαλλόμενα με τον χρόνο πρότυπα. Επίσης παρέχουν την δυνατότητα της συνεχούς εκπαίδευσης με νέα δεδομένα που δεν ήταν διαθέσιμα την χρονική στιγμή της πρώτης εκπαίδευσης του ΤΝΔ.

Τα αποτελέσματα της προσομοίωσης δείχνουν ότι ο προτεινόμενος αλγόριθμος παρέχει γρήγορη και σταθερή εκπαίδευση σε σύγκριση με άλλες μεθόδους εκπαίδευσης ανά πρότυπο εισόδου, καθώς επίσης και μεθόδους εκπαίδευσης ανά ομάδα προτύπων εισόδου. Συνεπώς παρέχει μεγαλύτερη πιθανότητα επιτυχημένης και γρήγορης εκπαίδευσης για προβλήματα του πραγματικού κόσμου.

Επίσης, παρουσιάσαμε μια νέα υβριδική μέθοδο και εξετάσαμε την απόδοσή της σε δύο πραγματικές εφαρμογές. Τα αποτελέσματα γενίκευσης των ΤΝΔ που αρχικά εκπαιδεύτηκαν από την προτεινόμενη μέθοδο εκπαίδευσης ανά πρότυπο εισόδου (Φάση 1) και στη συνέχεια εκπαιδεύτηκαν ξανά με ένα Εξελικτικό Αλγόριθμο (Φάση 2) είναι ικανοποιητικά και ανάλογα με τα καλύτερα αποτελέσματα μεθόδων που εκπαιδεύουν ανά ομάδα προτύπων εισόδου. Η προτεινόμενη υβριδική μέθοδος ανταποκρίθηκε με επιτυχία στο μη στατικό πρόβλημα της αναγνώρισης ανωμαλιών σε κολονοσκοπήσεις και αποδείχτηκε σθεναρή και προβλέψιμη. Τέλος, αξίζει να αναφερθεί ότι δεν παρατηρήθηκε το φαινόμενο της «καταστροφικής παρέμβασης» μεταξύ των προτύπων που προέρχονταν από διαφορετικές εικόνες.

Μη Μονότονοι Αλγόριθμοι Εκπαίδευσης

Ακούω και ξεχνώ.
Βλέπω και θυμάμαι.
Πράττω και καταλαβαίνω.

—Confucius (551–479 π.Χ.)

Σε αυτό το κεφάλαιο θα παρουσιάσουμε μη μονότονες αιτιοκρατικές στρατηγικές εκπαίδευσης ΤΝΔ. Το κύριο χαρακτηριστικό αυτών των στρατηγικών είναι ότι κατά την διάρκεια της εκπαίδευσης είναι δυνατό η ακολουθία των συναρτησιακών τιμών των διανυσμάτων των βαρών να μην είναι μονότονη (φθίνουσα) [34]. Οι στρατηγικές αυτές μπορούν να ενσωματωθούν σε οποιονδήποτε αλγόριθμο εκπαίδευσης ανά ομάδα προτύπων εισόδου και παρέχουν ταχεία, σταθερή και αξιόπιστη σύγκλιση και εκπαίδευση. Τα αποτελέσματά μας, σε πολλές και διαφορετικές κατηγορίες προβλημάτων, δείχνουν ότι αυτή η προσέγγιση βελτιώνει την ταχύτητα εκπαίδευσης και το ποσοστό επιτυχίας όλων των αλγορίθμων που δοκιμάσαμε και μας απαλλάσσει από το δύσκολο έργο της κατάλληλης ρύθμισης των ευρετικών παραμέτρων των αλγορίθμων εκπαίδευσης, που πολύ συχνά εξαρτώνται άμεσα από το δοθέν πρόβλημα.

8.1 Μη Μονότονες Στρατηγικές Εκπαίδευσης

Αν και οι μονότονες στρατηγικές εκπαίδευσης παρέχουν έναν αποδοτικό και αποτελεσματικό τρόπο για να εξασφαλιστεί ότι η συνάρτηση σφάλματος μειώνεται αρκετά, έχουν το μειονέκτημα ότι πληροφορίες που θα μπορούσαν να επιταχύνουν τη σύγκλισή τους, δεν αποθηκεύονται και δεν χρησιμοποιούνται [38]. Για να αξιοποιήσουμε αυτές τις πληροφορίες, προτείνουμε μια νέα μη μονότονη στρατηγική εκπαίδευσης που χρησιμοποιεί την συσσωρευμένη πληροφορία αναφορικά με τις M το πλήθος πιο πρόσφατες τιμές της συνάρτησης σφάλματος [106, 113, 123].

Χρησιμοποιούμε την ακόλουθη συνθήκη για τον καθορισμό ενός κριτηρίου αποδοχής οποιουδήποτε νέου διανύσματος βαρών:

$$E(w^k + \eta^k \varphi^k) - \max_{0 \leq j \leq M} \{E(w^{k-j})\} \leq \gamma \eta^k \langle \nabla E(w^k), \varphi^k \rangle, \quad (8.1)$$

όπου M είναι ένας μη αρνητικός ακέραιος αριθμός, ονομαζόμενος *μη μονότονος ορίζοντας εκπαίδευσης*, $0 < \gamma < 1$, η^k είναι ο ρυθμός εκπαίδευσης και φ^k είναι η κατεύθυνση ανίχνευσης στην k επανάληψη. Η ανωτέρω συνθήκη επιτρέπει μια αύξηση στις τιμές της συνάρτησης σφάλματος, χωρίς βλάβη της ιδιότητας της ευρείας σύγκλισης, σύμφωνα με το ακόλουθο θεώρημα [43, 127].

Θεώρημα 8.1 [43] Έστω $\{w^k\}$ μια ακολουθία βαρών που προκύπτει από το ακόλουθο επαναληπτικό σχήμα:

$$w^{k+1} = w^k + \eta^k \varphi^k, \quad \varphi^k \neq 0,$$

όπου φ^k είναι η κατεύθυνση ανίχνευσης και η^k ο ρυθμός εκπαίδευσης στην k επανάληψη. Έστω ότι:

- (i) το σύνολο $\Omega_0 = \{w : E(w) \leq E(w^0)\}$ είναι συμπαγές,
- (ii) υπάρχουν δύο θετικοί αριθμοί c_1 και c_2 , έτσι ώστε:

$$\begin{aligned} \nabla E(w^k)^\top \varphi^k &\leq -c_1 \|\nabla E(w^k)\|^2, \\ \|\varphi^k\| &\leq c_2 \|\nabla E(w^k)\|. \end{aligned}$$

- (iii) χρησιμοποιείται το ακόλουθο κριτήριο αποδοχής των νέων διανυσμάτων βαρών

$$E(w^k + \eta^k \varphi^k) - \max_{0 \leq j \leq M} \{E(w^{k-j})\} \leq \gamma \eta^k \langle \nabla E(w^k), \varphi^k \rangle,$$

όπου M είναι μη αρνητικός ακέραιος αριθμός και $\gamma \in (0, 1)$.

Τότε ισχύουν τα ακόλουθα:

- οι όροι της ακολουθίας $\{w^k\}$ παραμένουν εντός του συνόλου Ω_0 και για κάθε οριακό σημείο $\bar{w}$ ισχύει $\nabla E(\bar{w}) = 0$,
- κανένα οριακό σημείο της ακολουθίας $\{w^k\}$ δεν είναι τοπικό ελάχιστο της E ,
- εάν το πλήθος των στάσιμων σημείων της E εντός του συνόλου Ω_0 είναι πεπερασμένο, τότε η ακολουθία $\{w^k\}$ συγκλίνει.

Αν ικανοποιούνται οι υποθέσεις του παραπάνω θεωρήματος, τότε η μη μονότονη στρατηγική εκπαίδευσης παράγει μια ευρέως συγκλίνουσα ακολουθία διανυσμάτων βαρών, για οποιονδήποτε αλγόριθμο που ακολουθεί μια κατεύθυνση ανίχνευσης $\varphi^k \neq 0$. Στην συνέχεια, συνοψίζουμε τα βασικά βήματα της μη μονότονης στρατηγικής εκπαίδευσης στην επανάληψη k :

- 1:** Ενημέρωσε τα βάρη σύμφωνα με την σχέση: $w^{k+1} = w^k + \eta^k \varphi^k$.
- 2:** Αν $E(w^{k+1}) - \max_{0 \leq j \leq M^k} E(w^{k-j}) \leq \gamma \eta^k \langle \nabla E(w^k), \varphi^k \rangle$, αποθήκευσε το w^{k+1} , θέσε $k = k + 1$ και πήγαινε στο Βήμα 1. Αλλιώς πήγαινε στο επόμενο Βήμα.
- 3:** Χρησιμοποίησε μια οποιαδήποτε τεχνική ρύθμισης του η^k και επέστρεψε στο Βήμα 2.

Μια απλή τεχνική ρύθμισης για το η^k στο Βήμα 3 είναι να μειώνεται το ρυθμός εκπαίδευσης κατά έναν παράγοντα μείωσης $1/q$, όπου $q > 1$. Παρατηρούμε ότι η επιλογή του q δεν είναι κρίσιμη για την επιτυχή εκπαίδευση, εντούτοις μπορεί να έχει άμεση επίδραση στον αριθμό των υπολογισμών της συνάρτησης σφάλματος που απαιτούνται για να λάβουμε ένα αποδεκτό διάνυσμα βάρους. Κατά συνέπεια, για μερικά προβλήματα εκπαίδευσης αρκούν μια ή δύο μειώσεις του ρυθμού εκπαίδευσης κατά μέτρια ποσοστά (όπως $1/2$), ενώ άλλα απαιτούν πολλές τέτοιες μειώσεις, ή ίσως απαιτούν πιο δραστική μείωση του ρυθμού εκπαίδευσης (για παράδειγμα κατά $1/10$, ή ακόμα και $1/20$). Αφ' ετέρου, η υπερβολική μείωση η^k μπορεί να είναι δαπανηρή, δεδομένου ότι ο συνολικός αριθμός των επαναλήψεων θα αυξηθεί. Στη βιβλιογραφία συνήθως προτείνεται η τιμή $q = 2$ [7], η οποία επίσης επιβεβαιώθηκε στα πειράματά μας να λειτουργεί αποτελεσματικά και αποδοτικά. Η ανωτέρω διαδικασία αποτελεί μια αποδοτική μέθοδο καθορισμού ενός κατάλληλου ρυθμού εκπαίδευσης χωρίς πρόσθετους υπολογισμούς του διανύσματος των μερικών παραγώγων της

συνάρτησης σφάλματος. Κατά συνέπεια, το πλήθος των υπολογισμών του διανύσματος των μερικών παραγώγων είναι, γενικά, μικρότερο από ότι είναι το πλήθος των υπολογισμών της τιμής της συνάρτησης σφάλματος.

Η μη μονότονη στρατηγική εκπαίδευσης μπορεί να ενσωματωθεί σε οποιονδήποτε αλγόριθμο εκπαίδευσης ανά ομάδα προτύπων εισόδου. Μπορεί να χρησιμοποιηθεί ως μια τεχνική που διασφαλίζει και επιταχύνει τη σύγκλιση του αλγορίθμου, παρέχοντας την δυνατότητα αντιμετώπισης αυθαίρετα μεγάλων ρυθμών εκπαίδευσης, και κατ' αυτό τον τρόπο, για ένα δεδομένο πρόβλημα, γίνεται εφικτή η εκπαίδευση ΤΝΔ με την *πρώτη προσπάθεια*. Επιπλέον, επιλύει προβλήματα όπως αυτό της μειωμένης ταχύτητας σύγκλισης, του πρώτου κορεσμού της μεθόδου ή ακόμα και της απόκλισης [68], που δημιουργούνται από την εσφαλμένη επιλογή ευρετικών παραμέτρων των αλγορίθμων εκπαίδευσης.

8.1.1 Ο μη μονότονος ορίζοντας εκπαίδευσης

Πειραματικά αποτελέσματα που έχουμε παρουσιάσει στην εργασία [123], δείχνουν ότι η επιλογή της παραμέτρου M είναι κρίσιμη και εξαρτάται από τη φύση του προβλήματος. Εδώ, προτείνουμε μια διαδικασία εκτίμησης της τιμής του μη μονότονου ορίζοντα εκμάθησης M , που χρησιμοποιεί την έννοια της σταθεράς Lipschitz.

Είναι ευρέως γνωστό ότι η σταθερά Lipschitz συσχετίζεται άμεσα με τη μορφολογία της συνάρτησης [7, 22]. Παραδείγματος χάριν, η σταθερά Lipschitz για μια συνάρτηση που παρουσιάζει απότομες περιοχές έχει μεγάλη τιμή. Επίσης, όταν η συνάρτηση είναι επίπεδη, τότε η τιμή της σταθεράς Lipschitz είναι μικρή. Δυστυχώς, στην πράξη ούτε η μορφολογία της επιφάνειας σφάλματος αλλά ούτε και η τιμή της σταθεράς Lipschitz είναι γνωστά στην αρχή της εκπαίδευσης. Προκειμένου να επιλυθεί το πρόβλημα αυτό, μια τοπική εκτίμηση της σταθεράς Lipschitz έχει προταθεί [82], η οποία παρέχει πληροφορίες σχετικά με την τοπική μορφή της συνάρτησης σφάλματος (βλ. επίσης και [162] για τη χρησιμότητα αυτής της εκτίμησης).

Η ακόλουθη διαδικασία παρέχει μια σθεναρή αντιμετώπιση του δυναμικού υπολογισμού της τιμής του μη μονότονου ορίζοντα M σε κάθε επανάληψη:

$$M^k = \begin{cases} M^{k-1} + 1, & \Lambda^k < \Lambda^{k-1} < \Lambda^{k-2}, \\ M^{k-1} - 1, & \Lambda^k > \Lambda^{k-1} > \Lambda^{k-2}, \\ M^{k-1}, & \text{αλλιώς,} \end{cases} \quad (8.2)$$

όπου Λ^k είναι η τοπική εκτίμηση της σταθεράς Lipschitz στην k επανάληψη [82]:

$$\Lambda^k = \frac{\|\nabla E(w^k) - \nabla E(w^{k-1})\|}{\|w^k - w^{k-1}\|}. \quad (8.3)$$

Η τιμή Λ^k μπορεί να ληφθεί χωρίς πρόσθετους υπολογισμούς της τιμής ή του διανύσματος των μερικών παραγώγων της συνάρτησης σφάλματος.

Εάν το Λ^k αυξάνεται για δύο διαδοχικές επαναλήψεις, η ακολουθία των διανυσμάτων βαρών πλησιάζει μια απότομη περιοχή και η τιμή του M πρέπει να μειωθεί προκειμένου να ανιχνευτεί ένα πιθανό τοπικό ελάχιστο. Όταν το Λ^k μειώνεται για δύο διαδοχικές επαναλήψεις, η μέθοδος ενδεχομένως εισέρχεται σε μια επίπεδη περιοχή (κοιλιάδα) του χώρου των βαρών, έτσι η τιμή του M πρέπει να αυξηθεί. Αυτό επιτρέπει στη μέθοδο να αποδεχτεί μεγαλύτερους ρυθμούς εκπαίδευσης και να ξεφύγει γρηγορότερα από αυτή την περιοχή. Τέλος, όταν έχει η τιμή του Λ^k έχει μάλλον τυχαία συμπεριφορά (αυξάνεται ή μειώνεται για μόνο μια επανάληψη), η τιμή του M παραμένει αμετάβλητη.

Είναι προφανές ότι το M πρέπει να είναι θετικό. Κατά συνέπεια, εάν η Σχέση (8.2) δώσει μια μη θετική τιμή στο M , η τιμή του μη μονότονου ορίζοντα εκπαίδευσης τίθεται ίση με 1, προκειμένου να εξασφαλιστεί μείωση της συνάρτησης σφάλματος στην τρέχουσα επανάληψη.

8.1.2 Ανάπτυξη μη μονότονων αλγορίθμων εκπαίδευσης

Εδώ περιγράφουμε τις μη μονότονες τροποποιήσεις της μεθόδου οπισθοδρομικής διάδοσης του σφάλματος (BP) με ορμή (momentum) (BPM) [57, 133], καθώς επίσης και δύο μη μονότονες μεθόδους με προσαρμοστικό ρυθμό εκπαίδευσης.

1) *Η μη μονότονη BP μέθοδος με ορμή*: Μια απλή, ευρετική στρατηγική για την επιτάχυνση της μεθόδου BP έχει προταθεί στις εργασίες [57, 133] και είναι βασισμένη στη χρήση ενός όρου ορμής (momentum term). Ο όρος αυτός ενσωματώνεται στη μέθοδο της πιο απότομης καθόδου, ως εξής:

$$w^{k+1} = w^k - (1 - m)\eta \nabla E(w^k) + m(w^k - w^{k-1}),$$

όπου m είναι η σταθερά ορμής. Ένα μειονέκτημα του ανωτέρω σχήματος είναι ότι, εάν στο m δοθεί μια συγκριτικά μεγάλη τιμή, οι πληροφορίες κλίσης από τις προηγούμενες επαναλήψεις επηρεάζουν την ενημέρωση των βαρών περισσότερο από τις τρέχουσες πληροφορίες κλίσης. Μια λύση είναι να αυξηθεί ο ρυθμός εκπαίδευσης, εντούτοις στην πράξη αυτή η προσέγγιση αποδεικνύεται συχνά ατελέσφορη και οδηγεί στην αστάθεια της μεθόδου, στον κορεσμό και την απόκλιση της. Κατά συνέπεια, εάν η σταθερά ορμής m αυξάνεται, μπορεί να είναι απαραίτητο να εφαρμοστεί μια αντισταθμιστική μείωση στο ρυθμό εκπαίδευσης η για να διατηρηθεί η σταθερότητα της εκπαίδευσης. Στο πείραμα που αναφέρουμε στη συνέχεια, επιλύουμε αυτό το πρόβλημα με το συνδυασμό της μεθόδου BPM με τη μη μονότονη στρατηγική εκπαίδευσης. Αυτή η νέα μέθοδος ονομάζεται NMBPM.

2) *Η μη μονότονη BP με μεταβλητό βήμα*: Η μη μονότονη BP με μεταβλητό βήμα (Back-Propagation with Variable Stepsize - BPVS) [82], αξιοποιεί την τοπική μορφή της επιφάνειας σφάλματος με τον υπολογισμό της σταθεράς Lipschitz σε κάθε επανάληψη, και καθορίζει το ρυθμό εκπαίδευσης η^k , σύμφωνα με τον ακόλουθο τύπο:

$$\eta^k = \frac{1}{2\Lambda^k} = \frac{\|w^k - w^{k-1}\|}{2\|\nabla E(w^k) - \nabla E(w^{k-1})\|}, \quad (8.4)$$

όπου Λ^k η τοπική εκτίμηση της σταθεράς Lipschitz την k επανάληψη. Κατά συνέπεια, όταν η επιφάνεια σφάλματος έχει απότομες περιοχές, το Λ^k είναι μεγάλο, και μια μικρή τιμή για τον ρυθμό εκπαίδευσης είναι κατάλληλη προκειμένου να εγγυηθεί τη σύγκλιση. Αφ' ετέρου όταν έχει η επιφάνεια σφάλματος είναι επίπεδη, το Λ^k είναι μικρό και μεγάλος ρυθμός εκπαίδευσης επιλέγεται για να επιταχύνει τη σύγκλιση.

Προκειμένου να αποκλείσουμε την πιθανότητα μιας ακατάλληλης τοπικής εκτίμησης της σταθεράς Lipschitz, συνδυάζουμε τη μέθοδο BPVS με τη μη μονότονη στρατηγική εκπαίδευσης. Αυτή η νέα παραλλαγή της BPVS, που παρέχει την δυνατότητα μη μονότονης εκπαίδευσης, ονομάζεται NMBPVS.

3) *Η μη μονότονη μέθοδος των Barzilai-Borwein*: Στην εργασία [116] έχουμε προτείνει μια μέθοδο για την εκπαίδευση ΤΝΔ, αποκαλούμενη Barzilai-Borwein backPropagation (BBP), και η οποία είναι βασισμένη στη μέθοδο των Barzilai και Borwein [12].

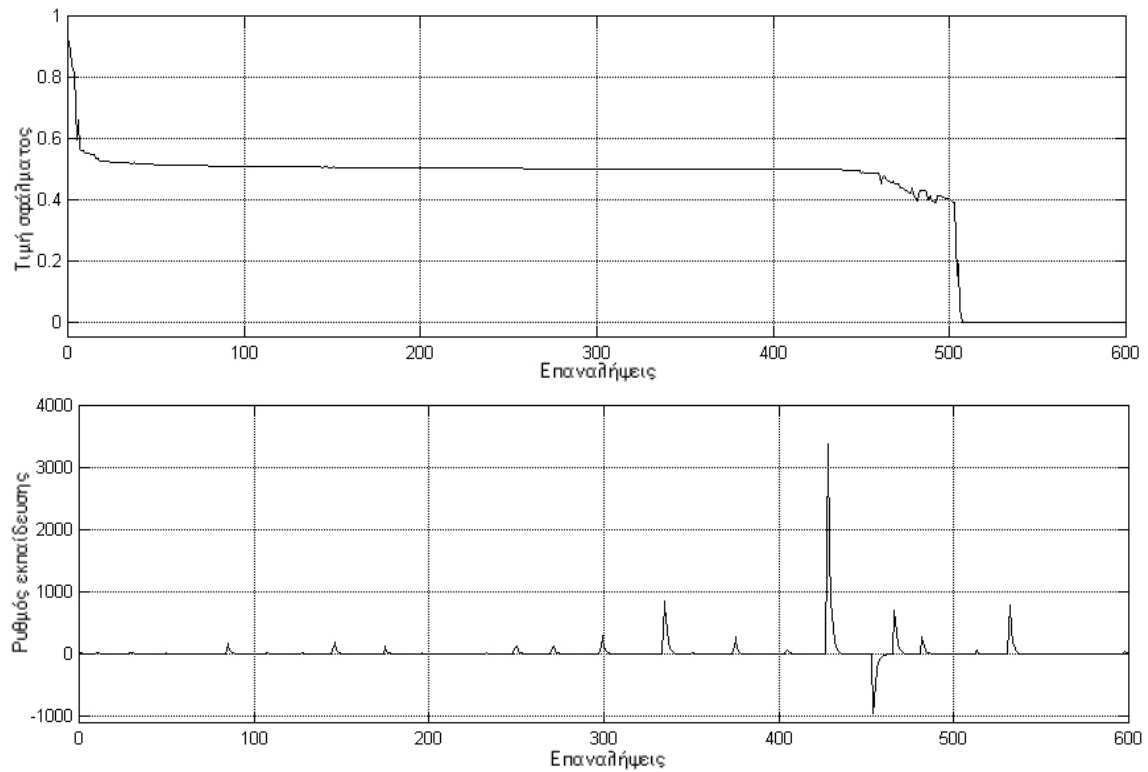
Αυτή είναι μια μέθοδος όπου η κατεύθυνση ανίχνευσης είναι πάντα η αντίθετη της κατεύθυνσης του διανύσματος των μερικών παραγώγων της συνάρτησης σφάλματος (κατεύθυνση της πιο απότομης καθόδου), και ο ρυθμός εκπαίδευσης υπολογίζεται χρησιμοποιώντας τον ακόλουθο τύπο:

$$\eta^k = \frac{\langle \delta^{k-1}, \delta^{k-1} \rangle}{\langle \delta^{k-1}, \psi^{k-1} \rangle}, \quad (8.5)$$

όπου $\delta^{k-1} = w^k - w^{k-1}$ και $\psi^{k-1} = \nabla E(w^k) - \nabla E(w^{k-1})$. Το κίνητρο για αυτήν την επιλογή είναι ότι παρέχει μια προσέγγιση δύο σημείων στη εξίσωση της χορδής των quasi-Newton μεθόδων [125]. Τα κύρια χαρακτηριστικά αυτής της μεθόδου είναι οι χαμηλές απαιτήσεις σε μνήμη και οι ανέξοδοι αριθμητικοί υπολογισμοί. Επιπλέον, δεν εγγυάται την μονότονη

μείωση της συνάρτησης σφάλματος E . Τα πειράματά μας (βλ. επίσης την εργασία [123]) έχουν δείξει ότι αυτή η ιδιότητα είναι πολύτιμη για την εκπαίδευση ΤΝΔ επειδή, πολύ συχνά, η μέθοδος ξεφεύγει από τοπικά ελάχιστα, όπου οι άλλες μέθοδοι παγιδεύονται. Για να εξασφαλίσουμε τη σύγκλιση της μεθόδου, ακόμα και όταν ο ανωτέρω τύπος δίνει ακατάλληλους ρυθμούς εκπαίδευσης, εφαρμόζουμε τη μη μονότονη στρατηγική εκπαίδευσης. Καλούμε αυτόν τον τροποποιημένο αλγόριθμο εκπαίδευσης, NMBBP.

