

Η Εκπαίδευση Τεχνητών Νευρωνικών Δικτύων με Επίβλεψη ως Πρόβλημα Ελαχιστοποίησης

M.N. Βραχάτης¹, Γ.Δ. Μαγουλάς², Β.Π. Πλαγιανάκος¹

¹Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών, 261 10, Πάτρα.

²Department of Information Systems and Computing, Brunel University, UB8 3PH, U.K.
vrahatis@math.upatras.gr, George.Magoulas@brunel.ac.uk, vpp@math.upatras.gr

1. Εισαγωγή

Ο όρος Τεχνητό Νευρωνικό Δίκτυο (ΤΝΔ), αναφέρεται σε μια αρχιτεκτονική που εκτελεί αριθμητικούς υπολογισμούς χρησιμοποιώντας δομή μαζικού παραλληλισμού (massively parallel structure) ή παράλληλης κατανεμημένης επεξεργασίας (parallel distributed processing). Τα ΤΝΔ απαρτίζονται από τεχνητούς νευρώνες που αλληλεπιδρούν μέσω συνδέσμων που ονομάζονται *συντελεστές βάρους* ή απλά *βάρη*. Θετικά ή αρνητικά βάρη αντιστοιχούν σε συνάψεις που μεταδίδουν ή αναστέλλουν ερεθίσματα από άλλους νευρώνες.

Η μάθηση στα ΤΝΔ αποτελεί ένα μέσο δυναμικής αναπαράστασης κωδικοποιημένης πληροφορίας στους νευρώνες ενός ΤΝΔ. Η προσέγγιση ότι η εκπαίδευση ΤΝΔ με επίβλεψη αντιστοιχεί στην ελαχιστοποίηση της συνάρτησης σφάλματος οδηγεί στην ανάπτυξη αλγορίθμων μάθησης που βασίζονται στην αριθμητική ελαχιστοποίηση χωρίς περιορισμούς, και ιδιαίτερα σε τεχνικές που χρησιμοποιούν πληροφορία σχετική με τη συνάρτηση σφάλματος, όπως η μέθοδος συζυγών κλίσεων και η μέθοδος του Newton (Luenberger 1969, Ortega & Rheinboldt 1970, Polak 1997). Οι δύο αυτές μέθοδοι χρησιμοποιούν, μολονότι με διαφορετικό τρόπο, το διάνυσμα των μερικών παραγώγων πρώτης τάξης και την Εσσιανή (δεύτερης τάξης μερικές παράγωγοι) της συνάρτησης σφάλματος.

Στην εργασία του Battiti (1992) παρουσιάζεται μια επισκόπηση τεχνικών ελαχιστοποίησης με χρήση παραγώγων πρώτης και δεύτερης τάξης, η οποίες εφαρμόζονται στην εκπαίδευση των ΤΝΔ με επίβλεψη. Περιγραφή σε ψευδοκώδικα των μεθόδων που καταγράφονται παρακάτω υπάρχει στις εργασίες των Beigi *et al.* (1993), Møller (1993) και van der Smagt (1994), καθώς και στα βιβλία των Kung (1993) και Haykin (1994).

- *Η μέθοδος μέγιστης κλίσης (steepest descent) και οι τροποποιήσεις της.* Η μέθοδος αυτή χαρακτηρίζεται από πολύ καλή απόδοση όταν οι αρχικές τιμές του διανύσματος βαρών είναι μακριά από το ελάχιστο, κάτι που ισχύει στις περισσότερες περιπτώσεις εκπαίδευσης ΤΝΔ. Ωστόσο, η σύγκλισή της στην περιοχή του ελαχίστου χαρακτηρίζεται από εξαιρετική βραδύτητα. Σημαντικοί περιορισμοί για τη χρήση της μεθόδου στα ΤΝΔ είναι η αδυναμία της για εγγύηση σύγκλισης στο ολικό ελάχιστο καθώς και η χρήση σταθερού μήκους βήματος που πολλές φορές εμποδίζει τη σύγκλιση ακόμα και σε ένα τοπικό ελάχιστο. Επιπλέον, η μέθοδος δεν εγγυάται τη μείωση της συνάρτησης σφάλματος σε κάθε επανάληψη του αλγορίθμου μάθησης. Για πρόσφατες τροποποιήσεις, βελτιώσεις και εφαρμογές αυτής της μεθόδου στα νευρωνικά δίκτυα, βλέπε Magoulas *et al.* (1997b), Magoulas *et al.* (1999), Magoulas *et al.* (2000a), Magoulas *et al.* (2000b), Magoulas &

Vrahatis (2000), Plagianakos *et al.* (1999a), Plagianakos *et al.* (1999b), Vrahatis *et al.* (2000a) και Vrahatis *et al.* (2000b).