Η νέα μέθοδος διατηρεί την δυνατότητα της BBP να διαφεύγει από τις ανεπιθύμητες περιοχές στο χώρο των βαρών, όπως φαίνεται στο Σχήμα 8.1. Στο πάνω μέρος του Σχήματος 8.1 παρουσιάζουμε τη συμπεριφορά της NMBBP στο πρόβλημα του αποκλειστικού-ΕΙΤΕ (βλ. Παράρτημα Α.1), ενώ στο κάτω μέρος απεικονίζεται ο ρυθμός εκπαίδευσης. Η μέθοδος ξεφεύγει επιτυχώς από μια περιοχή με σχεδόν σταθερή τιμή της συνάρτησης σφάλματος ($E \simeq 0.5$), χρησιμοποιώντας μεγάλα θετικά καθώς επίσης και αρνητικά βήματα. Να σημειωθεί ότι οι αρνητικοί ρυθμοί εκπαίδευσης είναι δυνατοί λόγω του τύπου προσαρμογής του ρυθμού εκπαίδευσης (8.5).



Σχήμα 8.1: Το αποκλειστικό-ΕΙΤΕ: (α) η μη μονότονη συμπεριφορά της μεθόδου NMBBP και (β) η συμπεριφορά του προσαρμοστικού ρυθμού εκπαίδευσης

8.1.3 Μοντέλο αλγόριθμου με μεταβλητό ρυθμό εκμάθησης με χρήση της μη μονότονης στρατηγικής

Στη συνέχεια περιγράφουμε μια υψηλού επιπέδου περιγραφή ενός γενικού αλγόριθμου πρώτης τάξης, που ενσωματώνει τη μη μονότονη στρατηγική και χρησιμοποιεί έναν προκαθορισμένο από το χρήστη ορίζοντα εκπαίδευσης M .

Αρχικοποίηση. Τυχαία επέλεξε το αρχικό διάνυσμα των βαρών w^0 , δώσε το μέγιστο αριθμό των επαναλήψεων ME , και θέσε τον αριθμό των επαναλήψεων $k = 0$. Επίσης, δώσε την

επιθυμητή ακρίβεια ε , το $\gamma \in (0, 1)$, την τιμή του $M \in [1, ME]$, και τον αρχικό ρυθμό εκπαίδευσης η^0 .

1: Για $k = 0$, υπολόγισε το $w^1 = w^0 - \eta^0 \nabla E(w^0)$.

Επαναλήψεις. Για $k = 1, 2, \dots, ME$.

2: Υπολόγισε το ρυθμό εκπαίδευσης η^k χρησιμοποιώντας τον τύπο υπολογισμού (8.4) ή (8.5).

3: Υπολόγισε το διάνυσμα των βαρών w^{k+1} , σύμφωνα με την σχέση:

$$w^{k+1} = w^k - \eta^k \nabla E(w^k).$$

4: Εάν $M > k$ τότε θέσε $M' = k$, αλλιώς θέσε $M' = M$.

5: Έλεγχε τη μη μονότονη στρατηγική:

$$E(w^k - \eta^k \nabla E(w^k)) - \max_{0 \leq j \leq M'} \{E(w^{k-j})\} \leq -\gamma \eta^k \|\nabla E(w^k)\|^2.$$

Εάν το παραπάνω κριτήριο ικανοποιείται, πήγαινε στο Βήμα 7.

6: Θέσε $\eta^k = \eta^k / 2$ και πήγαινε στο Βήμα 5.

7: Εάν το κριτήριο σύγκλισης $E(w^{k+1}) \leq \varepsilon$ ικανοποιείται τότε τερμάτισε.

8: Εάν $k < ME$, θέσε $k = k + 1$ και πήγαινε στο Βήμα 2, αλλιώς τερμάτισε.

Τερματισμός. Αποθήκευσε τα τελικά βάρη w^{k+1} και την αντίστοιχη τιμή της συνάρτησης σφάλματος $E(w^{k+1})$.

Είναι φανερό ότι αντί της χρησιμοποίησης μιας προκαθορισμένης από το χρήστη τιμής για τον μη μονότονο ορίζοντα εκπαίδευσης M , η Σχέση (8.2) μπορεί να εφαρμοστεί για τον δυναμικό υπολογισμό του.

8.2 Πειραματικά Αποτελέσματα

Εδώ αξιολογούμε την απόδοση της μη μονότονης στρατηγικής εκπαίδευσης με την εκτέλεση τεσσάρων συνολικά πειραμάτων: (α) συγκρίνουμε τους αλγόριθμους NMBPVS και NMBBP με μερικούς γνωστούς και ευρέως χρησιμοποιούμενους αλγόριθμους εκπαίδευσης ΤΝΔ, (β) μελετάμε την απόδοση των αλγορίθμων NMBPVS και NMBBP χρησιμοποιώντας διάφορες τιμές του μη μονότονου ορίζοντα M , καθώς επίσης και αυτόματα υπολογιζόμενο M , (γ) συγκρίνουμε την απόδοση των αλγορίθμων με και χωρίς τη χρήση της μη μονότονης στρατηγικής, και (δ) αξιολογούμε την ικανότητα γενίκευσης των αλγορίθμων σε τέσσερα δύσκολα προβλήματα ελέγχου γενίκευσης.

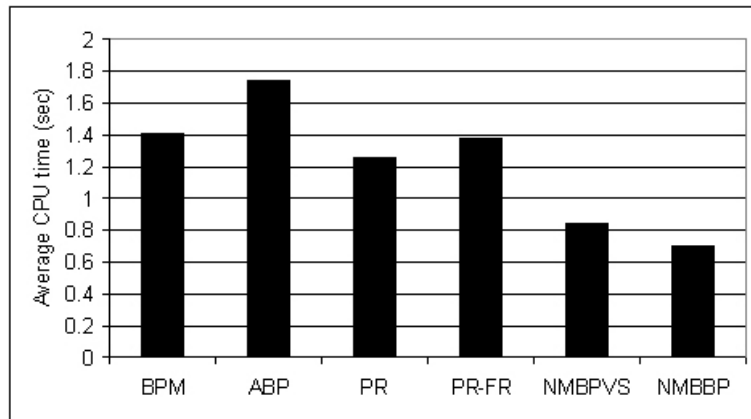
8.2.1 Συγκριτική μελέτη

Στην συνέχεια αξιολογούμε την απόδοση των μεθόδων NMBPVS και NMBBP, και τις συγκρίνουμε με τις μεθόδους οπισθοδρομικής διάδοσης του σφάλματος BP [133], οπισθοδρομικής διάδοσης του σφάλματος με ορμή BPM [57], οπισθοδρομικής διάδοσης του σφάλματος με προσαρμοστικό ρυθμό εκπαίδευσης και ορμή ABP [159], καθώς επίσης και με τις οι ευρέως συγκλίνουσες τροποποιήσεις των συζυγών μεθόδων κλίσης: Fletcher-Reeves (FR) [40], Polak-Ribiere (PR) [40], και Polak-Ribiere (PR) περιορισμένη από την FR μέθοδο (PR-FR) [40].

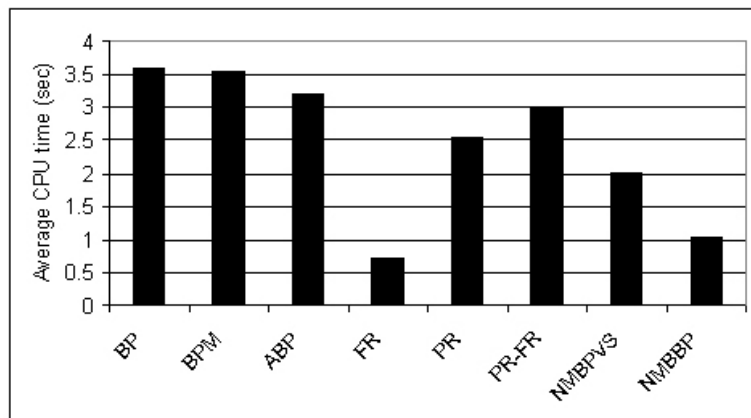
Για τους αλγόριθμους NMBPVS και NMBBP, έχουμε καθορίσει τις τιμές $M = 10$ και $\gamma = 10^{-5}$. Η μέθοδος των Nguyen-Widrow (βλ. την Ενότητα 1.5.2 του Κεφαλαίου 2 και την

εργασία [94]) έχει εφαρμοστεί για την αρχικοποίηση των βαρών και των πολώσεων για όλους τους αλγόριθμους, και η ίδια ακολουθία από πρότυπα έχει παρουσιαστεί στο κάθε ΤΝΔ. Να σημειωθεί ότι τα βάρη κάθε ΤΝΔ ενημερώνονται μόνο αφού έχει παρουσιαστεί ολόκληρο σύνολο των προτύπων εκπαίδευσης (εκπαίδευση ανά ομάδα προτύπων).

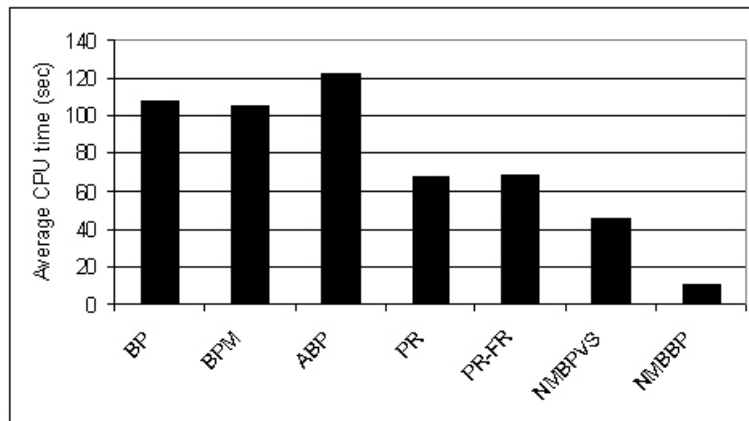
Για κάθε ένα από τα προβλήματα που περιγράφονται παρακάτω, παρουσιάζουμε έναν πίνακα που συνοψίζει την απόδοση των αλγορίθμων για τις προσομοιώσεις που έφθασαν σε λύση μέσα σε ένα προκαθορισμένο όριο υπολογισμών των τιμών της συνάρτησης σφάλματος (που διευκρινίζεται πιο κάτω). Οι αναφερόμενες στους πίνακες παράμετροι είναι: *Min* ο ελάχιστος αριθμός επαναλήψεων, μ η μέση τιμή των επαναλήψεων, *Max* ο μέγιστος αριθμός επαναλήψεων, σ η τυπική απόκλιση, και *Επιτυχία* ο αριθμός των επιτυχημένων προσομοιώσεων από ένα σύνολο 1000 δοκιμών. Εάν ένας αλγόριθμος αποτύχει να συγκλίνει μέσα στο όριο των υπολογισμών της συνάρτησης σφάλματος, θεωρείται ότι αποτυγχάνει να εκπαιδεύσει το ΤΝΔ και οι επαναλήψεις του (για τη συγκεκριμένη προσομοίωση) δεν συμπεριλαμβάνονται στην ανάλυση των αποτελεσμάτων του αλγορίθμου. Στα Σχήματα 8.2–8.5 παρουσιάζουμε τη συγκριτική ανάλυση του μέσου χρόνου σύγκλισης (CPU time) για τους παραπάνω αλγόριθμους σε τέσσερα προβλήματα εκπαίδευσης.



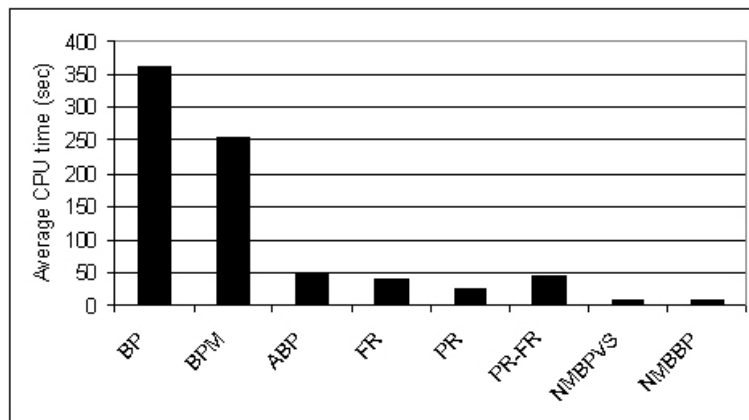
Σχήμα 8.2: Το πρόβλημα της ισοτιμίας 3-bit: μέσος χρόνος (CPU time) για τη σύγκλιση κάθε αλγορίθμου



Σχήμα 8.3: Το πρόβλημα προσέγγισης μιας συνεχούς συνάρτησης: μέσος χρόνος (CPU time) για τη σύγκλιση κάθε αλγορίθμου



Σχήμα 8.4: Το πρόβλημα αναγνώρισης των κεφαλαίων γραμμάτων: μέσος χρόνος (CPU time) για τη σύγκλιση κάθε αλγόριθμου



Σχήμα 8.5: Το πρόβλημα αναγνώρισης των αριθμών: μέσος χρόνος (CPU time) για τη σύγκλιση κάθε αλγόριθμου

Πρέπει εδώ να σημειώσουμε ότι για τις μεθόδους BP, BPM και ABP σε κάθε επανάληψη γίνεται ένας υπολογισμός της τιμής της συνάρτησης σφάλματος (Function Evaluation – FE) και ένας υπολογισμός του διανύσματος των μερικών παραγώγων της συνάρτησης σφάλματος (Gradient Evaluation – GE). Αφ’ ετέρου, ο αριθμός των υπολογισμών της τιμής της συνάρτησης σφάλματος των συζυγών μεθόδων κλίσης (PR, FR, και PR-FR) είναι, γενικά, μεγαλύτερος από τον αριθμό των υπολογισμών της κλίσης, λόγω της χρήσης μιας τεχνικής μη ακριβούς ευθύγραμμης (μονοδιάστατης) ανίχνευσης (inexact line search). Αυτό ισχύει και για τις μεθόδους NMBPVS και NMBBP, λόγω της χρήσης της μη μονότονης στρατηγικής που εφαρμόζεται.

Κατά συνέπεια, ακόμα και όταν οι μέθοδοι NMBPVS και NMBBP αποτυγχάνουν να συγκλίνουν μέσα στο προκαθορισμένο όριο των υπολογισμών των τιμών της συνάρτησης σφάλματος, ο αριθμός των υπολογισμών των κλίσεων είναι μικρότερος από τον αντίστοιχο αριθμό των άλλων μεθόδων. Λαμβάνοντας υπόψη ότι ένας υπολογισμός κλίσης είναι δαπανηρότερος από ένα υπολογισμό τιμής [92], είναι φανερό ότι αυτές οι μέθοδοι απαιτούν λιγότερες πράξεις κινητής υποδιαστολής (floating point operations) και είναι πραγματικά πολύ πιο γρήγορες. Από την ανωτέρω συζήτηση, είναι σαφές γιατί στους κατωτέρω πίνακες

υπάρχουν δύο σειρές αποτελεσμάτων για τις συζυγείς μεθόδους κλίσης και για τις μεθόδους NMBPVS και NMBBP· η πρώτη δείχνει την στατιστική ανάλυση για τους υπολογισμούς της τιμής της συνάρτησης σφάλματος (FE), ενώ η δεύτερη αναφέρεται στους υπολογισμούς του διανύσματος των μερικών παραγώγων της (GE).

1) *Ισοτιμία των 3-bit*: Για το πρόβλημα της ισοτιμίας (βλ. Παράρτημα Α.2), εκπαιδεύσαμε ένα 3-2-1 ΤΝΔ (8 βάρη και 3 πολώσεις). Το κρυφό στρώμα βασίστηκε σε νευρώνες με συνάρτηση ενεργοποίησης την υπερβολική εφαπτομένη. Για τους νευρώνες του στρώματος εξόδου χρησιμοποιήθηκαν γραμμικές συναρτήσεις ενεργοποίησης. Η συνθήκη τερματισμού ήταν να βρεθεί τιμή $E \leq 0.01$, μέσα σε 1000 υπολογισμούς τιμών της συνάρτησης σφάλματος. Τα αποτελέσματα φαίνονται στον Πίνακα 8.1.

Πίνακας 8.1: Αποτελέσματα από το πρόβλημα της ισοτιμίας 3-bit

Αλγόριθμος		<i>Min</i>	μ	<i>Max</i>	σ	<i>Επιτυχία</i>
BP		*	*	*	*	*
BPM		246	485.4	973	195.4	48.0%
ABP		465	599.2	924	103.9	45.0%
FR		*	*	*	*	*
PR	(FE)	189	465.3	972	188.2	53.2%
	(GE)	148	399.7	836	179.2	
PR-FR	(FE)	201	520.9	943	196.8	56.2%
	(GE)	170	432.3	829	187.2	
NMBPVS	(FE)	103	292.1	986	161.9	72.7%
	(GE)	98	285.5	960	156.8	
NMBBP	(FE)	47	298.9	978	212.4	67.9%
	(GE)	37	181.9	601	118.9	

* Ο αλγόριθμος απέτυχε να συγκλίνει μέσα στο όριο των υπολογισμών των τιμών της συνάρτησης σφάλματος.

Παρά την προσπάθεια που καταβάλλαμε για τη επιλογή του ρυθμού εκπαίδευσης, η μέθοδος BP απέτυχε να συγκλίνει μέσα στο όριο υπολογισμών των τιμών της συνάρτησης σφάλματος E , σε όλα τα πειράματα της προσομοίωσης. Αφ' ετέρου οι μέθοδοι BPM και ABP είχαν πολύ καλύτερα αποτελέσματα, με την μέθοδο BPM να είναι ελαφρώς καλύτερη από την μέθοδο ABP. Οι συζυγείς μέθοδοι κλίσης PR και PR-FR ήταν αποδοτικότερες και αποτελεσματικότερες από τις μεθόδους πρώτης τάξης, δηλαδή τις BP, BPM και ABP, αλλά η μέθοδος FR απέτυχε να συγκλίνει σε όλες τις δοκιμές μας. Όπως φαίνεται και στον Πίνακα 8.1, οι μέθοδοι NMBPVS και NMBBP είχαν την καλύτερη απόδοση, δεδομένου ότι έχουν το υψηλότερο ποσοστό επιτυχίας, καθώς επίσης και τον μικρότερο μέσο αριθμό υπολογισμών των τιμών της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της.

2) *Προσέγγιση μιας συνεχούς συνάρτησης*: Για το πρόβλημα της προσέγγισης μιας συνεχούς συνάρτησης (βλ. Παράρτημα Α.5), εκπαιδεύσαμε ένα 1-15-1 ΤΝΔ (30 βάρη και 16 πολώσεις), να προσεγγίσει την συνεχή συνάρτηση $f(x) = \sin(x) \cos(2x)$, χρησιμοποιώντας 20 ισαπέχουσες τιμές της από το διάστημα $[0, 2\pi]$.

Το ΤΝΔ εκπαιδεύεται έως ότου βρεθεί τιμή $E \leq 0.1$ μέσα σε 1000 υπολογισμούς τιμών της συνάρτησης σφάλματος E . Τα συγκριτικά αποτελέσματα παρουσιάζονται στον Πίνακα 8.2. Το κρυφό στρώμα είχε νευρώνες βασισμένους στην λογιστική συνάρτηση ενεργοποίησης, ενώ οι νευρώνες εξόδου είχαν γραμμικές ενεργοποιήσεις.

Αξίζει να σημειωθεί η αστάθεια στην απόδοση της μεθόδου FR, δεδομένου ότι φαίνεται να απαιτεί τον μικρότερο υπολογισμών, αλλά έχει πολύ μικρό ποσοστό επιτυχίας (17%). Αφ' ετέρου, οι προτεινόμενες μη μονότονες μέθοδοι (NMBPVS και NMBBP) έχουν πολύ καλή μέση απόδοση και επιδεικνύουν το υψηλότερο ποσοστό επιτυχίας.

Πίνακας 8.2: Αποτελέσματα από το πρόβλημα προσέγγισης μιας συνεχούς συνάρτησης

Αλγόριθμος		<i>Min</i>	μ	<i>Max</i>	σ	<i>Επιτυχία</i>
BP		328	706.7	998	175.6	13.8%
BPM		332	699.2	993	174.8	13.7%
ABP		166	628.1	994	216.8	26.9%
FR	(FE)	44	173.2	474	133.3	17.3%
	(GE)	38	87.8	164	43.4	
PR	(FE)	151	532.7	954	232.4	41.1%
	(GE)	119	465.5	940	235.1	
PR-FR	(FE)	80	597.5	997	283.3	43.0%
	(GE)	69	574.1	982	278.5	
NMBPVS	(FE)	64	443.9	991	238.5	66.8%
	(GE)	64	429.2	960	228.9	
NMBBP	(FE)	29	241.7	988	195.1	92.2%
	(GE)	26	158.2	625	114.7	

3) *Αναγνώριση των κεφαλαίων γραμμάτων:* Για το πρόβλημα αυτό, ένα 35-30-26 ΤΝΔ (1830 βάρη και 56 πολώσεις) εκπαιδεύεται να αναγνωρίζει τα κεφαλαία γράμματα της Αγγλικής αλφαβήτου (βλ. Παράρτημα Α.6). Το δίκτυο είναι βασισμένο σε νευρώνες με λογιστικές συναρτήσεις ενεργοποίησης στο κρυφό επίπεδο και γραμμικούς νευρώνες εξόδου. Το όριο εκπαίδευσης ήταν να βρεθεί τιμή $E \leq 0.1$, μέσα σε 2000 υπολογισμούς της συνάρτησης σφάλματος E . Τα αποτελέσματα στον Πίνακα 8.3 δείχνουν ότι οι μη μονότονοι αλγόριθμοι υπερέρχουν των άλλων μεθόδων που δοκιμάσαμε.

Πίνακας 8.3: Αποτελέσματα από το πρόβλημα αναγνώρισης γραμμάτων

Αλγόριθμος		<i>Min</i>	μ	<i>Max</i>	σ	<i>Επιτυχία</i>
BP		1098	1561.9	1999	202.8	76.8%
BPM		1142	1519.1	1931	169.3	4.9%
ABP		1119	1773.1	1999	168.9	37.2%
FR		*	*	*	*	*
PR	(FE)	340	1018.5	1673	275.7	100.0%
	(GE)	196	936.7	1669	317.1	
PR-FR	(FE)	388	998.5	1715	285.1	100.0%
	(GE)	244	988.3	1713	292.0	
NMBPVS	(FE)	337	669.4	1112	134.9	100.0%
	(GE)	322	633.3	1041	131.1	
NMBBP	(FE)	113	182.0	309	33.5	100.0%
	(GE)	73	119.4	193	20.5	

* Ο αλγόριθμος απέτυχε να συγκλίνει μέσα στο όριο των υπολογισμών των τιμών της συνάρτησης σφάλματος.

4) *Αναγνώριση αριθμών:* Σε αυτό το πείραμα εκπαιδεύεται ένα 64-6-10 ΤΝΔ (444 βάρη και 16 πολώσεις) ώστε να αναγνωρίζει τους αριθμούς από 0 έως 9 (βλ. Παράρτημα Α.7 και [82]). Το δίκτυο είναι βασισμένο σε νευρώνες με τη λογιστική συνάρτηση ενεργοποίησης, και τα βάρη αρχικοποιούνται με τους τυχαίους αριθμούς από την ομοιόμορφη κατανομή στο διάστημα $(-1, 1)$. Η συνθήκη τερματισμού είναι να βρεθεί τιμή της συνάρτησης σφάλματος τέτοια ώστε $E \leq 10^{-3}$.

Τα αποτελέσματα συνοψίζονται στον Πίνακα 8.4. Σαφώς, οι μέθοδοι NMBPVS και NMBBP έχουν την καλύτερη απόδοση και επιτυχία 100%. Παρατηρούμε επίσης ότι η μέ-

Θοδος NMBPVS έχει το μικρότερο μέσο αριθμό υπολογισμών του διανύσματος των μερικών παραγώγων της συνάρτησης σφάλματος.

Πίνακας 8.4: Αποτελέσματα από το πρόβλημα αναγνώρισης αριθμών

Αλγόριθμος		<i>Min</i>	μ	<i>Max</i>	σ	<i>Επιτυχία</i>
BP		9421	14489.2	19947	2783	66.2%
BPM		5328	10142.1	18756	1943	54.1%
ABP		228	1975.6	13822	2509	91.2%
FR	(FE)	366	2501.2	26560	5632.4	42.0%
	(GE)	260	620.3	3321	571	
PR	(FE)	806	1475.5	5585	763.7	96.1%
	(GE)	148	649.7	1099	109.1	
PR-FR	(FE)	1498	2723.5	5737	820.1	100.0%
	(GE)	533	750.3	1400	120.0	
NMBPVS	(FE)	104	380.4	1902	217.9	100.0%
	(GE)	88	268.9	1034	134.1	
NMBBP	(FE)	100	350.3	1532	182.5	100.0%
	(GE)	100	346.4	1488	178.5	

8.2.2 Μελέτη της επίδρασης του μη μονότονου ορίζοντα εκπαίδευσης

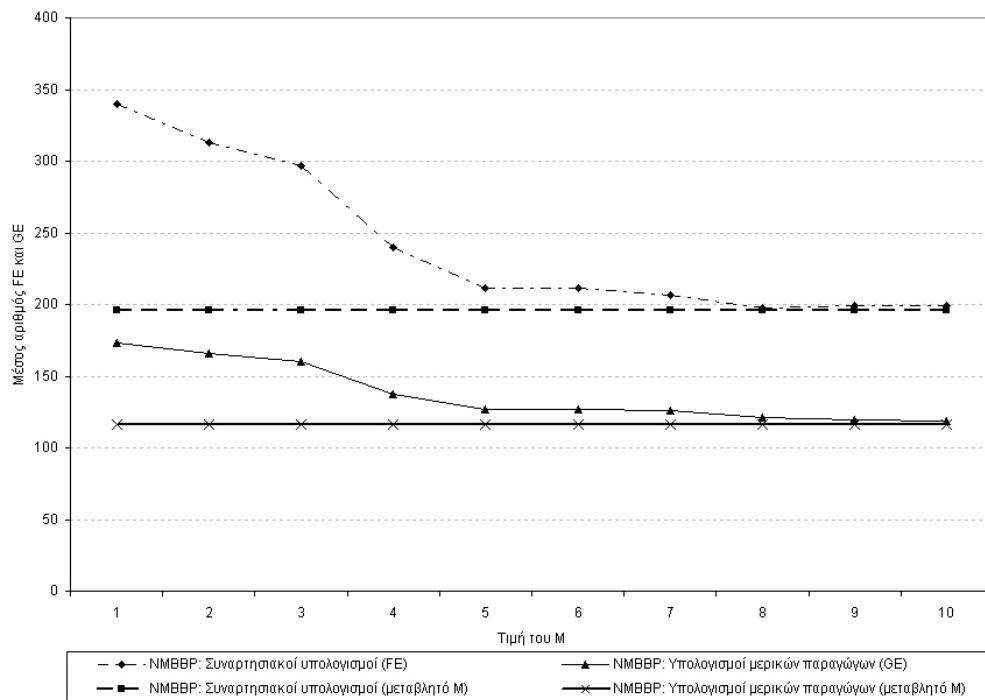
Για να μελετήσουν την επίδραση του μη μονότονου ορίζοντα εκπαίδευσης M στην απόδοση των μη μονότονων αλγορίθμων εκπαίδευσης, οι αλγόριθμοι NMBPVS και NMBBP εξετάστηκαν σε δύο προβλήματα. Για κάθε αλγόριθμο εκτελέστηκαν 100 προσομοιώσεις χρησιμοποιώντας τα ίδιες τυχαίες αρχικές τιμές για τα βάρη. Για τις δοκιμές έχουμε θέσει $\gamma = 10^{-5}$ και τον μέγιστο μη μονότονο ορίζοντα εκπαίδευσης $M = 10$.

Η πρώτη δοκιμή αφορά το προαναφερθέν πρόβλημα αναγνώρισης των κεφαλαίων γραμμάτων. Τα βάρη έχουν αρχικοποιηθεί με τη μέθοδο Nguyen-Widrow [94] και η συνθήκη τερματισμού είναι να βρεθεί σημείο με συναρτησιακή τιμή $E \leq 0.1$ μέσα στο όριο των 2000 συναρτησιακών υπολογισμών. Η δεύτερη δοκιμή είναι το πρόβλημα ταξινόμησης υφής (βλ. Παράρτημα Α.8). Η συνθήκη τερματισμού είναι το εκπαιδευμένο ΤΝΔ να έχει σφάλμα ταξινόμησης $CE \leq 3\%$, δηλαδή το ΤΝΔ να ταξινομεί σωστά τουλάχιστον 117 από τα 120 πρότυπα εκπαίδευσης.

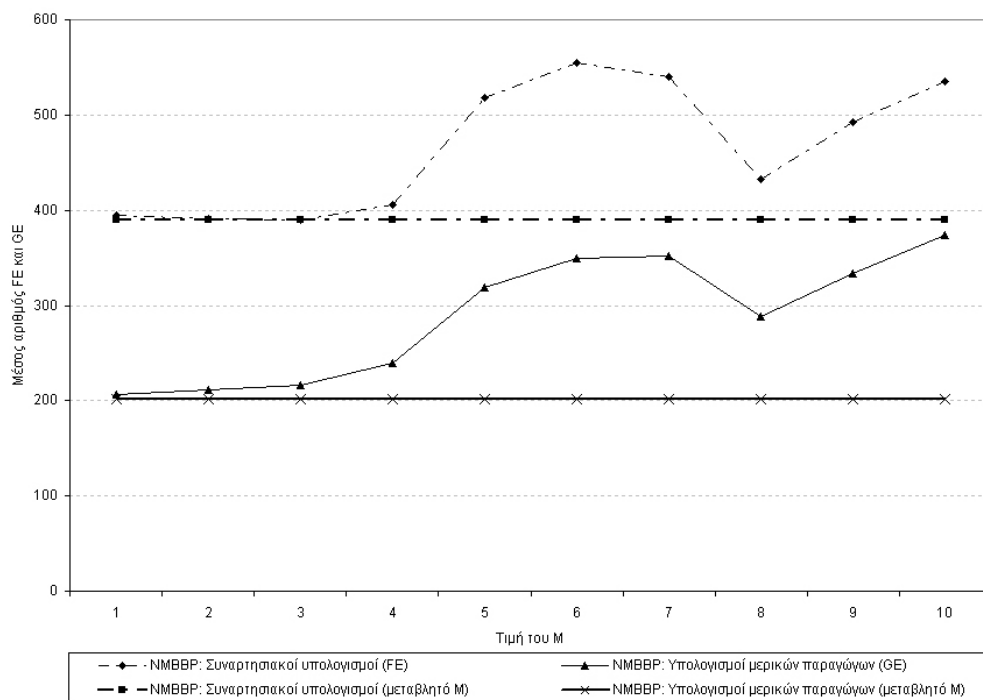
Τα λεπτομερή αποτελέσματα σχετικά με την επίδραση του μη μονότονου ορίζοντα εκπαίδευσης στις μεθόδους NMBPVS και NMBBP φαίνονται στα Σχήματα 8.6–8.9. Οι καμπύλες σε αυτές τις γραφικές παραστάσεις παρουσιάζουν την εξάρτηση της παραμέτρου M από τον μέσο αριθμό υπολογισμών των τιμών της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της. Η περίπτωση του προσαρμοζόμενου M , σύμφωνα με την Σχέση (8.2), απεικονίζεται ως ευθεία γραμμή από τα αριστερά προς στο δεξιά, προκειμένου να συγκρίνεται εύκολα με οποιαδήποτε από τις άλλες περιπτώσεις της δοκιμής. Από τα σχήματα είναι φανερό ότι η απόδοση των μεθόδων NMBPVS και NMBBP με ένα προσαρμοζόμενο μη μονότονο ορίζοντα M , είναι καλύτερη ή εξίσου καλή με τη μέση απόδοση που έχουν οι μέθοδοι για οποιαδήποτε προκαθορισμένη τιμή του M .

Αναφορικά με το ποσοστό επιτυχίας των μεθόδων, παρατηρούμε ότι η μέθοδος NMBPVS είχε ποσοστό επιτυχίας 100% και στα δύο πειράματα, ανεξάρτητα από την τιμή του μη μονότονου ορίζοντα M .

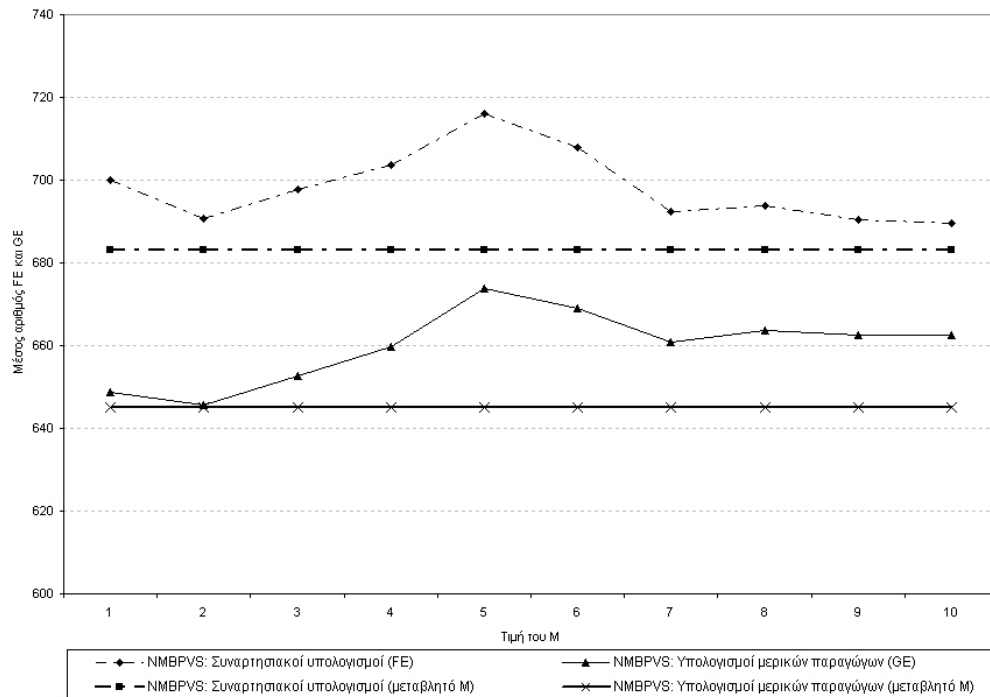
Από την άλλη μεριά, η απόδοση της NMBBP εξαρτάται από την επιλογή της τιμής του M . Στο πρόβλημα αναγνώρισης των κεφαλαίων γραμμάτων, η NMBBP έχει ποσοστό επιτυχίας 100% για κάθε τιμή του M , ενώ στο πρόβλημα ταξινόμησης υφής η μέθοδος έχει το υψηλότερο ποσοστό επιτυχίας όταν χρησιμοποιεί προσαρμοζόμενο M , όπως φαίνεται καθαρά στο



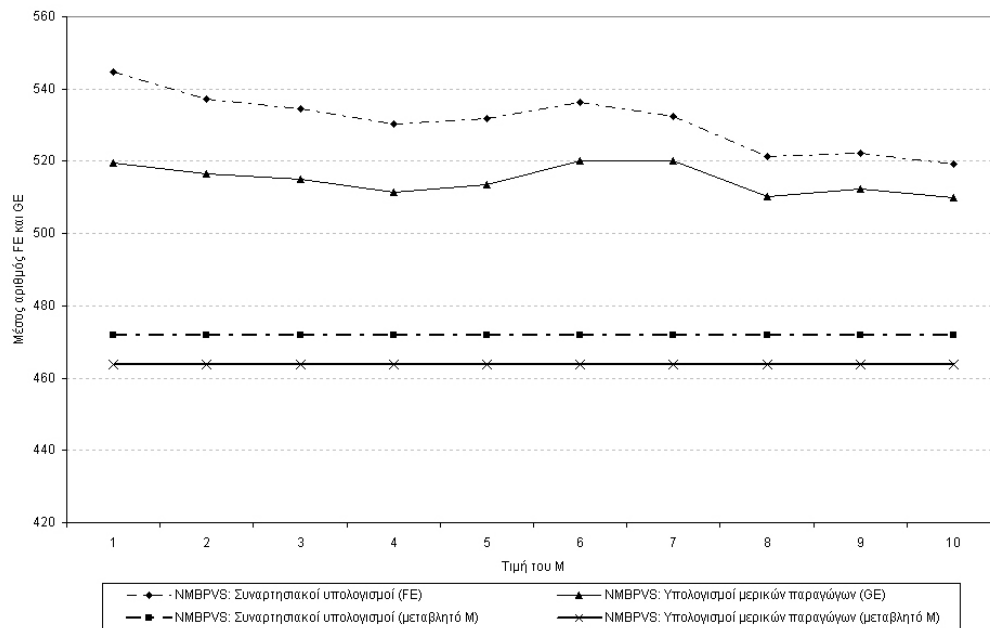
Σχήμα 8.6: Το πρόβλημα αναγνώρισης των κεφαλαίων γραμμάτων: Μέσος αριθμός υπολογισμών της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της, για τη μέθοδο NMBBP για διάφορες τιμές του M



Σχήμα 8.7: Το πρόβλημα ταξινόμησης υφής: Μέσος αριθμός υπολογισμών της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της, για τη μέθοδο NMBBP για διάφορες τιμές του M

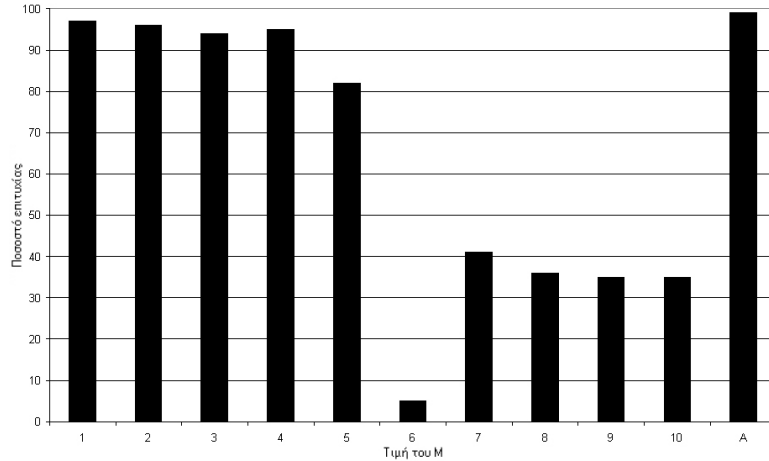


Σχήμα 8.8: Το πρόβλημα αναγνώρισης των κεφαλαίων γραμμάτων: Μέσος αριθμός υπολογισμών της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της, για τη μέθοδο NMBPVS για διάφορες τιμές του M



Σχήμα 8.9: Το πρόβλημα ταξινόμησης υφής: Μέσος αριθμός υπολογισμών της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της, για τη μέθοδο NMBPVS για διάφορες τιμές του M

Σχήμα 8.10. Στο Σχήμα αυτό, η σήμανση A στον οριζόντιο άξονα, δείχνει την περίπτωση όπου χρησιμοποιείται ο προσαρμοζόμενος οριζοντας εκπαίδευσης M . Γενικά, φαίνεται ότι η χρήση του προσαρμοζόμενου οριζοντα M , χωρίς να έχει πρόσθετο υπολογιστικό κόστος, βοηθά σημαντικά τη διαδικασία εκπαίδευσης.



Σχήμα 8.10: Το πρόβλημα ταξινόμησης υψής: Ποσοστό επιτυχίας για τη μέθοδο NMBBP για διάφορες τιμές του M

8.2.3 Μελέτη της επίδρασης της μη μονότονης στρατηγικής

Η μη μονότονη στρατηγική μπορεί να ενσωματωθεί σε *οποιοδήποτε* αλγόριθμο εκπαίδευσης. Παρακάτω, παρουσιάζουμε τα πειραματικά αποτελέσματα και συγκρίνουμε την απόδοση τριών αλγορίθμων εκπαίδευσης με τις μη μονότονες τροποποιήσεις τους. Σε όλες τις περιπτώσεις έχει χρησιμοποιηθεί ο προσαρμοζόμενος μη μονότονος οριζοντας εκμάθησης.

1) *Ο αλγόριθμος BPM*: Για να εξετάσουν την αποτελεσματικότητα της προτεινόμενης στρατηγικής εκπαίδευσης στην μη μονότονη παραλλαγή της BPM, έχουν εκτελεσθεί δύο πειράματα.

Το πρώτο πείραμα αναφέρεται στην εκπαίδευση ενός ΤΝΔ για την αναγνώριση των αριθμών από το 0 έως το 9. Οι λεπτομέρειες σχετικά με την αρχιτεκτονική του ΤΝΔ έχουν δοθεί παραπάνω. Για να αξιολογήσουν την σθεναρότητα της προτεινόμενης στρατηγικής, στις παραμέτρους εκπαίδευσης έχουν σκόπιμα δοθεί υψηλές (όχι βέλτιστες) τιμές, δηλαδή $\eta = 1.2$ και $m = 0.9$. Λεπτομερή αποτελέσματα σχετικά με την απόδοση των αλγορίθμων παρουσιάζονται στον Πίνακα 8.5, όπου μ είναι ο μέσος όρος των υπολογισμών των τιμών της συνάρτησης σφάλματος και του διανύσματος των μερικών παραγώγων της που απαιτούνται για την σύγκλιση, σ η αντίστοιχη τυπική απόκλιση, Min/Max ο ελάχιστος και ο μέγιστος αριθμός υπολογισμών τιμών και παραγώγων, και η *Επιτυχία* δείχνει το ποσοστό των προσομοιώσεων που συγκλίνουν σε ένα επιθυμητό ελάχιστο ($E \leq 0.1$).

Πίνακας 8.5: Αποτελέσματα από το πρόβλημα αναγνώρισης αριθμών

Αλγόριθμος	Υπολογισμοί μερικών παραγώγων			Συναρτησιακοί Υπολογισμοί			Επιτυχία %
	μ	σ	Min/Max	μ	σ	Min/Max	
BPM	560.2	684.9	239/3962	560.2	684.9	239/3962	39%
NMBPM	565.9	429.1	289/2823	571.6	428.9	295/2827	97%

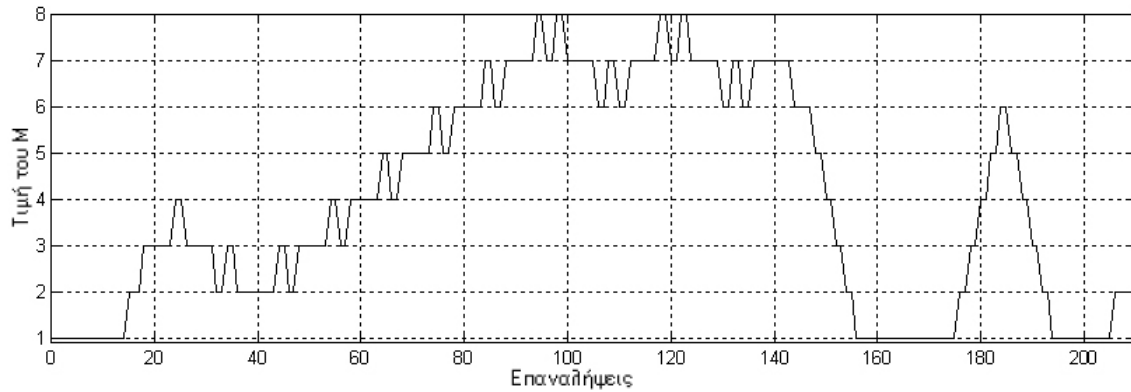
Το δεύτερο πρόβλημα που θα εξετάσουμε είναι το αποκλειστικό-ΕΙΤΕ (XOR). Για να λύσουμε αυτό το κλασικό πρόβλημα έχουμε χρησιμοποιήσει ένα ΤΝΔ με δύο κρυφούς νευρώνες

λογιστικών ενεργοποιήσεων και ένα γραμμικό νευρώνα εξόδου (6 βάρη και 3 πολώσεις). Τα βάρη αρχικοποιήθηκαν με τη μέθοδο Nguyen-Widrow [94] και οι παράμετροι εκπαίδευσης είχαν τις τιμές $\eta = 0.4$ και $m = 0.9$. Το κριτήριο τερματισμού ήταν $E \leq 0.1$, μέσα στο όριο των 2000 υπολογισμών της τιμής της συνάρτησης σφάλματος. Τα αποτελέσματα της προσομοίωσης παρουσιάζονται στον Πίνακα 8.6.

Πίνακας 8.6: Αποτελέσματα από το πρόβλημα του αποκλειστικού ΕΙΤΕ

Αλγόριθμος	Υπολογισμοί μερικών παραγώγων			Συναρτησιακοί Υπολογισμοί			Επιτυχία %
	μ	σ	Min/Max	μ	σ	Min/Max	
BPM	230.2	512.8	13/1764	230.2	512.8	13/1764	11%
NMBPM	187.8	365.1	16/1894	198.9	364.9	17/1903	80%

Η συμπεριφορά του M κατά τη διάρκεια μιας χαρακτηριστικής δοκιμής της μεθόδου NMBPM παρουσιάζεται στο Σχήμα 8.11.