- *Η μέθοδος συζυγών κλίσεων (conjugate gradient) και οι επεκτάσεις της.* Η εκπαίδευση των ΤΝΔ σε πολλές εφαρμογές απαιτεί την προσαρμογή αρκετών εκατοντάδων ή ακόμα χιλιάδων βαρών. Οι μέθοδοι συζυγών κλίσεων μπορούν να αντιμετωπίσουν προβλήματα μεγάλης κλίμακας και να τεθούν σε εφαρμογή σε υπολογιστές πολλαπλών επεξεργαστών. Αρκετοί αλγόριθμοι μάθησης βασίζόμενοι σε αυτές τις μεθόδους έχουν παρουσιαστεί (Battiti 1989, Kramer & Sangiovanni-Vincentelli 1989, Møller 1990) και τα αποτελέσματα δείχνουν αυξημένη ταχύτητα μάθησης σε σχέση με τις μεθόδους μέγιστης κλίσης. Ωστόσο, χρησιμοποιούν ευθύγραμμη ανίχνευση (line search) για τον καθορισμό του κατάλληλου μήκους βήματος αυξάνοντας την υπολογιστική πολυπλοκότητα της διεργασίας μάθησης με αρκετούς υπολογισμούς της συνάρτησης του ολικού τετραγωνικού σφάλματος μάθησης ή των παραγώγων της ενώ η απόδοσή τους εξαρτάται από την ακρίβεια της ευθύγραμμης ανίχνευσης (Johansson *et al.* 1990). Χαρακτηριστικό των μεθόδων είναι ότι δεν ακολουθούν πάντα κατευθύνσεις μείωσης (descent directions) του σφάλματος με αποτέλεσμα να εμφανίζεται αριθμητική αστάθεια. Επιπλέον αποτυγχάνουν όταν τα αρχικά βάρη είναι μακριά από το επιθυμητό ελάχιστο (συνηθισμένο φαινόμενο στα ΤΝΔ) εξαιτίας του ότι η Εσσιανή δεν είναι θετικά ορισμένη σε διάφορες περιοχές του χώρου των βαρών.
- *Η μέθοδος του Newton.* Η μέθοδος Newton θεωρείται ως η βασικότερη μέθοδος εύρεσης τοπικού ελαχίστου με χρήση παραγώγων δεύτερης τάξης, όταν οι αρχικές τιμές του διανύσματος παραμέτρων είναι κοντά στο ελάχιστο. Η χρήση της στα ΤΝΔ περιορίζεται από το γεγονός ότι απαιτεί γνώση της Εσσιανής, αναλυτικός υπολογισμός της οποίας είναι πολύπλοκος και κοπιώδης για ΤΝΔ με περισσότερα από εκατό βάρη. Ακόμα και στην περίπτωση που η Εσσιανή είναι διαθέσιμη, η αντιστροφή της παραμένει μια χρονοβόρα διαδικασία που τις περισσότερες φορές επιβαρύνει τη διαδικασία εκπαίδευσης πιο πολύ από μερικές ακόμα επαναλήψεις μιας απλούστερης μεθόδου. Επιπλέον μειονεκτήματα αποτελούν η πολυπλοκότητά της ανά επανάληψη και η υπόθεση της *θετικά ορισμένης* Εσσιανής, διότι στα ΤΝΔ η Εσσιανή μπορεί να είναι *αρνητικά ορισμένη*, να έχει *μηδενική ορίζουσα* (singular) ή ακόμα να έχει *μεγάλο συντελεστή αστάθειας* (ill-conditioned).
- *Η μέθοδος μεταβλητής μετρικής Broyden-Fletcher-Goldfarb-Shanno (BFGS).* Η μέθοδος αυτή έχει καλές ιδιότητες σύγκλισης τόσο θεωρητικά όσο και πρακτικά. Υπολογίζει μια προσέγγιση της Εσσιανής μειώνοντας έτσι την πολυπλοκότητα σε σχέση με την μέθοδο Newton. Επιπλέον η Εσσιανή της ορίζεται θεωρητικά ως *συμμετρική* και θετικά ορισμένη κάτι που εγγυάται την αριθμητική ευστάθεια του αλγορίθμου. Ο Watrous (1987) εφάρμοσε για πρώτη φορά αυτή τη μέθοδο στην εκπαίδευση των ΤΝΔ. Πέτυχε να επιταχύνει τη διαδικασία εκπαίδευσης σε σχέση με τις μεθόδους μέγιστης κλίσης και συζυγών κλίσεων, ωστόσο η υπολογιστική πολυπλοκότητα της BFGS είναι σημαντικά υψηλότερη της μεθόδου μέγιστης κλίσης αφού εκτελεί ευθύγραμμη ανίχνευση όπως η μέθοδος συζυγών κλίσεων, ενώ οι υπολογισμοί που εκτελούνται (υπολογισμός της κατεύθυνσης, αντιστροφή Εσσιανής, υπολογισμός λύσης) είναι ανάλογοι του τετραγώνου των βαρών. Πέρα από τα υπολογιστικά προβλήματα, σημαντικό μειονέκτημα της μεθόδου στα ΤΝΔ είναι ότι ο απαιτούμενος χώρος μνήμης για την αποθήκευση της Εσσιανής εξαρτάται από το τετράγωνο του αριθμού των βαρών του δικτύου κάτι που καθιστά την εφαρμογή της προβληματική για ΤΝΔ με μερικές εκατοντάδες βάρη. Επίσης στην πράξη είναι δύσκολο να αντιστραφεί η Εσσιανή μια και ο χρόνος που απαιτείται είναι, στις περισσότερες περιπτώσεις, μεγαλύτερος από αυτόν που χρειάζεται για μερικές ακόμα επαναλήψεις της μεθόδου μέγιστης κλίσης. Παρόμοια