Σχήμα 8.11: Το αποκλειστικό-ΕΙΤΕ: η συμπεριφορά της μεθόδου NMBPM με προσαρμοζόμενο M^k

Και στα δύο προβλήματα που ελέγξαμε, η χρήση της μη μονότονης στρατηγικής βελτιώνει σημαντικά το ποσοστό επιτυχίας της μεθόδου BPM. Εντούτοις, στο πρόβλημα αναγνώρισης αριθμών (βλ. τον Πίνακα 8.5), δεδομένου ότι λιγότερες δοκιμές έχουν συγκλίνει σε ένα επιθυμητό ελάχιστο για την μέθοδο BPM, ο αλγόριθμος φαίνεται να έχει μικρότερο μέσο αριθμό υπολογισμών τιμών και κλίσεων της συνάρτησης σφάλματος για τις δοκιμές που συνέκλιναν.

2) *Ο αλγόριθμος BPVS:* Για να παρουσιάσουμε την επίδραση της προτεινόμενης στρατηγικής εκμάθησης στην απόδοση της μεθόδου BPVS, την δοκιμάσαμε σε δύο εφαρμογές: (α) το πρόβλημα της ταξινόμησης υψών, και (β) το πρόβλημα της προσέγγισης μιας συνεχούς συνάρτησης (βλ. Παραρτήματα Α.8 και Α.5 αντίστοιχα). Τα αποτελέσματα παρατίθενται στους Πίνακες 8.7 και 8.8.

Πίνακας 8.7: Αποτελέσματα από το πρόβλημα ταξινόμησης υψών

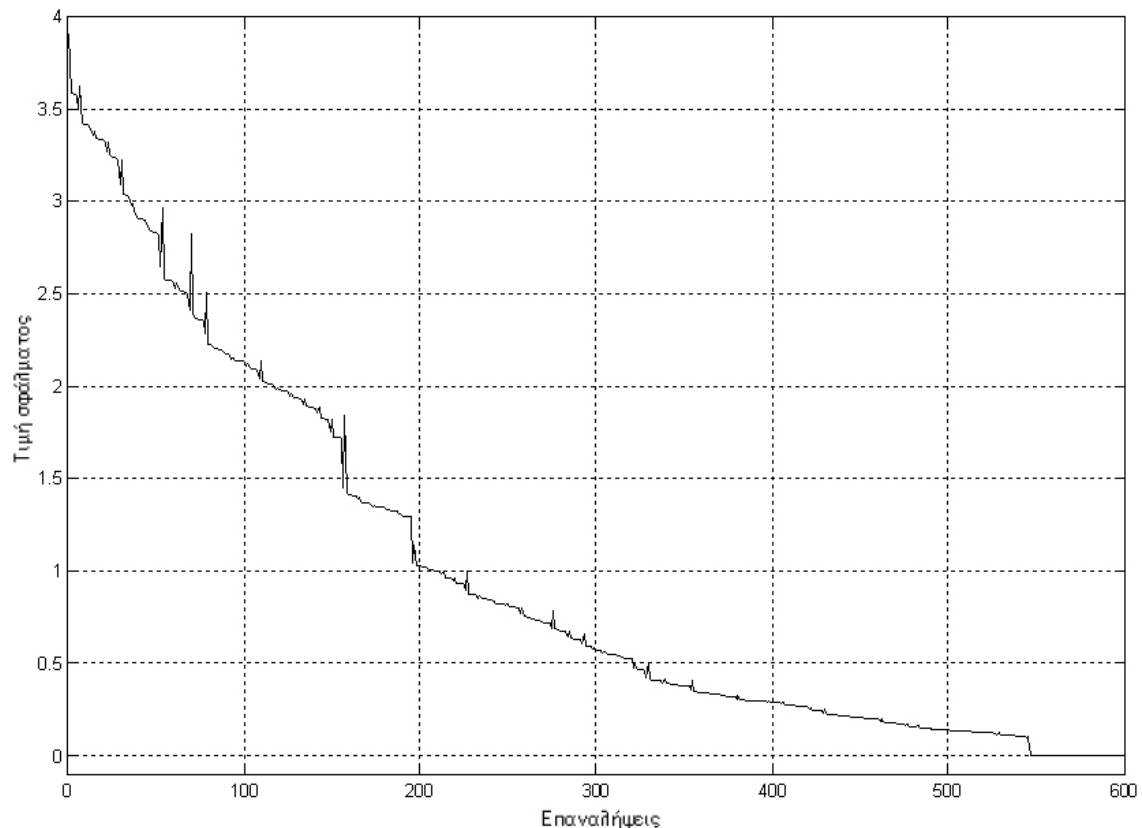
Αλγόριθμος	Υπολογισμοί μερικών παραγώγων			Συναρτησιακοί Υπολογισμοί			Επιτυχία %
	μ	σ	Min/Max	μ	σ	Min/Max	
BPVS	544.8	274.1	294/2227	519.5	257.7	283/2101	100%
NMBPVS	471.7	116.6	273/888	463.7	113.7	268/870	100%

Σύμφωνα με αυτά τα αποτελέσματα, η μη μονότονη στρατηγική βελτιώνει την απόδοση της μεθόδου BPVS. Η μείωση της συνάρτησης σφάλματος από τη μέθοδο NMBPVS σε

Πίνακας 8.8: Αποτελέσματα από το πρόβλημα της προσέγγισης μιας συνεχούς συνάρτησης

Αλγόριθμος	Υπολογισμοί μερικών παραγώγων			Συναρτησιακοί Υπολογισμοί			Επιτυχία %
	μ	σ	Min/Max	μ	σ	Min/Max	
BPVS	417.3	220.3	44/943	446.3	236.9	48/994	63.4%
NMBPVS	429.2	228.9	64/960	443.9	238.5	64/991	66.8%

μια χαρακτηριστική δοκιμή παρουσιάζεται στο Σχήμα 8.12, όπου διακρίνεται εύκολα η μη μονότονη συμπεριφορά σύγκλισης της μεθόδου NMBPVS.



Σχήμα 8.12: Το πρόβλημα αναγνώρισης των κεφαλαίων γραμμάτων: συμπεριφορά σύγκλισης της μεθόδου NMBPVS

3) *Ο αλγόριθμος BBP:* Στη συνέχεια δοκιμάσαμε τις προτεινόμενες στρατηγικές εκπαίδευσης στον αλγόριθμο BBP και στο πρόβλημα της προσέγγισης μιας συνεχούς συνάρτησης. Παραθέτουμε τα συγκριτικά αποτελέσματα στον Πίνακα 8.9. Υπάρχει μια αξιοπρόσεκτη βελτίωση της απόδοσης της μεθόδου BBP όταν εφαρμόζεται η μη μονότονη στρατηγική εκμάθησης. Συγκεκριμένα, η αρχική μέθοδος (BBP) έχει ένα ποσοστό επιτυχίας 79.6%, ενώ η μη μονότονη τροποποίηση (NMBBP) έχει ποσοστό επιτυχίας 92.2%.

Το δεύτερο πείραμα αφορά το πρόβλημα αναγνώρισης των κεφαλαίων γραμμάτων, που έχει περιγραφεί ανωτέρω. Τα αποτελέσματα δίνονται στον Πίνακα 8.10. Η μέθοδος BBP καθώς και η μέθοδος NMBBP έχουν επιτυχία 100%, εντούτοις η μη μονότονη τροποποίησή της είναι σαφώς γρηγορότερη.

Πίνακας 8.9: Αποτελέσματα από το πρόβλημα της προσέγγισης μιας συνεχούς συνάρτησης

Αλγόριθμος	Υπολογισμοί μερικών παραγώγων			Συναρτησιακοί Υπολογισμοί			Επιτυχία %
	μ	σ	Min/Max	μ	σ	Min/Max	
BBP	186.4	111.3	27/502	362.1	233.5	39/995	79.6%
NMBBP	158.2	114.7	26/625	241.7	195.1	29/988	92.2%

Πίνακας 8.10: Αποτελέσματα από το πρόβλημα αναγνώρισης των γραμμάτων

Αλγόριθμος	Υπολογισμοί μερικών παραγώγων			Συναρτησιακοί Υπολογισμοί			Επιτυχία %
	μ	σ	Min/Max	μ	σ	Min/Max	
BBP	169.8	35.9	90/373	332.6	70.4	167/758	100%
NMBBP	119.4	20.5	73/193	182.0	33.5	113/309	100%

8.2.4 Αποτελέσματα γενίκευσης

Η υψηλή ικανότητα γενίκευσης των εκπαιδευμένων ΤΝΔ είναι ένας αποφασιστικός παράγοντας για την επιλογή ενός αλγορίθμου εκπαίδευσης. Πρέπει λοιπόν να εξασφαλιστεί ότι τα αυξημένα ποσοστά σύγκλισης που έχουν οι μη μονότονοι αλγόριθμοι εκπαίδευσης δεν έχουν αρνητικές επιπτώσεις στην ικανότητα γενίκευσής τους. Για το λόγο αυτό, παρακάτω παρουσιάζουμε πειραματικά αποτελέσματα για την ικανότητα γενίκευσης των μεθόδων NMBPVS και NMBBP σε τέσσερα δύσκολα προβλήματα ελέγχου γενίκευσης. Τα αποτελέσματα αυτά δείχνουν ότι οι προτεινόμενες μέθοδοι έχουν καλύτερη γενίκευση από τους υπόλοιπους πολύ γνωστούς αλγόριθμους εκπαίδευσης που δοκιμάσαμε.

1) *Το πρόβλημα αναγνώρισης χειρόγραφων αριθμών:* Το πρόβλημα αναγνώρισης χειρόγραφων αριθμών [93] αποτελείται από ένα σύνολο 250 δειγμάτων από κάθε έναν από 44 ανεξάρτητους ανθρώπους. Συνολικά υπάρχουν 11000 πρότυπα, το σύνολο εκπαίδευσης αποτελείται από 7494 πρότυπα, ενώ τα υπόλοιπα χρησιμοποιούνται για τον έλεγχο της γενίκευσης. Πειραματικά βρέθηκε ότι η αρχιτεκτονική του ΤΝΔ με την καλύτερη απόδοση ήταν αυτή που είχε 50 νευρώνες στο κρυφό στρώμα. Έτσι χρησιμοποιήσαμε ένα 16-50-10 ΤΝΔ (1300 βάρη και 60 πολώσεις) με τους νευρώνες του κρυφού στρώματος να είναι βασισμένοι στη λογιστική συνάρτηση ενεργοποίησης, ενώ οι νευρώνες εξόδου να είναι γραμμικοί. Η εκπαίδευση σταμάτησε όταν το ΤΝΔ παρουσίασε 2% λαθεμένη ταξινόμηση στο σύνολο εκπαίδευσης.

Το μέσο ποσοστό επιτυχίας ταξινόμησης για κάθε αλγόριθμο στο πρόβλημα αυτό είναι: BP=91.0%, BPM=91.1%, ABP=92.0%, FR=96.8%, PR=97.1%, PR-FR=97.3%, NMBPVS=98.3%, και NMBBP=98.4%. Είναι φανερό ότι στο πρόβλημα αυτό οι μη μονότονες μέθοδοι έχουν πολύ υψηλή γενίκευση και υπερτερούν έναντι των υπολοίπων μεθόδων.

2) *Το πρόβλημα γενίκευσης MONK* [156]: Το πρόβλημα γενίκευσης MONK αποτελείται από 3 δύσκολα προβλήματα ρομποτικής, που σχεδιάστηκαν για την αξιολόγηση της ικανότητας γενίκευσης των αλγορίθμων εκπαίδευσης (βλ. και την Ενότητα 5.6 του Κεφαλαίου 5). Έχουμε δοκιμάσει τις μη μονότονες μεθόδους (NMBPVS και NMBBP) έναντι των πολύ γνωστών μεθόδων της οπισθοδρομικής διάδοσης του σφάλματος (BP) [133], της μεθόδου της οπισθοδρομικής διάδοσης του σφάλματος με εξασθένηση των βαρών (BP with Weight Decay – BPWD) [124], και της μεθόδου των διαδοχικών συσχετίσεων (Cascade Correlation – CC) [37].

Στην προσομοίωσή μας, έχουμε χρησιμοποιήσει τις ίδιες τοπολογίες δικτύων με εκείνες που βρίσκονται στην εργασία [156] για τη μέθοδο BP. Ο Πίνακας 8.11 σαφώς δείχνει ότι οι μέθοδοι NMBPVS και NMBBP έχουν άριστη γενίκευση, και κατορθώνουν να ταξινομήσουν σωστά όλα τα πρότυπα εισόδου σε όλα τα προβλήματα. Τα εκπαιδευμένα δίκτυα, σε όλα τα προβλήματα, φαίνεται να μπορούν να μαθαίνουν τις έννοιες που είναι ενσωματωμένες στα πρότυπα εισόδου. Αυτό είναι πιο εμφανές σε πρόβλημα MONK-3, όπου υπάρχει 5% εσκεμμένα λαθεμένη ταξινόμηση στα πρότυπα εισόδου. Τα δίκτυα που εκπαιδεύτηκαν με

τις μεθόδους BP, BPWD, και CC φαίνεται να αποτυγχάνουν να συλλάβουν τις έννοιες που ενσωματώνεται στα πρότυπα εισόδου, και μαθαίνουν τον θόρυβο αντ' αυτών.

Πίνακας 8.11: Αποτελέσματα από το πρόβλημα γενίκευσης MONK

Αλγόριθμος	MONK-1	MONK-2	MONK-3
BP	100%	100%	93.1%
BPWD	100%	100%	97.2%
CC	100%	100%	97.2%
NMBPVS	100%	100%	100%
NMBBP	100%	100%	100%

2) *Το πρόβλημα ταξινόμησης υφής:* Το πρόβλημα αυτό επιλέγεται, ως παράδειγμα πρακτικής εφαρμογής, για την σύγκριση και αξιολόγηση των αλγόριθμων εκπαίδευσης, χρησιμοποιώντας πραγματικά πρότυπα που περιέχουν τυχαίο θόρυβο.

Τα επιτυχώς εκπαιδευμένα ΤΝΔ εξετάζονται για την ικανότητα γενίκευσής τους χρησιμοποιώντας 20 τμήματα εικόνας μεγέθους 128×128 εικονοστοιχεία, που επιλέγονται τυχαία από κάθε εικόνα υφής. Για να αξιολογήσει τη μέση απόδοση γενίκευσης των ΤΝΔ ο κανόνας του μεγίστου (max rule) χρησιμοποιείται, δηλαδή ένα πρότυπο δοκιμής θεωρείται ότι έχει ταξινομηθεί σωστά εάν ο αντίστοιχος νευρώνας εξόδου έχει τη μέγιστη τιμή μεταξύ όλων των νευρώνων εξόδου.

Το μέσο ποσοστό επιτυχίας ταξινόμησης για κάθε αλγόριθμο στο πρόβλημα αυτό είναι: BP=90.0%, BPM=90.0%, ABP=93.5%, FR=92.0%, PR=92.6%, PR-FR=93.5%, NMBPVS=93.6%, και NMBBP=93.6%. Βλέπουμε και από αυτό το παράδειγμα ότι οι προτεινόμενες μη μονότονες μέθοδοι έχουν πολύ υψηλή γενίκευση.

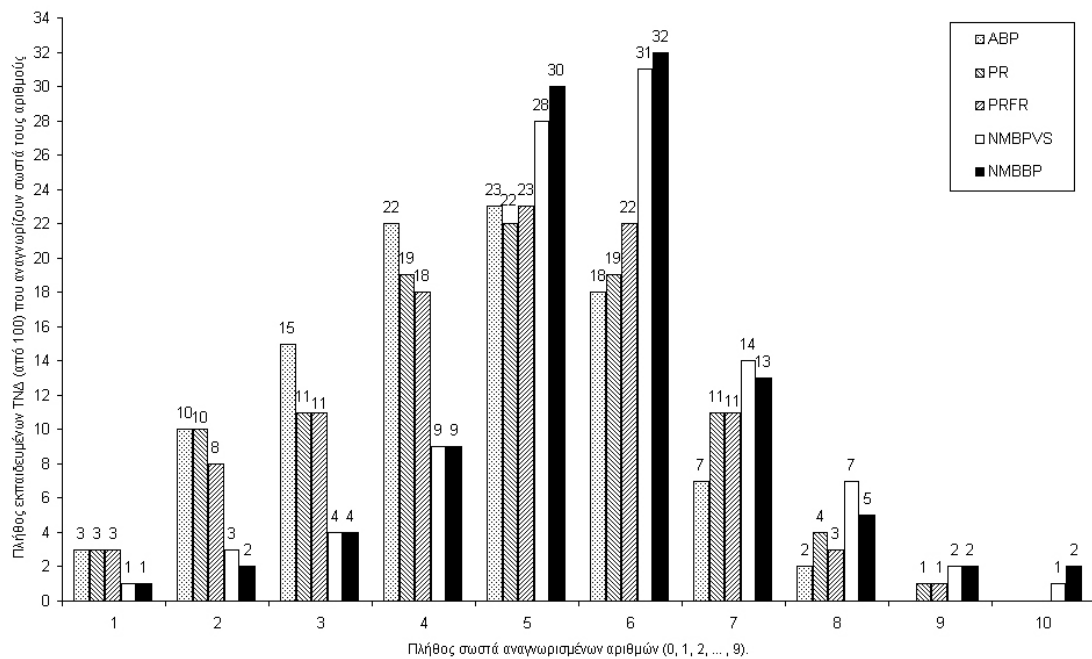
3) *Το πρόβλημα αναγνώρισης αριθμών:* Οι αριθμοί από το 0 έως το 9, από τη γραμματοσειρά helvetica, αποτελούν το σύνολο προτύπων εκπαίδευσης. Μετά από την φάση της εκπαίδευσης με τις μεθόδους BP, BPM, ABP, FR, PR, FR-PR, NMBPVS και NMBBP, τα ΤΝΔ εξετάζονται για την ικανότητα γενίκευσής τους χρησιμοποιώντας τους αριθμούς 0 έως 9 από τη γραμματοσειρά helvetica με πλάγιους χαρακτήρες (italics). Αξίζει να σημειωθεί, τα πρότυπα δοκιμής με πλάγιους χαρακτήρες αντιστρέφουν τις τιμές από 6 έως 14 bits σε σχέση με τα πρότυπα εκπαίδευσης και ότι 100 ΤΝΔ εκπαιδεύονται για κάθε περίπτωση. Για να αξιολογηθεί η μέση απόδοση γενίκευσης χρησιμοποιείται ο κανόνας του μεγίστου.

Τα ΤΝΔ που εκπαιδεύτηκαν με τις μεθόδους BP και BPM έχουν παρόμοιες ικανότητες γενίκευσης με αυτά που εκπαιδεύτηκαν με την μέθοδο ABP, αλλά περισσότερες επαναλήψεις ήταν απαραίτητες προκειμένου να συγκλίνουν. Το ίδιο ισχύει για τις μεθόδους FR και PR.

Για το λόγο αυτό στο Σχήμα 8.13 παρουσιάζουμε μόνο η απόδοση των μεθόδων ABP, PR, PR-FR, NMBPVS, και NMBBP. Όπως φαίνεται σε αυτό το σχήμα, οι μέθοδοι NMBPVS και NMBBP εκπαιδυσαν ΤΝΔ με την μεγαλύτερη ικανότητα γενίκευσης σε σχέση με όλους τους άλλους αλγόριθμους που δοκιμάσαμε. Είναι αξιοσημείωτο ότι μόνο οι μέθοδοι NMBPVS και NMBBP εκπαιδυσαν ΤΝΔ με ποσοστό επιτυχίας 100%.

8.3 Συμπεράσματα – Συνεισφορά

Σε αυτό το κεφάλαιο προτάθηκαν αιτιοκρατικές στρατηγικές μη μονότονης εκπαίδευσης για ΤΝΔ. Σύμφωνα με αυτήν την προσέγγιση, η τιμή της συνάρτησης σφάλματος πρέπει να ικανοποιεί ένα κριτήριο σχετικά με τη μέγιστη τιμή των προηγούμενων M επαναλήψεων, που αποτελεί και τον μη μονότονο ορίζοντα εκπαίδευσης.



Σχήμα 8.13: Το πρόβλημα αναγνώρισης των αριθμών: ο αριθμός των εκπαιδευμένων ΤΝΑ (από σύνολο 100), που ταξινομούν σωστά τους αριθμούς 0, 1, ..., 9

Επίσης, προτάθηκε μια διαδικασία για την κατάλληλη προσαρμογή του ορίζοντα εκπαίδευσης βασισμένη στην τοπική εκτίμηση της σταθεράς Lipschitz. Τα πειράματά μας δείχνουν ότι η χρήση ενός προσαρμοζόμενου M βοηθά στην μείωση του αριθμού υπολογισμών της τιμής της συνάρτησης σφάλματος και της κλίσης της, που απαιτούνται για τη σύγκλιση. Οι μη μονότονες στρατηγικές μπορούν να ενσωματωθούν σε *οποιοδήποτε* αλγόριθμο εκπαίδευσης ανά ομάδα προτύπων εισόδου, παρέχοντας σταθερή εκπαίδευση και, επομένως, μεγαλύτερη πιθανότητα επιτυχίας.

Τα αριθμητικά αποτελέσματά μας δείχνουν ότι η χρήση των στρατηγικών αυτών βελτιώνει την αποδοτικότητα και την αποτελεσματικότητα των μεθόδων εκπαίδευσης πρώτης τάξης, επιταχύνει σημαντικά τη σύγκλισή τους, και καταργεί την ανάγκη για ευρετικές παραμέτρους, που πιθανά εξαρτώνται από το εκάστοτε πρόβλημα. Επιπλέον, η μη μονότονη στρατηγική βελτίωσε ελαφρώς την μέση ικανότητα γενίκευσης των αλγορίθμων που δοκιμάσαμε.

Οι μη μονότονι αλγόριθμοι συγκρίθηκαν και με κάποιες πολύ γνωστές συζυγείς μεθόδους κλίσης, οι οποίες χρησιμοποιούν τεχνικές μη ακριβούς ευθύγραμμης (μονοδιάστατης) ανίχνευσης, για να εξασφαλίσουν την μονότονη μείωση του σφάλματος εκπαίδευσης.

Αξίζει να σημειωθεί ότι, σε ορισμένες περιπτώσεις, οι μη μονότονες μέθοδοι συγκλίνουν γρηγορότερα, ή εξίσου γρήγορα, με τις μεθόδους συζυγών κλίσεων. Τέλος, αποδεικνύονται σθεναρές ενάντια στις ταλαντώσεις λόγω των ακατάλληλων παραμέτρων εκπαίδευσης και μπορούν να χειριστούν επιτυχώς αυθαίρετα μεγάλους ρυθμούς εκπαίδευσης.

Μέρος V

**Συμπεράσματα – Παραρτήματα –
Βιβλιογραφία – Ευρετήριο**

Συμπεράσματα Διατριβής

Είναι σημαντικό να μην σταματάς να εξετάζεις.
Η περιέργεια έχει λόγο που υπάρχει.

—*Albert Einstein (1879-1955)*

Στο κεφάλαιο αυτό ανακεφαλαιώνουμε τις βασικές μεθόδους εκπαίδευσης ΤΝΔ που προτείνουμε, καθώς επίσης και τα βασικά συμπεράσματα που προέκυψαν από την χρήση τους.

Η συνεισφορά της παρούσας διατριβής επικεντρώνεται στη μελέτη και τη Μαθηματική θεμελίωση νέων μεθόδων εκπαίδευσης ΤΝΔ με επίβλεψη. Εκπαίδευση με επίβλεψη είναι η διαδικασία της προσαρμογής ενός ΤΝΔ έτσι ώστε να έχει συγκεκριμένη απόκριση σε συγκεκριμένες εισόδους. Έτσι η πραγματική έξοδος του ΤΝΔ συγκρίνεται με την επιθυμητή έξοδο και υπολογίζεται η διαφορά τους (σφάλμα). Στη συνέχεια τα βάρη του ΤΝΔ μεταβάλλονται με τέτοιο τρόπο ώστε στην επόμενη επανάληψη να μειωθεί η τιμή του σφάλματος.

Οι αλγόριθμοι εκπαίδευσης με επίβλεψη μπορούν να διαιρεθούν σε δύο βασικές κατηγορίες:

- αλγόριθμοι εκπαίδευσης ανά πρότυπο εισόδου (on-line ή stochastic training), και
- αλγόριθμοι εκπαίδευσης ανά ομάδα προτύπων εισόδου (batch ή off-line training).

Η εκπαίδευση ανά ομάδα προτύπων εισόδου είναι η κλασική προσέγγιση, όπου ένα σύνολο προτύπων λαμβάνεται και χρησιμοποιείται προκειμένου να εκπαιδευτεί το ΤΝΔ, προτού αυτό χρησιμοποιηθεί σε κάποια εφαρμογή. Αντίθετα, στην εκπαίδευση ανά πρότυπο εισόδου τα στοιχεία που συγκεντρώνονται κατά τη διάρκεια της κανονικής λειτουργίας του συστήματος χρησιμοποιούνται για την συνεχή εκπαίδευση και προσαρμογή του ΤΝΔ. Η εκπαίδευση ανά πρότυπο εισόδου μπορεί να επιλεχθεί για προβλήματα που έχουν είτε πολύ μεγάλο αριθμό προτύπων (πιθανά και κάποιο αριθμό περιττών ή λανθασμένων προτύπων), είτε όταν προσπαθούμε να προσεγγίσουμε ένα αργά μεταβαλλόμενο σύστημα και συχνά βοηθά στην αποφυγή τοπικών ελαχίστων. Δυστυχώς όμως πάσχει από κάποια μειονεκτήματα όπως για παράδειγμα η μεγάλη ευαισθησία στις παραμέτρους εκπαίδευσης. Το μεγαλύτερο μέρος της παρούσας διατριβής πραγματεύεται μεθόδους εκπαίδευσης ΤΝΔ ανά ομάδα προτύπων εισόδου.

Αρχίζουμε την παρουσίαση του ερευνητικού μέρους της διατριβής δίνοντας κάποια θεωρητικά αποτελέσματα σχετικά με την σύγκλιση μεθόδων εκπαίδευσης ΤΝΔ. Έτσι, στο Κεφάλαιο 3 προτείνουμε και αποδεικνύουμε ένα νέο θεωρητικό αποτέλεσμα που υποστηρίζει την ανάπτυξη αιτιοκρατικών αλγορίθμων εκπαίδευσης ΤΝΔ ευρείας σύγκλισης με τοπικούς ρυθμούς εκπαίδευσης. Το θεώρημα αυτό παρέχει σε *οποιαδήποτε* μέθοδο την ιδιότητα της ευρείας σύγκλισης, με την προϋπόθεση να ακολουθεί την προτεινόμενη τεχνική προσαρμογής της κατεύθυνσης ανίχνευσης και ρύθμισης του ρυθμού εκπαίδευσης. Οι νέες μέθοδοι, έχουν σημαντικά βελτιωμένα ποσοστά επιτυχίας και είναι ικανές να ανιχνεύσουν ελάχιστα

με μεγαλύτερη ακρίβεια. Τα αποτελέσματα δείχνουν ότι η στρατηγική του θεωρήματος έχει την αναμενόμενη συμπεριφορά και πρακτική εφαρμογή, αφού αυξάνει σημαντικά τη σθεναρότητα και το ποσοστό επιτυχίας.

Στο επόμενο κεφάλαιο δώσαμε ένα θεωρητικό αποτέλεσμα σχετικά με την σύγκλιση της γνωστής μεθόδου εκπαίδευσης ΤΝΔ QuickProp και προτείναμε μια τροποποίησή της, που δεν απαιτεί την ρύθμιση ευρετικών και εξαρτωμένων από το πρόβλημα παραμέτρων εκπαίδευσης. Επίσης, αποδείξαμε ένα νέο θεώρημα που εγγυάται την σύγκλιση της προτεινόμενης τροποποίησης. Η νέα αυτή μέθοδος παρουσιάζει ταχύτατη, ομαλή και σθεναρή σύγκλιση, με αποτέλεσμα να έχει μεγαλύτερα ποσοστά επιτυχίας. Τέλος, η αύξηση στην ταχύτητα σύγκλισης, δεν έχει αρνητικές συνέπειες στην γενίκευση των εκπαιδευμένων ΤΝΔ, αφού αυτά παρουσίασαν πολύ υψηλή ικανότητα γενίκευσης.

Τα θεωρητικά αυτά αποτελέσματα βοηθούν την δημιουργία νέων κλάσεων μεθόδων εκπαίδευσης ΤΝΔ, με πολύ καλές ιδιότητες σύγκλισης. Αυτό πιστοποιείται στην πράξη και από τις προσομοιώσεις μας και τις συγκρίσεις με άλλες γνωστές μεθόδους εκπαίδευσης.

Στα επόμενα δύο κεφάλαια μελετήσαμε την εκπαίδευση ΤΝΔ με μεθόδους ολικής Βελτιστοποίησης. Πιο συγκεκριμένα, στο Κεφάλαιο 5 προτείναμε και μελετήσαμε διεξοδικά μια νέα κλάση μεθόδων που είναι ικανές να εκπαιδεύσουν ΤΝΔ με περιορισμένα ακέραια βάρη με χρήση συναρτήσεων κατωφλίων και επεκτείναμε τις μεθόδους αυτές έτσι ώστε να υλοποιούνται και να εκτελούνται αποδοτικά από παράλληλες υπολογιστικές μηχανές με περισσότερους του ενός επεξεργαστές. Η χρήση κατωφλίων για όλους τους νευρώνες μειώνει κατά πολύ την πολυπλοκότητα της υλοποίησης σε υλικό, επειδή δεν υπάρχει ανάγκη να σχεδιαστούν και να εφαρμοστούν περίπλοκες μη-γραμμικές συναρτήσεις ενεργοποίησης. Ένα ακόμα βασικό χαρακτηριστικό των ΤΝΔ που χρησιμοποιούν κατώφλια είναι ότι η εκπαίδευση μπορεί να συνεχιστεί στο υλικό εάν το σύνολο των προτύπων έχει αλλάξει. Τέλος, ένα ακόμα πλεονέκτημα των ΤΝΔ με ακέραια βάρη και πολώσεις, καθώς και κατώφλια είναι ότι το εκπαιδευμένο ΤΝΔ μπορεί να είναι σε πολλές περιπτώσεις ανθεκτικό στο θόρυβο που περιέχεται στα πρότυπα εκπαίδευσης. Τέτοια δίκτυα είναι ικανά να συλλάβουν το βασικό χαρακτηριστικό γνώρισμα των προτύπων εκπαίδευσης (όπως φαίνεται και από τα αποτελέσματα γενίκευσης) και γενικά θόρυβος χαμηλής έντασης δεν μπορεί να διαταράξει τα ακέραια βάρη, αφού απαιτούνται σχετικά μεγάλες διακυμάνσεις, έτσι ώστε τα βάρη και οι πολώσεις να μετακινηθούν από μια ακέραια τιμή στην επόμενη ή στην προηγούμενη.

Το Κεφάλαιο 6 πραγματεύεται την εφαρμογή γνωστών μεθόδων Ολικής Βελτιστοποίησης, όπως οι Γενετικοί Αλγόριθμοι, η μέθοδος της προσομοιωμένης απόσβεσης (Simulated Annealing) και η μέθοδος βελτιστοποίησης με σμήνος σωματιδίων (Particle Swarm Optimization) στην εκπαίδευση ΤΝΔ. Οι μέθοδοι αυτές συχνά καταφέρνουν να αποφύγουν την σύγκλιση σε ανεπιθύμητα τοπικά ελάχιστα. Η αποφυγή τέτοιων ελαχίστων στην πράξη, δεν είναι πάντα εφικτή, αλλά υπάρχει μεγαλύτερη πιθανότητα να καταλήξουν σε μια αποδεκτή λύση και κατά αυτή την έννοια βελτιώνουν την διαδικασία εκπαίδευσης. Επίσης στο ίδιο κεφάλαιο, αναλύσαμε δύο μετασχηματισμούς της συνάρτησης σφάλματος, που έχουν σαν σκοπό την απαλοιφή τοπικών ελαχίστων της. Η τεχνική της παρεκκλίνουσας τροχιάς και η τεχνική του «εφελκυσμού» της αντικειμενικής συνάρτησης παρέχουν σταθερή σύγκλιση και πολύ μεγάλη πιθανότητα επιτυχίας. Τα πειράματά μας δείχνουν ότι αλγόριθμοι εκπαίδευσης που βοηθούνται από τους μετασχηματισμούς αυτούς είναι δυνατόν να ανακαλύπτουν επιθυμητά ελάχιστα με πολύ μεγαλύτερη πιθανότητα. Τα αποτελέσματα είναι ενθαρρυντικά και δείχνουν τη χρησιμότητα και εφαρμοσιμότητα όλων των τεχνικών που παρουσιάσαμε.

Στο Κεφάλαιο 7, μελετήσαμε μεθόδους εκπαίδευσης ανά πρότυπο εισόδου και προτείνουμε μια νέα τέτοια μέθοδο. Τα αποτελέσματα της προσομοίωσης δείχνουν ότι η προτεινόμενη μέθοδος παρέχει γρήγορη και σταθερή εκπαίδευση σε σύγκριση με άλλες μεθόδους εκπαίδευσης ανά πρότυπο εισόδου, καθώς επίσης και μεθόδους εκπαίδευσης ανά ομάδα προτύπων εισόδου. Συνεπώς παρέχει μεγαλύτερη πιθανότητα επιτυχημένης και γρήγορης εκπαίδευσης για προβλήματα του πραγματικού κόσμου. Επίσης, παρουσιάσαμε μια νέα

υβριδική μέθοδο και εξετάσαμε την απόδοσή της σε δύο πραγματικές εφαρμογές. Τα αποτελέσματα γενίκευσης των ΤΝΔ που αρχικά εκπαιδεύτηκαν από την προτεινόμενη μέθοδο εκπαίδευσης ανά πρότυπο εισόδου (Φάση 1) και στη συνέχεια εκπαιδεύτηκαν ξανά με ένα Εξελικτικό Αλγόριθμο (Φάση 2) είναι ικανοποιητικά και ανάλογα με τα καλύτερα αποτελέσματα μεθόδων που εκπαιδεύουν ανά ομάδα προτύπων εισόδου. Η προτεινόμενη υβριδική μέθοδος ανταποκρίθηκε με επιτυχία στο μη στατικό πρόβλημα της αναγνώρισης ανωμαλιών σε κολονοσκοπήσεις και αποδείχτηκε σθεναρή και προβλέψιμη. Τέλος, αξίζει να αναφερθεί ότι δεν παρατηρήθηκε το φαινόμενο της «καταστροφικής παρέμβασης» μεταξύ των προτύπων που προέρχονταν από διαφορετικές εικόνες.

Η παρουσίαση του ερευνητικού μέρους αυτής της διατριβής ολοκληρώνεται με το Κεφάλαιο 8, όπου προτείνουμε και μελετάμε αιτιολογικές στρατηγικές μη μονότονης εκπαίδευσης για ΤΝΔ, που μπορούν να ενσωματωθούν σε *οποιοδήποτε* αλγόριθμο εκπαίδευσης ανά ομάδα προτύπων εισόδου παρέχοντας σταθερότερη εκπαίδευση και μεγαλύτερη πιθανότητα επιτυχίας. Σύμφωνα με αυτήν την προσέγγιση, η τιμή της συνάρτησης σφάλματος πρέπει να ικανοποιεί ένα κριτήριο σχετικά με τη μέγιστη τιμή των προηγούμενων M επαναλήψεων, που αποτελεί και τον μη μονότονο ορίζοντα εκπαίδευσης. Επίσης, προτάθηκε μια διαδικασία για την κατάλληλη προσαρμογή του ορίζοντα εκπαίδευσης βασισμένη στην τοπική εκτίμηση της σταθεράς Lipschitz. Τα πειράματά μας δείχνουν ότι η χρήση ενός προσαρμοζόμενου ορίζοντα εκπαίδευσης βοηθά στην μείωση του αριθμού υπολογισμών της τιμής της συνάρτησης σφάλματος και της κλίσης της, που απαιτούνται για τη σύγκλιση. Η χρήση των στρατηγικών αυτών βελτιώνει την αποδοτικότητα και την αποτελεσματικότητα των μεθόδων εκπαίδευσης πρώτης τάξης, επιταχύνει σημαντικά τη σύγκλισή τους, και καταργεί την ανάγκη για ευρετικές παραμέτρους, που πιθανά εξαρτώνται από το εκάστοτε πρόβλημα. Αξίζει να σημειωθεί ότι, σε πολλές περιπτώσεις, οι μη μονότονες μέθοδοι συγκλίνουν γρηγορότερα από τις μονότονες μεθόδους που εξετάσαμε. Επιπλέον, βελτιώνουν ελαφρώς την μέση ικανότητα γενίκευσης των αλγορίθμων που δοκιμάσαμε. Τέλος, είναι σθεναρές ενάντια στις ταλαντώσεις λόγω των ακατάλληλων παραμέτρων εκπαίδευσης και μπορούν να χειριστούν επιτυχώς αυθαίρετα μεγάλους ρυθμούς εκπαίδευσης.

Προβλήματα Εκπαίδευσης Νευρωνικών Δικτύων

Στο παράρτημα αυτό θα παρουσιάσουμε μια σύντομη περιγραφή των διαφόρων προβλημάτων εκπαίδευσης ΤΝΔ που χρησιμοποιήθηκαν για την σύγκριση και την αξιολόγηση της ταχύτητας, της σθεναρότητας, της γενίκευσης και της αξιοπιστίας των αλγορίθμων που παρουσιάσαμε.

A.1 Αποκλειστικό-ΕΙΤΕ (XOR)

Το πρώτο πρόβλημα που θα περιγράψουμε είναι το αποκλειστικό-ΕΙΤΕ [133]. Αυτό είναι ένα πολύ γνωστό και δύσκολο πρόβλημα ταξινόμησης, που χρησιμοποιείται για την αξιολόγηση αλγορίθμων εκπαίδευσης και επιλογής ΤΝΔ. Το αποκλειστικό-ΕΙΤΕ απεικονίζει δύο δυαδικές εισόδους σε μία δυαδική έξοδο. Τα πρότυπα εκπαίδευσης για αυτό το πρόβλημα φαίνονται στον Πίνακα Α.1. Αξίζει να σημειωθεί ότι η δυαδική αυτή συνάρτηση δεν είναι γραμμικά διαχωρίσιμη, και συνεπώς απαιτεί την χρήση τουλάχιστον ενός κρυφού στρώματος. Τέλος, είναι εξαιρετικά ευαίσθητο στην επιλογή των αρχικών βαρών, στις διακυμάνσεις του ρυθμού εκπαίδευσης, και η επιφάνεια σφάλματος παρουσιάζει πολλά τοπικά ελάχιστα με σχετικά μεγάλες συναρτησιακές τιμές. Η αναλυτική έκφραση της αντικειμενικής συνάρτησης που πρέπει να ελαχιστοποιηθεί είναι η ακόλουθη:

$$\begin{aligned}
 f(\mathbf{x}) = & \left(1 + e^{-\frac{x_7}{1 + e^{-x_1 - x_2 - x_5}} - \frac{x_8}{1 + e^{-x_3 - x_4 - x_6}} - x_9} \right)^{-2} + \\
 & + \left(1 + e^{-\frac{x_7}{1 + e^{-x_5}} - \frac{x_8}{1 + e^{-x_6}} - x_9} \right)^{-2} + \\
 & + \left(1 - \left(1 + e^{-\frac{x_7}{1 + e^{-x_1 - x_5}} - \frac{x_8}{1 + e^{-x_3 - x_6}} - x_9} \right)^{-1} \right)^2 + \\
 & + \left(1 - \left(1 + e^{-\frac{x_7}{1 + e^{-x_2 - x_5}} - \frac{x_8}{1 + e^{-x_4 - x_6}} - x_9} \right)^{-1} \right)^2.
 \end{aligned}$$

A.2 Ισοτιμία των 3-bit

Το πρόβλημα της ισοτιμίας 3-bit, μπορεί να θεωρηθεί μια γενίκευση του προβλήματος του Αποκλειστικού-ΕΙΤΕ, αλλά είναι πιο δύσκολο. Το δίκτυο πρέπει να εκπαιδευτεί ώστε να υπολογίζει την πρόσθεση modulo 2, τριών εισόδων ή διαφορετικά να υπολογίζει την συνάρτηση της περιττής ισοτιμίας (odd parity function). Η επιφάνεια σφάλματος του προβλήματος αυτού

Πίνακας A.1: Τα πρότυπα εκπαίδευσης του αποκλειστικού-ΕΙΤΕ

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

είναι γνωστό ότι έχει πολλά τοπικά ελάχιστα. Τα 8 πρότυπα εκπαίδευσης παρουσιάζονται στον Πίνακα A.2.

Πίνακας A.2: Τα πρότυπα εκπαίδευσης της ισοτιμίας 3-bit

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

A.3 4-2-4 Κωδικοποιητής/Αποκωδικοποιητής

Σε αυτό το πρόβλημα το δίκτυο δέχεται 4 διαφορετικά πρότυπα εισόδου μήκους 4 bit, που το καθένα έχει μόνο ένα bit με την τιμή ένα (1) και τα υπόλοιπα με την τιμή μηδέν (0). Στον Πίνακα A.3 παρουσιάζονται τα πρότυπα εκπαίδευσης για αυτό το πρόβλημα. Το ζητούμενο είναι η έξοδος του δικτύου να είναι η ίδια με την είσοδο. Αφού η πληροφορία της εισόδου περνούν από το κρυφό στρώμα των νευρώνων, το δίκτυο πρέπει να κατασκευάσει μια μοναδική κωδικοποίηση για καθένα από τα 4 πρότυπα, στους 2 κρυφούς νευρώνες και βάρη τέτοια ώστε να κάνουν την κωδικοποίηση και την αποκωδικοποίηση δυνατή. Το πρόβλημα αυτό αν και φαίνεται απλό προσομοιάζει καλά πραγματικά προβλήματα ταξινόμησης προτύπων, όπου μικρές αλλαγές στις εισόδους έχουν σαν αποτέλεσμα μικρές αλλαγές στην έξοδο [35].

Πίνακας A.3: Τα πρότυπα εκπαίδευσης του 4-2-4 Κωδικοποιητή/Αποκωδικοποιητή

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)

Α.4 Το πρόβλημα γενίκευσης MONK

Το πρόβλημα γενίκευσης MONK αποτελείται από 3 δύσκολα προβλήματα ρομποτικής, που σχεδιάστηκαν για την αξιολόγηση της ικανότητας γενίκευσης των αλγορίθμων εκπαίδευσης. Συγκεκριμένα, σε κάθε πρόβλημα, ένα ρομπότ περιγράφεται από 6 διαφορετικές ιδιότητες. Το καθένα από αυτά τα προβλήματα ορίζεται από μια περιγραφή της κλάσης του, όπως φαίνεται εδώ:

MONK-1: ($Ιδιότητα_1 == Ιδιότητα_2$) ΕΙΤΕ ($Ιδιότητα_5 == 1$).

124 πρότυπα έχουν επιλεγεί τυχαία μέσα από το σύνολο εκπαίδευσης, ενώ τα υπόλοιπα 308 χρησιμοποιούνται για τον έλεγχο της γενίκευσης. Σε αυτό το πρόβλημα δεν υπάρχει εσκεμμένη λαθεμένη ταξινόμηση.

MONK-2: ($Ακριβώς\ δύο\ ιδιότητες == 1$).

Το πρόβλημα αυτό μοιάζει με το πρόβλημα Α.2. Συγκεκριμένα, 169 πρότυπα αποτελούν το σύνολο εκπαίδευσης, ενώ τα υπόλοιπα αποτελούν το σύνολο ελέγχου της γενίκευσης. Και εδώ δεν υπάρχει εσκεμμένη λαθεμένη ταξινόμηση.

MONK-3: ($Ιδιότητα_5 == 3$ ΚΑΙ $Ιδιότητα_4 == 1$) ΕΙΤΕ ($Ιδιότητα_5 \neq 4$ ΚΑΙ $Ιδιότητα_2 \neq 3$).

Στο πρόβλημα αυτό υπάρχει 5% εσκεμμένα λαθεμένη ταξινόμηση στο σύνολο εκπαίδευσης, το οποίο αποτελείται από 122 πρότυπα. Τα υπόλοιπα 310 πρότυπα χρησιμοποιούνται για τον έλεγχο της γενίκευσης του εκπαιδευμένου δικτύου.

Κάθε μια από τις 6 ιδιότητες μπορεί να έχει 3, 3, 2, 3, 4, και 2 τιμές αντίστοιχα. Έτσι υπάρχουν 432 δυνατοί συνδυασμοί οι οποίοι και αποτελούν το σύνολο δεδομένων (περισσότερες πληροφορίες σχετικά με τις ιδιότητες αυτές υπάρχουν στην εργασία [156]). Τέλος, σε κάθε δυνατή τιμή κάθε ιδιότητας έχουμε και μια είσοδο του δικτύου που αντιστοιχεί σε αυτή. Συνεπώς, το δίκτυο έχει 17 εισόδους ($3 + 3 + 2 + 3 + 4 + 2 = 17$).