συμπεράσματα παρουσιάστηκαν και σε άλλες εργασίες (Battiti & Masulli 1990, van der Smagt 1994). Παρόλα αυτά η μέθοδος παραμένει ανταγωνιστική όταν ο αριθμός των προτύπων προς μάθηση είναι πολύ μεγάλος. Τότε, ο υπολογισμός της συνάρτησης σφάλματος είναι κοπιώδης και η ταχύτητα μείωσης του σφάλματος μάθησης παίζει το σπουδαιότερο ρόλο.

- Η μέθοδος BFGS με περιορισμένη μνήμη για προβλήματα μεγάλης κλίμακας. Αν και το πρόβλημα της διαθέσιμης μνήμης αποθήκευσης δεν είναι πλέον τόσο σημαντικό όσο πριν μια δεκαετία, η μέθοδος αυτή, που ομοιάζει της μεθόδου BFGS, αποφεύγει την αποθήκευση των πινάκων και έτσι είναι πιο εύχρηστη σε προβλήματα μεγάλης κλίμακας (Liu & Nocedal 1989, Kung 1993). Ο Battiti (1990, 1992) στηριζόμενος σε αυτή τη μέθοδο και αναγνωρίζοντας ότι το υπολογιστικό κόστος παραμένει εξαιρετικά υψηλό για δίκτυα με περισσότερα από εκατό βάρη πρότεινε μια τροποποίησή της με πολυπλοκότητα ανάλογη του αριθμού των βαρών. Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί μόνο για μάθηση ανά ομάδα προτύπων, κάτι που αποτελεί περιοριστικό παράγοντα για τη χρήση της, ενώ η ταχύτητα εκπαίδευσης βελτιώνεται σε σχέση με τη μέθοδο μέγιστης κλίσης μόνο στην περίπτωση που απαιτείται υψηλή ακρίβεια στην εύρεση της λύσης. Στην πράξη ωστόσο, η ακριβής εύρεση του ελαχίστου αντιστοιχεί σε απομνημόνευση του συνόλου των προτύπων και σε πολλές εφαρμογές περιορίζει την ικανότητα γενίκευσης του ΤΝΔ.

2. Μάθηση με επαναληπτική προσαρμογή των βαρών

Η εκπαίδευση με επίβλεψη μπορεί να ταυτιστεί με το πρόβλημα της ελαχιστοποίησης της συνάρτησης σφάλματος του ΤΝΔ, η οποία θα συμβολίζεται παρακάτω με E , θεωρείται συνεχής και παραγωγίσιμη και εξαρτάται από q βάρη. Το ζητούμενο είναι η εύρεση ενός \mathbf{w}^* τέτοιο ώστε:

$$\mathbf{w}^* = \min_{\mathbf{w}} E(\mathbf{w}).$$

Στη συνέχεια θα ασχοληθούμε με αλγόριθμους μάθησης που προσαρμόζουν τα βάρη με την επαναληπτική σχέση:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \alpha_k \mathbf{d}_k, \quad (1)$$

όπου \mathbf{d}_k είναι η κατεύθυνση ανίχνευσης (search direction) και α_k είναι το μήκος του βήματος, το οποίο λαμβάνεται μέσω μονοδιάστατης ανίχνευσης και καθορίζει το ρυθμό μάθησης.