Α.5 Προσέγγιση μιας συνεχούς συνάρτησης

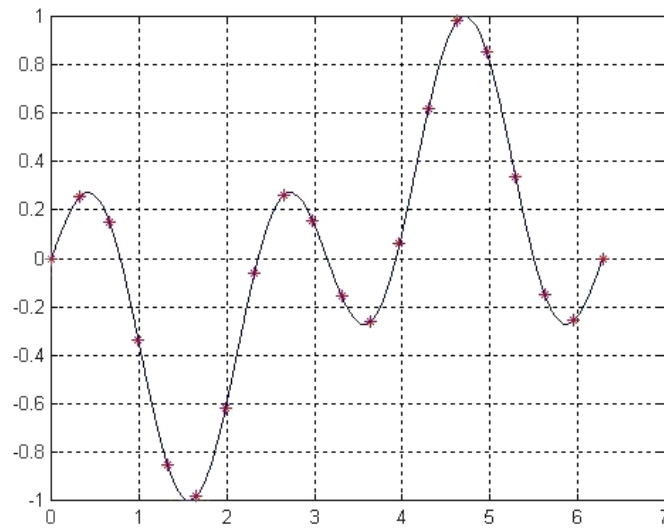
Το πρόβλημα προσέγγισης μιας (γενικά άγνωστης) συνάρτησης, χρησιμοποιώντας μόνο κάποιες γνωστές τιμές της, είναι ένα από τα βασικά προβλήματα που καλούνται να επιλύσουν τα ΤΝΔ. Για το συγκεκριμένο πρόβλημα επιλέξαμε την συνεχή συνάρτηση $f(x) = \sin(x) \cos(2x)$ [146]. Το σύνολο των προτύπων εισόδου αποτελείται από τιμές της συνάρτησης σε ισαπέχοντα (συνήθως 20) σημεία, στο διάστημα $[0, 2\pi]$. Στο Σχήμα Α.1 απεικονίζεται αυτή η συνάρτηση και τα 20 σημεία που αποτελούν τα πρότυπα εκπαίδευσης.

Α.6 Αναγνώριση των κεφαλαίων γραμμάτων

Για το πρόβλημα της αναγνώρισης των γραμμάτων, 26 πίνακες που έχουν τα κεφαλαία γράμματα της Αγγλικής αλφαβήτου, αποτελούν το σύνολο των προτύπων εισόδου. Κάθε γράμμα ορίζεται με δυαδικές τιμές σε ένα πλέγμα 5×7 . Στον Πίνακα Α.4 και στο Σχήμα Α.2, φαίνεται η κωδικοποίηση του κεφαλαίου γράμματος Άλφα.

Α.7 Αναγνώριση αριθμών

Ο σκοπός αυτού του προβλήματος είναι το δίκτυο να εκπαιδευτεί να αναγνωρίζει εκτυπωμένους αριθμούς από το 0 έως το 9 [82, 148]. Κάθε αριθμός ορίζεται από ένα 8×8 πίνακα. Στον Πίνακα Α.5 και στο Σχήμα Α.3, φαίνεται η κωδικοποίηση του αριθμού 6.



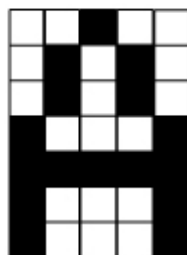
Σχήμα A.1: Γραφική παράσταση της συνάρτησης $f(x) = \sin(x) \cos(2x)$ και τα 20 πρότυπα εκπαίδευσης

Πίνακας A.4: Ο δυαδικός πίνακας κωδικοποίησης για το γράμμα Άλφα

0	0	1	0	0
0	1	0	1	0
0	1	0	1	0
1	0	0	0	1
1	1	1	1	1
1	0	0	0	1
1	0	0	0	1

A.8 Ταξινόμηση υφής

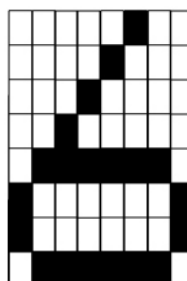
Για το πρόβλημα της ταξινόμησης υφής 12 εικόνες διαφορετικής υφής (Brodatz texture images) [19] μεγέθους 512×512 σαρώνονται σε ανάλυση 150dpi. Από κάθε εικόνα υφής, 10 υποπεριοχές της μεγέθους 128×128 επιλέγονται τυχαία και σε αυτές εφαρμόζουμε την μέθοδο εξαγωγής προτύπων που πρότεινε ο Haralick [46]. Τελικά, από κάθε εικόνα εξάγονται 10 πρότυπα εκπαίδευσης, με 16 στοιχεία το καθένα. Έτσι τα πρότυπα παρουσιάζονται ως μια πεπερασμένη σειρά $C = (c_1, c_2, \dots, c_p)$ από ζεύγη εισόδου-εξόδου $c_p = (u_p, t_p)$, όπου u_p είναι τα πρότυπα εισόδου που ανήκουν στο $\mathbb{R}^{16}$, και t_p είναι οι επιθυμητές έξοδοι που ανήκουν στο $\{0, 1\}^{12}$, για $p = 1, \dots, 120$. Στο πρόβλημα αυτό το ΤΝΔ πρέπει να εκπαιδευτεί να ξεχωρίζει κάθε υφή (ξύλο, ύφασμα κτλ.) και να ταξινομεί σωστά (γνωστά και άγνωστα) πρότυπα εισόδου. Στην Εικόνα A.4, φαίνονται οι 12 εικόνες που χρησιμοποιήθηκαν για την εξαγωγή των προτύπων εκπαίδευσης.



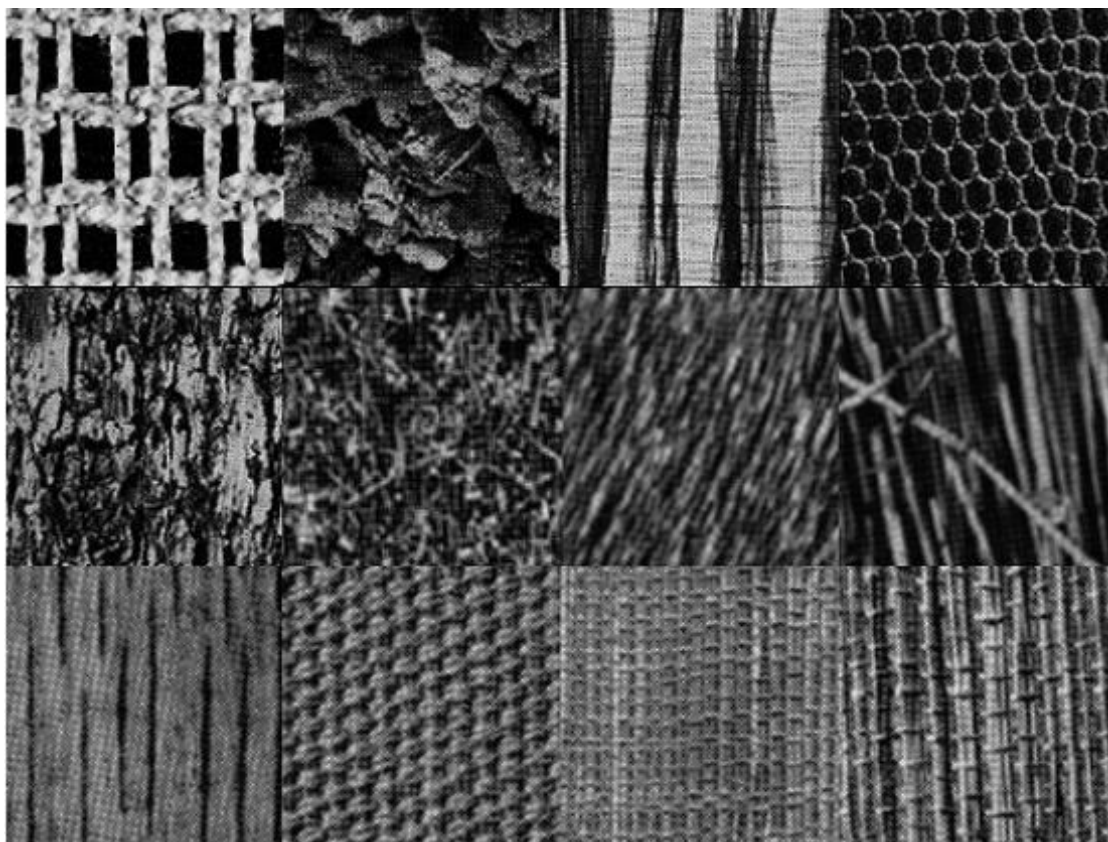
Σχήμα Α.2: Η κωδικοποίηση για το γράμμα Άλφα

Πίνακας Α.5: Ο δυαδικός πίνακας κωδικοποίησης για τον αριθμό 6

0	0	0	0	1	0	0
0	0	0	1	0	0	0
0	0	0	1	0	0	0
0	0	1	0	0	0	0
0	1	1	1	1	1	0
1	0	0	0	0	0	1
1	0	0	0	0	0	1
0	1	1	1	1	1	0



Σχήμα Α.3: Η κωδικοποίηση για τον αριθμό 6



Σχήμα A.4: Οι 12 εικόνες υφής που χρησιμοποιήθηκαν για την εξαγωγή των προτύπων εκπαίδευσης

Απόδειξη της Μεθόδου Οπισθοδρομικής Διάδοσης του Σφάλματος

Στο παράρτημα αυτό θα δώσουμε την απόδειξη της μεθόδου της οπισθοδρομικής διάδοσης του σφάλματος (Back Propagation – BP) [133] για τον υπολογισμό του διανύσματος των μερικών παραγώγων της συνάρτησης σφάλματος E ενός ΤΝΔ με ένα κρυφό επίπεδο. Υποθέτουμε ότι οι συναρτήσεις ενεργοποίησης που χρησιμοποιούνται είναι παραγωγίσιμες. Έστω ότι συμβολίζουμε με X_I τον I -στο νευρώνα εισόδου, με Z_J τον J -στο κρυφό νευρώνα, και με Y_K τον K -στο νευρώνα εξόδου. Συμβολίζουμε w_{JK} το βάρος της σύνδεσης από τον Z_J στον νευρώνα Y_K και v_{IJ} το βάρος από τον X_I στον Z_J .

Τότε αν συμβολίζουμε x_I το πρότυπο εκπαίδευσης του X_I νευρώνα εισόδου, η είσοδος του Z_J κρυφού νευρώνα θα είναι:

$$z_{inJ} = \sum_i x_i v_{iJ},$$

και η έξοδος του θα είναι:

$$z_J = f_1(z_{inJ}),$$

όπου f_1 είναι η συνάρτηση ενεργοποίησης για τους νευρώνες του κρυφού επιπέδου.

Ομοίως, ο νευρώνας εξόδου Y_K θα έχει είσοδο y_{inK} και ενεργοποίηση y_K , που δίνονται από τους ακόλουθους τύπους:

$$y_{inK} = \sum_j z_j w_{jK},$$

$$y_K = f_2(y_{inK}),$$

όπου f_2 είναι η συνάρτηση ενεργοποίησης των νευρώνων εξόδου. Συνεπώς, η συνάρτηση σφάλματος E του ΤΝΔ θα δίνεται από τον τύπο:

$$E = \frac{1}{2} \sum_k (t_k - y_k)^2.$$

Αφού δώσαμε τον συμβολισμό που θα ακολουθήσουμε, ξεκινάμε υπολογίζοντας την μερική παράγωγο:

$$\begin{aligned} \frac{\partial E}{\partial w_{JK}} &= \frac{\partial}{\partial w_{JK}} \frac{1}{2} \sum_k (t_k - y_k)^2 = \\ &= \frac{1}{2} \frac{\partial}{\partial w_{JK}} (t_K - y_K)^2 = \\ &= -(t_K - y_K) \frac{\partial}{\partial w_{JK}} f_2(y_{inK}) = \\ &= -(t_K - y_K) f_2'(y_{inK}) z_J, \end{aligned}$$

και αν θέσουμε:

$$\delta_K = (t_K - y_K) f'_2(y_{in_K}),$$

έχουμε τελικά:

$$\frac{\partial E}{\partial w_{JK}} = -\delta_K z_J. \quad (B.1)$$

Με παρόμοιο τρόπο υπολογίζουμε και την μερική παράγωγο:

$$\begin{aligned} \frac{\partial E}{\partial v_{IJ}} &= \frac{\partial}{\partial v_{IJ}} \frac{1}{2} \sum_k (t_k - y_k)^2 = \\ &= - \sum_k (t_k - y_k) \frac{\partial}{\partial v_{IJ}} y_k = \\ &= - \sum_k (t_k - y_k) f'_2(y_{in_K}) \frac{\partial}{\partial v_{IJ}} y_{in_K} = \\ &= - \sum_k \delta_K w_{Jk} \frac{\partial}{\partial v_{IJ}} z_J = \\ &= - \sum_k \delta_K w_{Jk} \frac{\partial}{\partial v_{IJ}} f_1(z_{in_J}) = \\ &= - \sum_k \delta_K w_{Jk} f'_1(z_{in_J}) x_I, \end{aligned}$$

και αν θέσουμε:

$$\delta_J = \sum_k \delta_K w_{Jk} f'_1(z_{in_J}),$$

έχουμε τελικά:

$$\frac{\partial E}{\partial v_{IJ}} = -\delta_J x_J. \quad (B.2)$$

Τελικά, χρησιμοποιώντας τις Σχέσεις (B.1) και (B.2) υπολογίζουμε το διάνυσμα των μερικών παραγώγων της συνάρτησης σφάλματος, ∇E .

Βιβλιογραφία

- [1] “The beowulf project”, <http://www.beowulf.org>, last accessed 01/09/2002.
- [2] D.J. Albers, J.C. Sprott, and W.D. Dechert, “Routes to chaos in neural networks with random weights”, *International Journal of Bifurcation and Chaos*, vol. 8, 1463–1478, 1998.
- [3] L.B. Almeida, T. Langlois, J.D. Amaral, and A. Plankhov, “Parameter adaptation in stochastic optimization”, In *On-line Learning in Neural Networks*, edited by D. Saad, Cambridge University Press, 111–134, 1998.
- [4] J. Anderson, “A simple neural network generating an interactive memory”, *Mathematical Biosciences*, vol. 14, 197–220, 1972.
- [5] J.A. Anderson and E. Rosenfeld, *Neurocomputing: Foundations of research*, MIT Press, Cambridge, 1989.
- [6] P. Angeline, “Tracking extrema in dynamic environments”, In the proceedings of the *Sixth Annual conference on Evolutionary Programming VI*, Springer, 335–345, 1997.
- [7] L. Armijo, “Minimization of functions having lipschitz-continuous first partial derivatives”, *Pacific Journal of Mathematics*, vol. 16, 1–3, 1966.
- [8] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, New York, 1996.
- [9] S.W. Baik and P. Pachowicz, “Adaptive object recognition based on the radial basis function paradigm”, In the proceedings of the *IEEE International Joint Conference on Neural Networks (IJCNN’99)*, CD-ROM Proceedings, Paper No.215, Session 9.4, Washington, U.S.A., 1999.
- [10] S. Baluja, “Evolution of an artificial neural network based autonomous land vehicle controller”, *IEEE Transactions on System, Man and Cybernetics-Part B*, vol. 26, 450–463, 1996.
- [11] A.G. Barto, “Reinforcement learning and adaptive critic methods”, In *Handbook of Intelligent Control*, edited by D. White and D. Sofge, chap. 12, Van Nostrand Reinhold, New York, 1992.
- [12] J. Barzilai and J.M. Borwein, “Two point step size gradient methods”, *IMA Journal of Numerical Analysis*, vol. 8, 141–148, 1988.
- [13] R. Battiti, “Accelerated backpropagation learning: two optimization methods”, *Complex Systems*, vol. 3, 331–342, 1989.
- [14] R. Battiti, “First- and second-order methods for learning: between steepest descent and newton’s method”, *Neural Computation*, vol. 4, 141–166, 1992.
- [15] S. Becker and Y. Le Cun, “Improving the convergence of the back-propagation learning with second order methods”, In the proceedings of the *1988 Connectionist Models Summer School*, edited by D. Touretzky, G. Hinton, and T. Sejnowski, Morgan Kaufmann, San Mateo, 29–37, 1988.

- [16] A.L. Blum and R. Rivest, "Training a 3-node neural network is np-complete", *Neural Networks*, vol. 5, 117-127, 1992.
- [17] E.K. Blum, "Approximation of boolean functions by sigmoidal networks: Part i: Xor and other two variable functions", *Neural Computation*, vol. 1, 532-540, 1989.
- [18] B. Boutsinas and M.N. Vrahatis, "Artificial nonmonotonic neural networks", *Artificial Intelligence*, vol. 132, 1-38, 2001.
- [19] P. Brodatz, *Textures - a photographic album for artists and designers*, Dover, New York, 1966.
- [20] C.G. Broyden, "A class of methods for solving nonlinear simultaneous equations", *Math. Comp.*, vol. 19, 577-593, 1965.
- [21] M. Burton, Jr. and G.J. Mpitsos, "Event dependent control of noise enhances learning in neural networks", *Neural Networks*, vol. 5, 627-637, 1992.
- [22] A. Cauchy, "Méthode générale pour la résolution des systèmes d'équations simultanées", *Comp. Rend. Acad. Sci. Paris*, vol. 25, 536-538, 1847.
- [23] L.W. Chan and F. Fallside, "An adaptive training algorithm for back-propagation networks", *Computers Speech and Language*, vol. 2, 205-218, 1987.
- [24] P.S. Churchland, *Neurophilosophy: Toward a unified science of the mind/brain*, MIT Press, Cambridge, 1986.
- [25] L. Coetzee and E.C. Botha, "An analysis of coarse-grain parallel training of a neural net", *Network: Computation in Neural Systems*, vol. 6, 73-91, 1995.
- [26] E.M. Corwin, A.M. Logar, and W.J.B. Oldham, "An iterative method for training multilayer networks with threshold functions", *IEEE Transactions on Neural Networks*, vol. 5, 507-508, 1994.
- [27] J.E. Dennis, Jr. and J. Moré, "A characterization of superlinear convergence and its applications to quasi newton methods", *Math. Comp.*, vol. 28, 577-593, 1974.
- [28] J.E. Dennis, Jr. and J.J. Moré, "Quasi-newton methods, motivation and theory", *SIAM Review*, vol. 19, 46-89, 1977.
- [29] J.E. Dennis, Jr. and R.B. Schnabel., "A view of unconstrained optimization", In *Handbooks in OR & MS, vol. 1*, edited by G. N. et al., Elsevier Science Publishers B.V., North-Holland, 1989.
- [30] J.E. Dennis, Jr. and R.B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, SIAM, Philadelphia, 1996, Originally published: Prentice Hall, New Jersey, 1983.
- [31] A.D. Doulamis, N.D. Doulamis, and S.D. Kollias, "On-line retrainable neural networks: improving the performance of neural networks in image analysis problems", *IEEE Transactions on Neural Networks*, vol. 11, 137-155, 2000.
- [32] R.C. Eberhart and Y.H. Shi, "Evolving artificial neural networks", In the proceedings of the *International Conference on Neural Networks and Brain*, Beijing, P.R. China, 1998.
- [33] R.C. Eberhart, P.K. Simpson, and R.W. Dobbins, *Computational intelligence PC tools*, Academic Press Professional, Boston, 1996.
- [34] J.L. Elman, E.A. Bates, M.H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett, *The shape of change*, chap. 6, MIT Press, Cambridge, Massachusetts, 1997.
- [35] S.E. Fahlman, "An empirical study of learning speed in back-propagation networks", Technical Report CMU-CS-88-162, Carnegie Mellon University, Pittsburgh, PA 15213, September 1988.

- [36] S.E. Fahlman, "Faster-learning variations on back-propagation: an empirical study", In the proceedings of the *1988 Connectionist Models Summer School*, edited by D. Touretzky, G. Hinton, and T. Sejnowski, Morgan Kaufmann, San Mateo, 38-51, 1988.
- [37] S.E. Fahlman and C. Lebiere, "The cascade-correlation learning architecture", Technical Report CMU-CS-90-100, Carnegie Mellon University, Pittsburgh, PA 15213, February 1990.
- [38] A.V. Fiacco and G.P. McCormick, *Nonlinear programming: Sequential unconstrained minimization techniques*, 1990, Philadelphia, SIAM.
- [39] A. Geist, A. Beguelin, J. Dongarra, W. Jiang, R. Manchek, and V. Sunderam, *PVM: Parallel Virtual Machine. A User's Guide and Tutorial for Networked Parallel Computing*, MIT Press, Cambridge, 1994.
- [40] J.C. Gilbert and J. Nocedal, "Global convergence properties of conjugate gradient methods for optimization", *SIAM Journal Optimization*, vol. 2, 21-42, 1992.
- [41] A.A. Goldstein, "On steepest descent", *SIAM Journal of Control*, vol. 3, 147-151, 1965.
- [42] M. Gori and A. Tesi, "On the problem of local minima in backpropagation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, 76-85, 1992.
- [43] L. Grippo, F. Lampariello, and S. Lucidi, "A nonmonotone line search technique for newton's method", *SIAM Journal on Numerical Analysis*, vol. 23, 707-716, 1986.
- [44] S. Grossberg, "Some physiological and biochemical consequences of psychological postulates", In the proceedings of the *National Academy of Sciences*, vol. 60, 758-765, 1968.
- [45] M.T. Hagan, H.B. Demuth, and M. Beale, *Neural network design*, PWS Publishing Company, Boston, 1996.
- [46] R. Haralick, K. Shanmugan, and I. Dinstein, "Textural features for image classification", *IEEE Transactions on System, Man and Cybernetics*, vol. 3, 610-621, 1973.
- [47] S. Haykin, *Neural networks: a comprehensive foundation*, Macmillan College Publishing Company, New York, 1994.
- [48] D.O. Hebb, *The organization of behavior*, Wiley, New York, 1949.
- [49] R. Hecht-Nielsen, "Kolmogorov's mapping neural network existence theorem", In the proceedings of the *IEEE First International Conference on Neural Networks*, vol. III, 11-14, 1987.
- [50] M.E. Hohil, D. Liu, and S.H. Smith, "Solving the n-bit parity problem using neural networks", *Neural Networks*, vol. 12, 1321-1323, 1999.
- [51] J.H. Holland, *Adaptation in neural and artificial systems*, University of Michigan Press, 1975.
- [52] J. Holt and J. Hwang, "Finite precision error analysis of neural network hardware implementations", *IEEE Transactions on Computers*, vol. 42, 281-290, 1993.
- [53] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators", *Neural Networks*, vol. 2, 359-366, 1989.
- [54] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks", *Neural Networks*, vol. 3, 551-560, 1990.

- [55] C.R. Houck, J.A. Joines, and M.G. Kay, "A genetic algorithm for function optimization: a matlab implementation", Technical Report NCSU-IE TR, 95-09, North Carolina State University, 1995.
- [56] D.R. Hush, B. Horne, and J.M. Salas, "Error surfaces for multi-layer perceptrons", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, 1152-1161, 1992.
- [57] R.A. Jacobs, "Increased rates of convergence through learning rate adaptation", *Neural Networks*, vol. 1, 295-307, 1988.
- [58] S. Karkanis, G.D. Magoulas, and N. Theofanous, "Image recognition and neuronal networks: intelligent systems for the improvement of imaging information", In the proceedings of the *Minimally Invasive Therapy and Allied Technologies*, 225-230, 2000.
- [59] N.A. Katsanos, E. Iliopoulou, V.P. Plagianakos, and H. Mangou, "Interrelations between absorption energies and local isotherms, local monolayer capabilities, and energy distribution functions, as determined for heterogeneous surfaces by inverse gas chromatography", *Journal of Colloid and Interface Science*, vol. 239, 10-19, 2001.
- [60] N.A. Katsanos, F. Roubani-Kalantzopoulou, E. Iliopoulou, I. Bassiotis, V. Siokos, M.N. Vrahatis, and V.P. Plagianakos, "Lateral molecular interactions on heterogeneous surfaces experimentally measured", *Colloid and Surfaces A*, vol. 210, 173-180, 2002.
- [61] J. Kennedy and R.C. Eberhart, "Particle swarm optimization", In the proceedings of the *IEEE International Conference on Neural Networks*, Piscataway, NJ, IV:1942-1948, 1995.
- [62] S. Kirkpatrick, C.D. Gelatt Jr., and M.P. Vecchi, "Optimization by simulated annealing", *Science*, vol. 220, 671-680, 1983.
- [63] T. Kohonen, "Correlation matrix memories", *IEEE Transactions on Computers*, vol. 21, 353-359, 1972.
- [64] T. Kohonen, *Self-organization and associative memory*, Springer-Verlag, Berlin, 1987.
- [65] A.N. Kolmogorov, "On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition", *Doklady Akademii Nauk SSSR*, vol. 144, 679-681, 1957, (American Mathematical Society Translation, vol. 28, 55-59).
- [66] C.M. Kuan and K. Hornik, "Convergence of learning algorithms with constant learning rates", *IEEE Transactions on Neural Networks*, vol. 2, 484-488, 1991.
- [67] S. Kudo, *Early colorectal cancer*, Igaku-Shoin Publishers, Tokyo, 1996.
- [68] Y. Lee, S.H. Oh, and M.W. Kim, "An analysis of premature saturation in backpropagation learning", *Neural Networks*, vol. 6, 719-728, 1993.
- [69] C. Leopold, *Parallel and Distributed Computing: A Survey of Models, Paradigms and Approaches*, John Wiley and Sons, 2000.
- [70] A. Levy, A. Montalvo, S. Gomez, and A. Galderon, *Topics in Global Optimization, Lecture Notes in Mathematics No. 909*, Springer-Verlag, New York, 1981.
- [71] R. Liu, G. Dong, and X. Ling, "A convergence analysis for neural networks with constant learning rates and non-stationary inputs", In the proceedings of the *34th Conference on Decision and Control*, New Orleans, 1278-1283, 1995.

- [72] C.G. Looney, *Pattern recognition using neural networks*, Oxford University Press, New York, 1997.
- [73] G.D. Magoulas, V.P. Plagianakos, G.S. Androulakis, and M.N. Vrahatis, "A framework for the development of globally convergent adaptive learning rate algorithms", *International Journal of Computer Research*, vol. 1, 1-10, 2001.
- [74] G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "Effective neural network training with a different learning rate for each weight", In the proceedings of the *6th IEEE International Conference on Electronics, Circuits and Systems (ICECS '99)*, Pafos, Cyprus, 591-594, 1999.
- [75] G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "Sign-methods for training with imprecise error function and gradient values", In the proceedings of the *IEEE International Joint Conference on Neural Networks (IJCNN'99)*, CD-ROM Proceedings, Paper No.2020, Session 5.1, Washington, U.S.A., 1999.
- [76] G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "Development and convergence analysis of training algorithms with local learning rate adaptation", In the proceedings of the *IEEE International Joint Conference on Neural Networks (IJCNN'2000)*, Como, Italy, 2000.
- [77] G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "Adaptive stepsize algorithms for on-line training of neural networks", *Nonlinear Analysis, Theory, Methods and Applications*, vol. 47, 3425-3430, 2001.
- [78] G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "Hybrid methods using evolutionary algorithms for on-line training", In the proceedings of the *INNS-IEEE International Joint Conference on Neural Networks (IJCNN'2001)*, Washington D.C., U.S.A., 2001.
- [79] G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "Improved neural network-based interpretation of colonoscopy images through on-line learning and evolution", In the proceedings of the *EUNITE 2001 Conference*, Tenerife, Spain, 2001.
- [80] G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "On globally convergent algorithms with local learning rates", *IEEE Transactions on Neural Networks*, vol. 13, 774-779, 2002.
- [81] G.D. Magoulas, M.N. Vrahatis, and G.S. Androulakis, "A new method in neural network supervised training with imprecision", In the proceedings of the *IEEE 3rd International Conference on Electronics, Circuits and Systems*, 287-290, 1996.
- [82] G.D. Magoulas, M.N. Vrahatis, and G.S. Androulakis, "Effective back-propagation with variable stepsize", *Neural Networks*, vol. 10, 69-82, 1997.
- [83] G.D. Magoulas, M.N. Vrahatis, and G.S. Androulakis, "On the alleviation of local minima in backpropagation", *Nonlinear Analysis, Theory, Methods and Applications*, vol. 30, 4545-4550, 1997.
- [84] G.D. Magoulas, M.N. Vrahatis, and G.S. Androulakis, "Improving the convergence of the back-propagation algorithm using learning rate adaptation methods", *Neural Computation*, vol. 11, 1769-1796, 1999.
- [85] G.D. Magoulas, M.N. Vrahatis, T.N. Grapsa, and G.S. Androulakis, "Neural network supervised training based on a dimension reducing method", In *Mathematics of Neural Networks: Models, Algorithms and Applications*, edited by S. Ellacot, J. Mason, and I. Anderson, Kluwer, 245-249, 1997.
- [86] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity", *Bulletin of Mathematical Biophysics*, vol. 5, 115-133, 1943.

- [87] G.C. Meletiou, D.K. Tasoulis, and M.N. Vrahatis, "A first study of the neural network approach to the RSA cryptosystem", In the proceedings of the *6th IASTED International Conference on Artificial Intelligence and Soft Computing*, Alberta, Canada, 483-488, 2002.
- [88] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equations of state calculations by fast computing machines", *Journal of Chemical Physics*, vol. 21, 1087-1092, 1953.
- [89] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer-Verlag, 1996.
- [90] Z. Michalewicz and D.B. Fogel, *How to solve it: Modern heuristics*, Springer, Berlin, 2000.
- [91] M. Minsky and S. Papert, *Perceptrons*, MIT Press, Cambridge, 1969.
- [92] M.F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning", *Neural Networks*, vol. 6, 525-533, 1993.
- [93] P.M. Murphy and D.W. Aha, "UCI repository of machine learning databases", 1994, <http://www.ics.uci.edu/~mllearn/MLRepository.html>, last accessed 01/09/2002.
- [94] D. Nguyen and B. Widrow, "Improving the learning speed of 2-layer neural network by choosing initial values of the adaptive weights", In the proceedings of the *IEEE First International Joint Conference on Neural Networks*, vol. 3, 21-26, 1990.
- [95] J. Nocedal, "Theory of algorithms for unconstrained optimization", In *Acta Numerica*, 199-242, 1992.
- [96] J.M. Ortega and W.C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York, 1970.
- [97] P.W. Pachowicz and S.W. Baik, "Adaptive rbf classifier for object recognition in images sequences", In the proceedings of the *IEEE International Joint Conference on Neural Networks (IJCNN'2000)*, vol. VI-600, Como, Italy, 2000.
- [98] N.G. Panagiotidis, D. Kalogeras, S.D. Kollias, and A. Stafylopatis, "Neural network-assisted effective lossy compression of medical images", *Proc. IEEE*, vol. 84, 1474-1487, 1996.
- [99] D.B. Parker, "Optimal algorithms for adaptive networks: Second order back-propagation, second order direct propagation, and second order hebbian learning", In the proceedings of the *IEEE International Conference on Neural Networks*, vol. 2, 593-600, 1987.
- [100] K.E. Parsopoulos, V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Improving the particle swarm optimizer by function stretching", In *Advances in Convex Analysis and Global Optimization, Honoring the memory of C. Caratheodory (1873-1950)*, edited by N. Hadjisavvas and P. Pardalos, chap. 28, Kluwer Academic Publishers, Dordrecht, The Netherlands, 445-457, 2001.
- [101] K.E. Parsopoulos, V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Objective function "stretching" to alleviate convergence to local minima", *Nonlinear Analysis, Theory, Methods and Applications*, vol. 47, 3419-3424, 2001.
- [102] K.E. Parsopoulos, V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Stretching technique for obtaining global minimizers through particle swarm optimization", In the proceedings of the *Particle Swarm Optimization Workshop*, Indianapolis, U.S.A., 22-29, 2001.

- [103] M. Pfister and R. Rojas, "Speeding-up backpropagation – a comparison of orthogonal techniques", In the proceedings of the *Joint Conference on Neural Networks*, Nagoya, Japan, 517–523, 1993.
- [104] V.P. Plagianakos, G.D. Magoulas, N.K. Nouis, and M.N. Vrahatis, "Pvm-based training of large neural architectures", In the proceedings of the *INNS-IEEE International Joint Conference on Neural Networks (IJCNN'2001)*, Washington D.C., U.S.A., 2001.
- [105] V.P. Plagianakos, G.D. Magoulas, N.K. Nouis, and M.N. Vrahatis, "Training multi-layer networks with discrete activation functions", In the proceedings of the *INNS-IEEE International Joint Conference on Neural Networks (IJCNN'2001)*, Washington D.C., U.S.A., 2001.
- [106] V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Nonmonotone learning rules for backpropagation networks", In the proceedings of the *6th IEEE International Conference on Electronics, Circuits and Systems (ICECS '99)*, Pafos, Cyprus, 291–294, 1999.
- [107] V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Optimization strategies and backpropagation neural networks", In the proceedings of the *Seventh Hellenic Conference on Informatics*, Ioannina, Greece, 1999.
- [108] V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Global learning rate adaptation in on-line neural network training", In the proceedings of the *Second International ICSC Symposium on Neural Computation (NC'2000)*, Berlin, Germany, 2000.
- [109] V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Learning in multilayer perceptrons using global optimization strategies", *Nonlinear Analysis, Theory, Methods and Applications*, vol. 47, 3431–3436, 2001.
- [110] V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Learning rate adaptation in stochastic gradient descent", In *Advances in Convex Analysis and Global Optimization, Honoring the memory of C. Caratheodory (1873-1950)*, edited by N. Hadjisavvas and P. Pardalos, chap. 27, Kluwer Academic Publishers, Dordrecht, The Netherlands, 433–444, 2001.
- [111] V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Supervised trained using global search methods", In *Advances in Convex Analysis and Global Optimization, Honoring the memory of C. Caratheodory (1873-1950)*, edited by N. Hadjisavvas and P. Pardalos, chap. 26, Kluwer Academic Publishers, Dordrecht, The Netherlands, 421–432, 2001.
- [112] V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Tumor detection in colonoscopic images using hybrid methods for on-line neural network training", In the proceedings of the *Neural Networks and Expert Systems in Medicine and HealthCare*, Milos Island, Greece, 2001.
- [113] V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Deterministic nonmonotone strategies for effective training of multi-layer perceptrons", *IEEE Transactions on Neural Networks*, 2002, in press.
- [114] V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Improved learning of neural nets through global search", In *Global Optimization – Selected Case Studies*, edited by J. Pintér, Kluwer Academic Publishers, 2002, in press.
- [115] V.P. Plagianakos, N.K. Nouis, and M.N. Vrahatis, "Locating and computing in parallel all the simple roots of special functions using pvm", *Journal of Computational and Applied Mathematics*, vol. 133, 545–554, 2001.

- [116] V.P. Plagianakos, D.G. Sotiropoulos, and M.N. Vrahatis, "An improved backpropagation method with adaptive learning rate", In *Recent Advances in circuits and systems*, edited by N. Mastorakis, World Scientific, 1998.
- [117] V.P. Plagianakos, D.G. Sotiropoulos, and M.N. Vrahatis, "Integer weight training by differential evolution algorithms", In *Recent Advances in circuits and systems*, edited by N. Mastorakis, World Scientific, 1998.
- [118] V.P. Plagianakos and E. Tzanaki, "Chaotic analysis of seismic time series and short term forecasting using neural networks", In the proceedings of the *INNS-IEEE International Joint Conference on Neural Networks (IJCNN'2001)*, Washington D.C., U.S.A., 2001.
- [119] V.P. Plagianakos and M.N. Vrahatis, "Neural network training with constrained integer weights", In the proceedings of the *Congress of Evolutionary Computation (CEC'99)*, edited by M. S. X. Y. P.J. Angeline, Z. Michalewicz and A. Zalala, IEEE Press, Washington D.C., U.S.A., 2007-2013, 1999.
- [120] V.P. Plagianakos and M.N. Vrahatis, "Training neural networks with 3-bit integer weights", In the proceedings of the *Genetic and Evolutionary Computation Conference (GECCO'99)*, edited by W. Banzhaf, J. Daida, A. Eiben, M. Garzon, V. Honavar, M. Jakiela, and R. Smith, Morgan Kaufmann, Orlando, U.S.A., 910-915, 1999.
- [121] V.P. Plagianakos and M.N. Vrahatis, "Training neural networks with threshold activation functions and constrained integer weights", In the proceedings of the *IEEE International Joint Conference on Neural Networks (IJCNN'2000)*, Como, Italy, 2000.
- [122] V.P. Plagianakos and M.N. Vrahatis, "Parallel evolutionary training algorithms for 'hardware-friendly' neural networks", *Natural Computing*, vol. 1, 307-322, 2002.
- [123] V.P. Plagianakos, M.N. Vrahatis, and G.D. Magoulas, "Nonmonotone methods for backpropagation training with adaptive learning rate", In the proceedings of the *IEEE International Joint Conference on Neural Networks (IJCNN'99)*, CD-ROM Proceedings, Paper No.2001, Session 5.1, Washington, U.S.A., 1999.
- [124] D.C. Plaut, S.J. Nowlan, and G.E. Hinton, "Experiments on learning by back propagation", Technical Report CMU-CS-86-126, Carnegie Mellon University, Pittsburgh, PA 15213, 1986.
- [125] E. Polak, *Optimization: algorithms and consistent approximations*, Springer-Verlag, New York, 1997.
- [126] M.J.D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives", *Computer Journal*, vol. 7, 155-162, 1964.
- [127] M. Raydan, "The barzilai and borwein gradient method for the large scale unconstrained minimization problem", *SIAM Journal on Optimization*, vol. 7, 26-33, 1997.
- [128] R.D. Reed and R.J. Marks II, *Neural smithing: Supervised learning in feedforward artificial neural networks*, MIT Press, Cambridge, 1999.
- [129] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: the rprop algorithm", In the proceedings of the *IEEE International Conference on Neural Networks*, San Francisco, 586-591, 1993.
- [130] A.K. Rigler, J.M. Irvine, and T.P. Vogl, "Rescaling of variables in backpropagation learning", *Neural Networks*, vol. 4, 225-229, 1991.
- [131] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain", *Psychological Review*, vol. 65, 386-408, 1958.

- [132] F. Roubani-Kalantzopoulou, T. Artemiadi, N.A. Katsanos, and V.P. Plagianakos, "Time separation of absorption sites on heterogeneous surfaces by inverse gas chromatography", *Chromatographia*, vol. 53, 315–320, 2001.
- [133] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning internal representations by error propagation", In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, edited by D. Rumelhart and J. McClelland, MIT Press, Cambridge, Massachusetts, 318–362, 1986.
- [134] D.E. Rumelhart and J.L. McClelland, editors, *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1, Cambridge, Massachusetts, MIT Press, 1986.
- [135] A.P. Russo, "Neural networks for sonar signal processing", In the proceedings of the *IEEE Conference on Neural Networks for Ocean Engineering*, Washington D.C., 1991.
- [136] D. Saad, *On-line learning in neural networks*, Cambridge University Press, 1998.
- [137] S. Saarinen, R. Bramley, and G. Cybenko, "Neural networks, back-propagation and automatic differentiation", In *Automatic differentiation of algorithms: theory, implementation and application*, edited by A. Griewank and G. Gorliss, SIAM, Philadelphia, 31–42, 1992.
- [138] S. Saarinen, R. Bramley, and G. Cybenko, "Ill-conditioning in neural network training problems", *SIAM Journal of Scientific Computing*, vol. 14, 693–714, 1993.
- [139] R. Salomon and P. Eggenberger, "Adaptation on the evolutionary time scale: a working hypothesis and basic experiments", In *Third European Conference on Artificial Evolution (AE'97)*, *Lecture Notes in Computer Science vol. 1363*, Springer, Nimes, France, 1998.
- [140] N.N. Schraudolph, "Online local gain adaptation for multi-layer perceptrons", Technical Report IDSIA-09-98, IDSIA, Lugano, Switzerland, 1998.
- [141] N.N. Schraudolph, "Local gain adaptation in stochastic gradient descend", Technical Report IDSIA-09-99, IDSIA, Lugano, Switzerland, 1999.
- [142] H.P. Schwefel, *Evolution and Optimum Seeking*, John Wiley & Sons, New York, 1995.
- [143] G.M. Sepherd and C. Koch, "Introduction to synaptic circuits", In *The Synaptic Organization of the Brain*, edited by G. Sepherd, Oxford University Press, New York, 3–31, 1990.
- [144] K. Sikorski, "Bisection is optimal", *Numerische Mathematik*, vol. 40, 111–117, 1982.
- [145] F. Silva and L. Almeida, "Lecture notes in computer science", vol. 412, chap. Acceleration techniques for the back-propagation algorithm, Springer-Verlag, Berlin, 110–119, 1990.
- [146] P.P. Van der Smagt, "Minimisation methods for training feedforward neural networks", *Neural Networks*, vol. 7, 1–11, 1994.
- [147] D.G. Sotiropoulos, V.P. Plagianakos, and M.N. Vrahatis, "An evolutionary algorithm for minimizing multimodal functions", In the proceedings of the *Hellenic European Research on Computer Mathematics and its Applications Conference (HERCMA'2001)*, Athens, Greece, 2001.
- [148] A. Sperduti and A. Starita, "Speed up learning and network optimization with extended back-propagation", *Neural Networks*, vol. 6, 365–383, 1993.

- [149] D.A. Sprecher, "On the structure of continuous functions of several variables", *Transactions of the American Mathematical Society*, vol. 115, 340-355, 1965.
- [150] T.L. Sterling, J. Salmon, D.J. Becker, and D.F. Savarese, *How to build a Beowulf: A Guide to Implementation and Application of PC Clusters*, MIT Press, Cambridge, 1999.
- [151] G.W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
- [152] R. Storn, "System design by constraint adaptation and differential evolution", *IEEE Transactions on Evolutionary Computation*, vol. 3, 22-34, 1999.
- [153] R. Storn and K. Price, "Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces", *Journal of Global Optimization*, vol. 11, 341-359, 1997.
- [154] R.S. Sutton, "Adapting bias by gradient descent: an incremental version of delta-bar-delta", In the proceedings of the *Tenth National Conference on Artificial Intelligence*, MIT Press, 171-176, 1992.
- [155] R.S. Sutton and S.D. Whitehead, "Online learning with random representations", In the proceedings of the *Tenth International Conference on Machine Learning*, Morgan Kaufmann, 314-321, 1993.
- [156] S.B. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. De Jong, S. Dzeroski, S.E. Fahlman, D. Fisher, R. Hamann, K. Kaufmann, S. Keller, I. Kononenko, J. Kreuziger, R.S. Michalski, T. Mitchell, P. Pachowicz, Y. Reich, H. Vafaie, W. Vande Welde, W. Wenzel, J. Wnek, and J. Zhang, "The monk's problems: A performance comparison of different learning algorithms", Technical Report CMU-CS-91-197, Carnegie Mellon University, 1991.
- [157] R. Varga, *Matrix Iterative Analysis, Second Edition*, Springer-Verlag, Berlin, 2000.
- [158] F. Vavak and T.C. Fogarty, "A comparative study of steady state and generational genetic algorithms", In *Evolutionary Computing: AISB Workshop, Lecture Notes in Computer Science vol. 1143*, Springer, 1996.
- [159] 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*, vol. 59, 257-263, 1988.
- [160] R.G. Voigt, "Rates of convergence for a class of iterative procedures", *SIAM Journal of Numerical Analysis*, vol. 8, 127-134, 1971.
- [161] M.N. Vrahatis, "Solving systems of nonlinear equations using the nonzero value of the topological degree", *ACM Transactions Mathematical Software*, vol. 14, 312-329, 1988.
- [162] M.N. Vrahatis, G.S. Androulakis, J.N. Lambrinos, and G.D. Magoulas, "A class of gradient unconstrained minimization algorithms with adaptive stepsize", *Journal of Computational and Applied Mathematics*, vol. 114, 367-386, 2000.
- [163] M.N. Vrahatis, B. Boutsinas, P. Alevizos, and G. Pavlides, "The new k -windows algorithm for improving the k -means clustering algorithm", *Journal of Complexity*, vol. 18, 375-391, 2002.
- [164] M.N. Vrahatis and K.I. Iordanidis, "A rapid generalized method of bisection for solving systems of nonlinear equations", *Numerische Mathematik*, vol. 49, 123-138, 1986.

- [165] M.N. Vrahatis, G.D. Magoulas, and V.P. Plagianakos, "Convergence analysis of the quickprop method", In the proceedings of the *IEEE International Joint Conference on Neural Networks (IJCNN'99)*, CD-ROM Proceedings, Paper No.848, Session 5.3, Washington, U.S.A., 1999.
- [166] M.N. Vrahatis, G.D. Magoulas, and V.P. Plagianakos, "Globally convergent modification of the quickprop method", *Neural Processing Letters*, vol. 12, 159-170, 2000.
- [167] M.N. Vrahatis, G.D. Magoulas, and V.P. Plagianakos, "Neural network supervised training as an optimization problem", In *Dynamical Systems*, edited by A. Bountis and S. Pneumatikos, vol. 6, Pneumatikos publications, Athens, 243-262, 2000.
- [168] M.N. Vrahatis, G.D. Magoulas, and V.P. Plagianakos, "From linear to nonlinear iterative methods", *Applied Numerical Mathematics*, 2002, in press.
- [169] R.L. Watrous, "Learning algorithms for connectionist networks: applied gradient of nonlinear optimization", In the proceedings of the *IEEE International Conference on Neural Networks*, vol. 2, 619-627, 1987.
- [170] S.T. Weslstead, *Neural network and fuzzy logic applications in C/C++*, Wiley, 1994.
- [171] B. Widrow and M.E. Hoff, "Adaptive switching circuits", In the proceedings of the *1960 IRE WESCON Convention Record*, New York, 96-104, 1960.
- [172] J. Wilkinson, *Rounding errors in algebraic processes*, Prentice-Hall, Englewood Cliffs, NJ, 1963.
- [173] G.G. Wilknsn, "Open questions in neurocomputing for earth observation", In the proceedings of the *First COMPARES Workshop*, York, U.K., 1996.
- [174] P. Wolfe, "Convergence conditions for ascent methods", *SIAM Review*, vol. 11, 226-235, 1969.
- [175] P. Wolfe, "Convergence conditions for ascent methods ii: Some corrections", *SIAM Review*, vol. 13, 185-188, 1971.
- [176] J. Wray and G. Green, "Neural networks, approximation theory and finite precision computation", *Neural Networks*, vol. 8, 31-37, 1995.
- [177] D. Young, "Iterative methods for solving partial difference equations of elliptic type", *Transactions of the American Mathematical Society*, vol. 76, 92-111, 1954.
- [178] X. H. Yu and G. A. Chen, "On the local minima free condition of backpropagation learning", *IEEE Transactions on Neural Networks*, vol. 6, 1300-1303, 1995.
- [179] W.I. Zangwill, "Minimizing a function without calculating derivatives", *Computer Journal*, vol. 10, 293-296, 1967.
- [180] G. Zoutendijk, "Nonlinear programming, computational methods", In *Integer and Nonlinear Programming*, edited by J. Abadie, North-Holland, Amsterdam, 37-86, 1970.
- [181] M. Zurada, *Artificial Neural Systems*, West Publishing, St. Paul, 1992.

Κατάλογος Δημοσιεύσεων Υποψηφίου

A. Δημοσιεύσεις σε Επιστημονικά Περιοδικά με Σύστημα Κριτών.

1. M.N. Vrahatis, G.D. Magoulas, and V.P. Plagianakos, "Globally Convergent Modification of the Quickprop Method", *Neural Processing Letters*, Vol. 12, No.2, 159-170, (2000).
2. V.P. Plagianakos, N.K. Nosis, and M.N. Vrahatis, "Locating and Computing in Parallel all the Simple Roots of Special Functions Using PVM", *Journal of Computational and Applied Mathematics*, Vol. 133, 545-554, (2001).
3. G.D. Magoulas, V.P. Plagianakos, G.S. Androulakis and M.N. Vrahatis, "A Framework for the Development of Globally Convergent Adaptive Learning Rate Algorithms", *International Journal of Computer Research*, Vol. 10, No. 1, 1-10, (2001).
4. G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "Adaptive stepsize algorithms for on-line training of neural networks", *Nonlinear Analysis, Theory, Methods and Applications*, Vol. 47, 3425-3430, (2001).
5. K.E. Parsopoulos, V.P. Plagianakos, G.D. Magoulas and M.N. Vrahatis, "Objective function "stretching" to alleviate convergence to local minima", *Nonlinear Analysis, Theory, Methods and Applications*, Vol. 47, 3419-3424, (2001).
6. V.P. Plagianakos, G.D. Magoulas and M.N. Vrahatis, "Learning in multilayer perceptrons using global optimization strategies", *Nonlinear Analysis, Theory, Methods and Applications*, Vol. 47, 3431-3436, (2001).
7. N.A. Katsanos, E. Iliopoulou, V.P. Plagianakos, and H. Mangou, "Interrelations between Absorption Energies and Local Isotherms, Local Monolayer Capabilities, and Energy Distribution Functions, as Determined for Heterogeneous Surfaces by Inverse Gas Chromatography", *Journal of Colloid and Interface Science*, Vol. 239, 10-19, (2001).
8. F. Roubani-Kalantzopoulou, T. Artemiadi, N.A. Katsanos, and V.P. Plagianakos, "Time separation of Absorption Sites on Heterogeneous Surfaces by Inverse Gas Chromatography", *Chromatographia*, Vol. 53, 315-320, (2001).
9. N.A. Katsanos, F. Roubani-Kalantzopoulou, E. Iliopoulou, I. Bassiotis, V. Sio-kos, M.N. Vrahatis, and V.P. Plagianakos, "Lateral Molecular Interactions on Heterogeneous Surfaces Experimentally Measured", *Colloid and Surfaces A*, Vol. 210, 173-180, (2002).
10. G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "On Globally Convergent Algorithms with Local Learning Rates", *IEEE Transactions on Neural Networks*, Vol. 13, 774-779, (2002).
11. V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Deterministic Nonmonotone Strategies for Effective Training of Multi-Layer Perceptrons", *IEEE Transactions on Neural Networks*, in press, (2002).
12. V.P. Plagianakos and M.N. Vrahatis, "Parallel Evolutionary Training Algorithms for 'Hardware-Friendly' Neural Networks", *Natural Computing*, Vol. 1, 307-322, (2002).

13. M.N. Vrahatis, G.D. Magoulas, and V.P. Plagianakos, “From linear to nonlinear iterative methods”, *Applied Numerical Mathematics*, in press, (2002).

Β. Εργασίες σε Επιστημονικά Περιοδικά υπό Κρίση.

1. V.P. Plagianakos, G.D. Magoulas, N.K. Nouis, and M.N. Vrahatis, “Distributed Computing Methodology for Training Neural Networks with Large Training Sets”, (2001).
2. V.P. Plagianakos, G.D. Magoulas, N.K. Nouis, and M.N. Vrahatis, “Evolutionary Training of Multilayer Perceptrons with Threshold Activations”, (2001).

Γ. Άρθρα και Κεφάλαια σε Βιβλία με Σύστημα Κριτών.

1. V.P. Plagianakos, D.G. Sotiropoulos, and M.N. Vrahatis, “An Improved Backpropagation Method with Adaptive Learning Rate”, In: *Recent Advances in circuits and systems*, N.E. Mastorakis (ed.), World Scientific, (1998).
2. V.P. Plagianakos, D.G. Sotiropoulos, and M.N. Vrahatis, “Integer Weight Training by Differential Evolution Algorithms”, In: *Recent Advances in circuits and systems*, N.E. Mastorakis (ed.), World Scientific, (1998).
3. V.P. Plagianakos, G.D. Magoulas, G.S. Androulakis, and M.N. Vrahatis, “Global Search Methods for Neural Network Training”, In: *Advances in Intelligent Systems and Computer Science*, N.E. Mastorakis (ed.), World Scientific and Engineering Society Press, 47–52, (1999).
4. G.D. Magoulas, V.P. Plagianakos, G.S. Androulakis, and M.N. Vrahatis, “A Framework for the Development of Globally Convergent Adaptive Learning Rate Algorithms”, In: *Advances in Intelligent Systems and Computer Science*, N.E. Mastorakis (ed.), World Scientific and Engineering Society Press, 207–212, (1999).
5. V.P. Plagianakos and M.N. Vrahatis, “A Derivative Free Minimization Method For Noisy Functions”, In: *Advances in Combinatorial and Global Optimization*, A. Migdalas, P. Pardalos, and R. Burkard (eds.), World Scientific, River Edge, 283–296, (2001).
6. V.P. Plagianakos, G.D. Magoulas and M.N. Vrahatis, “Supervised trained using global search methods”, In: *Advances in Convex Analysis and Global Optimization*, Honoring the memory of C. Caratheodory (1873–1950), N. Hadjisavvas and P.M. Pardalos, (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, Chapter 26, 421–432, (2001).
7. V.P. Plagianakos, G.D. Magoulas and M.N. Vrahatis, “Learning rate adaptation in stochastic gradient descent”, In: *Advances in Convex Analysis and Global Optimization*, Honoring the memory of C. Caratheodory (1873–1950), N. Hadjisavvas and P.M. Pardalos, (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, Chapter 27, 433–444, (2001).
8. K.E. Parsopoulos, V.P. Plagianakos, G.D. Magoulas and M.N. Vrahatis, “Improving the particle swarm optimizer by function stretching”, In: *Advances in Convex Analysis and Global Optimization*, Honoring the memory of C. Caratheodory (1873–1950), N. Hadjisavvas and P.M. Pardalos, (eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, Chapter 28, 445–457, (2001).
9. V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, “Improved learning of neural nets through global search”, In: *Global Optimization - Selected Case Studies*, J.D. Pintér (ed.), Kluwer Academic Publishers, to appear.

Δ. Δημοσιεύσεις σε Διεθνή Επιστημονικά Συνέδρια με Σύστημα Κριτών.

1. V.P. Plagianakos and M.N. Vrahatis, "Training Neural Networks with 3-bit Integer Weights", In: *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'99)*, W. Banzhaf, J. Daida, A.E. Eiben, M.H. Garzon, V. Honavar, M. Jakiela, and R.E. Smith (eds.), Morgan Kaufmann, 910-915, Orlando, U.S.A., (1999).
2. V.P. Plagianakos and M.N. Vrahatis, "Neural Network Training with Constrained Integer Weights", In: *Proceedings of the Congress of Evolutionary Computation (CEC'99)*, P.J. Angeline, Z. Michalewicz, M. Schoenauer, X. Yao, and A. Zalzal (eds.), IEEE Press, 2007-2013, Washington D.C., U.S.A., (1999).
3. V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Nonmonotone Learning Rules For Backpropagation Networks", In: *Proceedings of the 6th IEEE International Conference on Electronics, Circuits and Systems (ICECS '99)*, 291-294, Pafos, Cyprus, (1999).
4. G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "Effective Neural Network Training with a Different Learning Rate for each Weight", In: *Proceedings of the 6th IEEE International Conference on Electronics, Circuits and Systems (ICECS '99)*, 591-594, Pafos, Cyprus, (1999).
5. G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "Sign-methods for Training with Imprecise Error Function and Gradient Values", In: *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'99)*, Washington D.C., U.S.A., (1999).
6. M.N. Vrahatis, G.D. Magoulas, and V.P. Plagianakos, "Convergence Analysis of the Quickprop Method", In: *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'99)*, Washington D.C., U.S.A., (1999).
7. V.P. Plagianakos, M.N. Vrahatis, and G.D. Magoulas, "Nonmonotone Methods for Backpropagation Training with Adaptive Learning Rate", In: *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'99)*, Washington D.C., U.S.A., (1999).
8. V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Optimization Strategies and Backpropagation Neural Networks", In: *Proceedings of the Seventh Hellenic Conference on Informatics*, Ioannina, Greece, (1999).
9. V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Global Learning Rate Adaptation in On-line Neural Network Training", In: *Proceedings of the Second International ICSC Symposium on Neural Computation (NC'2000)*, Berlin, Germany, (2000).
10. G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "Development and Convergence Analysis of Training Algorithms with Local Learning Rate Adaptation", In: *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'2000)*, Como, Italy, (2000).
11. V.P. Plagianakos and M.N. Vrahatis, "Training Neural Networks with Threshold Activation Functions and Constrained Integer Weights", In: *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'2000)*, Como, Italy, (2000).
12. K.E. Parsopoulos, V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Stretching technique for obtaining global minimizers through Particle Swarm Optimization", In: *Proceedings of the Particle Swarm Optimization Workshop*, 22-29, Indianapolis, U.S.A., (2001).

13. G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "Hybrid Methods Using Evolutionary Algorithms for On-line Training", In: *Proceedings of the INNS-IEEE International Joint Conference on Neural Networks (IJCNN'2001)*, Washington D.C., U.S.A., (2001).
14. V.P. Plagianakos, G.D. Magoulas, N.K. Nouis, and M.N. Vrahatis, "PVM-based Training of Large Neural Architectures", In: *Proceedings of the INNS-IEEE International Joint Conference on Neural Networks (IJCNN'2001)*, Washington D.C., U.S.A., (2001).
15. V.P. Plagianakos, G.D. Magoulas, N.K. Nouis, and M.N. Vrahatis, "Training Multilayer Networks with Discrete Activation Functions", In: *Proceedings of the INNS-IEEE International Joint Conference on Neural Networks (IJCNN'2001)*, Washington D.C., U.S.A., (2001).
16. V.P. Plagianakos and E. Tzanaki, "Chaotic analysis of seismic time series and short term forecasting using neural networks", In: *Proceedings of the INNS-IEEE International Joint Conference on Neural Networks (IJCNN'2001)*, Washington D.C., U.S.A., (2001).
17. V.P. Plagianakos, G.D. Magoulas, and M.N. Vrahatis, "Tumor detection in colonoscopic images using hybrid methods for on-line neural network training", In: *Proceedings of the Neural Networks and Expert Systems in Medicine and HealthCare*, Milos Island, Greece, (2001).
18. D.G. Sotiropoulos, V.P. Plagianakos, and M.N. Vrahatis, "An Evolutionary Algorithm for Minimizing Multimodal Functions", In: *Proceedings of the Hellenic European Research on Computer Mathematics and its Applications Conference (HERCMA'2001)*, Athens, Greece, (2001).
19. G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, Improved Neural Network-based Interpretation of Colonoscopy Images Through On-line Learning and Evolution, In: *Proceedings of European Symposium on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems, EUNITE 2001*, 402-407, Tenerife, Spain, (2001).

E. Άλλες Δημοσιεύσεις.

1. Μ.Ν. Βραχάτης, Γ. Μαγουλάς, και Β.Π. Πλαγιανάκος, "Η Εκπαίδευση Τεχνητών Νευρωνικών Δικτύων με Επίβλεψη", Στο: *Τάξη και Χάος στα Μη Γραμμικά Συστήματα*, Αν. Μπούνη και Σ. Πνευματικός, Τόμος 6, 243-262, Εκδόσεις Γ. Πνευματικού, Αθήνα, (2000).
2. Μ.Ν. Βραχάτης, Β.Π. Πλαγιανάκος, και Γ. Μαγουλάς, "Εισαγωγή στα Τεχνητά Νευρωνικά Δίκτυα", Στο: *Τάξη και Χάος στα Μη Γραμμικά Συστήματα*, Αν. Μπούνη, Δ. Έλληνας, και Ι. Γρυσιολάκης, Τόμος 7, 225-247, Εκδόσεις Γ. Πνευματικού, Αθήνα, (2002).

Z. Αναφορές από άλλους ερευνητές.

1. Για την εργασία Α.2
A. Elbert, "Some recent results on the zeros of Bessel functions and orthogonal polynomials", *Journal of Computational and Applied Mathematics*, Vol. 133, 65-83, (2001).
2. Για την εργασία Δ.12
X. Hu and R.C. Eberhart, "Tracking dynamic systems with PSO: Where's the cheese?", In: *Proceedings of the Particle Swarm Optimization Workshop*, Indianapolis, Indiana, U.S.A., 80-83, (2001).

3. Για την εργασία Δ.12
“Fundamentals of Particle Swarm Optimization techniques”, In: *IEEE Power engineering society tutorial on modern heuristic optimization techniques with application to power systems*, K.Y. Lee and M.A. El-Sharkawi (eds.), (2002).
4. Για την εργασία Δ.12
“Particle Swarm Optimization: Developments, applications and resources”, In: *Proceedings of the Congress on Evolutionary Computation (CEC'2001)*, Seoul, Korea, 81-86, (2001).
5. Για την εργασία Δ.11
A.C.L. Corbalan, A.C.M. Pisano, A.C.G. Osella Masa, L.L. Lanzarini, “Crituras virtuales especificadas a través de redes neuronales evolutivas”, (στον δικτυακό τόπο: <http://170.210.92.2:300/CACIC2001/trabajos/pdf/SI-00075.pdf>).
6. Για την εργασία Δ.2
P. Grim, P.S. Denis, and T. Kokalis, “Learning to Communicate: The emergence of signaling in spatialized arrays of neural nets”, *Adaptive Behavior*, in press.
7. Για την εργασία Α.5
F. Bergh, “An analysis of Particle Swarm Optimizers”, *Ph.D. Thesis*, (2001).
8. Για την εργασία Γ.8
F. Bergh, “An analysis of Particle Swarm Optimizers”, *Ph.D. Thesis*, (2001).
9. Για την εργασία Δ.12
F. Bergh, “An analysis of Particle Swarm Optimizers”, *Ph.D. Thesis*, (2001).
10. Για την εργασία Δ.2
D. Braendler, “Implementing Neural Hardware with On Chip Training on Field Programmable Gate Arrays”, *Ph.D. Thesis*, (2002).
11. Για την εργασία Δ.1
Z. Tóth, “The Generic Evolutionary Algorithms Programming Library”, *Ph.D. Thesis*, (2000).
12. Για την εργασία Δ.12
R. Brits, A.P. Engelbrecht and F. Bergh, “A niching Particle Swarm Optimizer”, In: *Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution and Learning*, (2002).

Η. Ηλεκτρονικές αναφορές από άλλους ερευνητές.

1. Για την εργασία Δ.12
Y. Shi, “Particle Swarm Optimization Bibliography”,
<http://www.engr.iupui.edu/~shi/PSO/bibliography.html>
2. Για τις εργασίες Γ.2, Δ.1, Δ.2, Δ.11, και Δ.13
J. Lampinen, “A Bibliography of Differential Evolution Algorithm”,
<http://www.lut.fi/~jlampine/debiblio.htm>

Θ. Βραβεία-Διακρίσεις.

1. IEEE Neural Network Council Student Travel Grant, (2000).
2. IEEE Neural Network Council Student Travel Grant, (2001).
3. Για την εργασία Δ.19
του έχει απονεμηθεί από το “European Network of Excellence on Intelligent Technologies for Smart Adaptive Systems (EUNITE)” το τρίτο βραβείο “Best Paper Awards in Human, Medical and Healthcare Track”, (2001).

Α

Αλγόριθμος εκπαίδευσης, 3, 5, 10
Ανάδραση, 7
Ανεπιθύμητο τοπικό ελάχιστο, 71
Αρχικοποίηση των βαρών, 11
Αρχιτεκτονική, 6
Άξονας, 4

Β

Βάρος, 3, 5
Βάρος αδράνειας, 76, 82
Βιολογικό νευρωνικό δίκτυο, 4

Γ

Γενετικοί αλγόριθμοι, 55, 74, 93
Γενιά, 54, 74
Γενίκευση, 10, 49
Γήρανση του πληθυσμού, 63
Γονίδια, 74

Δ

Δενδρίτες, 4
Διαφοροεξελικτικοί Αλγόριθμοι, 53

Ε

Εκπαίδευση
 ανά ομάδα προτύπων εισόδου, 29, 41, 88, 97, 103, 115
 ανά πρότυπο εισόδου, 19, 87, 88
 με ακέραια βάρη, 53
 με ενεργοποιήσεις κατώφλια, 60
 με ενίσχυση, 14
 με επίβλεψη, 14, 88
 με παράλληλους ΔΕΑ, 62
 με περιορισμένα ακέραια βάρη, 58
 παράλληλη, 14
 χωρίς επίβλεψη, 14
Εκπαίδευση ΤΝΔ, 5, 10
Ενεργοποίηση, 5
Εξελικτικοί αλγόριθμοι, 53, 93
Επιθυμητό τοπικό ελάχιστο, 71
Επίπεδες περιοχές, 73, 88, 99
Επίπεδο, 6
Ευθύγραμμη ανίχνευση, 20, 73, 82, 104, 115
Ευρεία σύγκλιση, 30, 32, 34, 44

Θ

Θεώρημα
 Hecht-Nielsen, 26
 Kolmogorov, 26
 Sprecher, 26
 Wolfe-Zoutendijk, 32

Κ

Καθολικός προσεγγιστής, 25
Κανόνας του μεγίστου, 49, 114
Καταστροφική παρέμβαση, 89, 95, 96
Κίνηση Metropolis, 73
Κολονοσκόπηση, 94
Κολοσκόπηση, 94
Κρυφό στρώμα νευρώνων, 7

Μ

Μέγιστος παράγοντας αύξησης, 44
Μέθοδοι ευρείας σύγκλισης, 19, 29
Μέθοδοι ολικής ανίχνευσης, 71
Μέθοδοι ολικής βελτιστοποίησης, 71
Μέθοδος
 Broyden, 43
 Fletcher-Reeves, 47, 102
 Polak-Ribiere, 47, 102
 Powell, 23
 Quickprop, 35, 41, 43
 Silva-Almeida, 35
 βελτιστοποίησης με σμήνος σωματιδίων, 75
 για επανεκπαίδευση, 92
 διχοτόμησης, 24
 μη γραμμική Jacobi, 21
 μη γραμμική SOR, 22
 μη μονότονη, 97
 οπισθοδρομικής διάδοσης του σφάλματος, 18, 47, 66, 81, 87, 90, 102, 113
 οπισθοδρομικής διάδοσης του σφάλματος Barzilai-Borwein, 100, 112
 εκπαίδευσης ανά πρότυπο εισόδου, 90
 με εξασθένηση των βαρών, 66, 113
 με μεταβλητό βήμα, 100, 111
 με ορμή, 47, 81, 90, 100, 102, 110
 με προσαρμοστικό ρυθμό εκπαίδευσης και ορμή, 47, 90, 102
 πιο απότομης καθόδου, 18
 πιο απότομης καθόδου (στοχαστική), 87, 89, 93
 προσομοιωμένης ανόπτησης, 73
 τροποποιημένη Quickprop, 41
 των διαδοχικών συσχετίσεων, 66, 113
 χορδής, 42
Μετανάστευση, 63
Μη μονότονος ορίζοντας εκπαίδευσης, 97, 99
Μη στατικά προβλήματα, 10, 60, 87

Ν

Νευρώνας, 3, 4

$\mathcal{O}$

Ολική ανίχνευση, 72

 Π

Παράλληλοι Διαφοροεξελικτικοί Αλγόριθμοι, 62
 Πληθυσμός, 54, 58, 62, 74
 Πόλωση, 5
 Πρόβλημα
 4-2-4 Κωδικοποιητή/Αποκωδικοποιητή, 57, 65, 124
 αναγνώρισης ανωμαλιών σε κολονοσκοπήσεις, 39, 93
 αναγνώρισης των αριθμών, 36, 49, 92, 107, 110, 114, 125
 αναγνώρισης των κεφαλαίων γραμμάτων, 92, 106, 107, 112, 125
 αναγνώρισης χειρόγραφων αριθμών, 113
 αποκλειστικού-ΕΠΕ, 47, 56, 59, 61, 64, 82, 91, 101, 111, 123
 γενίκευσης MONK, 66, 113, 125
 ισοτιμίας 3-bit, 56, 59, 61, 64, 82, 105, 123
 προσέγγισης μιας συνεχούς συνάρτησης, 38, 105, 111, 112, 125
 ταξινόμησης υφής, 48, 93, 107, 111, 114, 126
 Πρότυπα εισόδου, 9
 Πρότυπα εκπαίδευσης, 9
 Πρότυπα ελέγχου, 49

 $\mathcal{P}$

Ρυθμός εκπαίδευσης, 18
 Ρυθμός εκπαίδευσης ανά κατεύθυνση, 29

 Σ

Σμήνος, 75
 Σταθερά Lipschitz, 19, 32, 99
 Σταθερά ανασυνδυασμού, 55
 Σταθερά μετάλλαξης, 55
 Σταθερά μετανάστευσης, 63
 Σταθερά ορμής, 100
 Στιγμιαίο σφάλμα, 88
 Στρατηγική οπισθοδρόμησης, 35
 Στρώμα, 3, 6
 Στρώμα εισόδου, 6
 Στρώμα εξόδου, 6
 Συνάρτηση ενεργοποίησης, 5
 Συνάρτηση ενεργοποίησης
 Διπολική με κατώφλι θ , 5
 Διαδική με κατώφλι θ , 5
 Λογιστική, 5
 Ταυτοτική, 5
 Υπερβολική, 5
 Συνάρτηση μεταφοράς, 5
 Σύναψη, 4
 Συνθήκες του Wolfe, 20, 32, 44
 Συνθήκη του Zoutendijk, 32
 Συντελεστής βάρους, 3, 5

Σφάλμα ταξινόμησης, 91, 93, 95, 107
 Σώμα, 4
 Σωματίδιο, 75

 $\mathcal{T}$

Ταχύτητα σωματιδίου, 76
 Τελεστής
 ανασυνδυασμού, 54, 55, 74
 επιλογής, 54, 74
 μετάλλαξης, 54, 58, 74
 Τεχνητό νευρωνικό δίκτυο, 3, 4
 Τεχνητό Νευρωνικό Δίκτυο
 δυναμικό, 7
 με ακέραια βάρη, 53
 πλήρως διασυνδεδεμένο, 7
 πολυστρωματικό, 7
 πρόσθιας τροφοδότησης, 7
 Τεχνική της παρεκκλίνουσας τροχιάς, 77
 Τεχνική του «εφελκυσμού» της συνάρτησης σφάλματος, 79
 Τοπικά ελάχιστα, 11, 71, 77, 87, 88
 Τοπική ανίχνευση, 71
 Τοπική εκτίμηση της σταθεράς Lipschitz, 99, 100, 115
 Τοπικός ρυθμός εκπαίδευσης, 29
 Τοπολογία, 6

 Υ

Υλοποίηση σε υλικό, 10, 53, 60
 Υποπληθυσμός, 63

 $\mathcal{X}$

Χρωμοσώματα, 74