Η κατεύθυνση ανίχνευσης στην περίπτωση των μεθόδων συζυγών κλίσεων έχει την παρακάτω μορφή:

$$\mathbf{d}_k = -\nabla E(\mathbf{w}_k) + \beta_k \mathbf{d}_{k-1}$$

όπου το βαθμωτό β_k επιλέγεται έτσι ώστε να προκύπτει η γραμμική μέθοδος συζυγών κλίσεων όταν η συνάρτηση είναι τετραγωνική και η ευθύγραμμη ανίχνευση ακριβής. Μια άλλη κατηγορία μεθόδων ακολουθεί την κατεύθυνση ανίχνευσης που ορίζεται ως εξής:

$$\mathbf{d}_k = -\mathbf{B}_k^{-1} \nabla E(\mathbf{w}_k), \quad (2)$$

όπου \mathbf{B}_k είναι ένας συμμετρικός πίνακας μη ιδιάζων (nonsingular), δηλαδή έχει μη μηδενική ορίζουσα. Ειδικές περιπτώσεις δίνονται από τις σχέσεις:

$$\mathbf{B}_k = \mathbf{I} \text{ (για τη μέθοδο μέγιστης κλίσης)}$$

$$\mathbf{B}_k = \nabla^2 E(\mathbf{w}_k) \text{ (για τη μέθοδο Newton)}$$

Οι μέθοδοι μεταβλητής μετρικής είναι επίσης της μορφής (2), αλλά στην περίπτωση αυτή ο πίνακας \mathbf{B}_k δεν είναι μόνο συνάρτηση του \mathbf{w}_k , αλλά εξαρτάται και από τα \mathbf{B}_{k-1} και \mathbf{w}_{k-1} .

Σημαντικό για όλους τους αλγόριθμους μάθησης είναι η κατεύθυνση ανίχνευσης \mathbf{d}_k να είναι κατεύθυνση μείωσης της συνάρτησης σφάλματος, δηλαδή:

$$\mathbf{d}_k^T \nabla E(\mathbf{w}_k) < 0$$

ώστε η τιμή της συνάρτησης σφάλματος να μειώνεται ακολουθώντας ένα μικρό βήμα κατά μήκος της \mathbf{d}_k . Η κατεύθυνση της μεθόδου μέγιστης κλίσης θεωρείται τοπικά ως η κατεύθυνση της ταχύτερης μείωσης, ενώ για μεθόδους που ομοιάζουν της μεθόδου Newton (2) η \mathbf{d}_k είναι κατεύθυνση μείωσης της συνάρτησης όταν ο πίνακας \mathbf{B}_k είναι θετικά ορισμένος. Στην περίπτωση των αλγόριθμων μάθησης που βασίζονται στη μέθοδο συζυγών κλίσεων απαιτείται προσεκτική επιλογή της στρατηγικής ευθύγραμμης ανίχνευσης που θα χρησιμοποιηθεί ώστε να επιτύχουμε κατεύθυνση μείωσης. Η απαίτηση αυτή σχετικά με την κατεύθυνση ανίχνευσης αποτελεί και τη βάση της ανάπτυξης των αλγόριθμων ευρείας σύγκλισης, δηλαδή με σύγκλιση σε ένα ελάχιστο από οποιαδήποτε αρχική συνθήκη (Dennis & Schnabel 1983, Byrd *et al.* 1987) και αποτελεί επιθυμητό χαρακτηριστικό κάθε αλγόριθμου μάθησης (Battitti 1992).

Τέλος είναι εμφανές από τα παραπάνω ότι η ελαχιστοποίηση της συνάρτησης σφάλματος $E(\mathbf{w})$ απαιτεί τον υπολογισμό των μερικών παραγώγων $\nabla E(\mathbf{w})$ της $E(\mathbf{w})$ ως προς τα βάρη του ΤΝΔ. Επομένως, είναι απαραίτητο η συνάρτηση σφάλματος να ικανοποιεί τις υποθέσεις για την ύπαρξη των παραγώγων πρώτης τάξης. Ασφαλώς αυτό περιορίζει τη μορφή της $E(\mathbf{w})$ και καθιστά απαραίτητο οι νευρώνες του ΤΝΔ να ακολουθούν μοντέλα που επιτρέπουν τον ορισμό της παραγώγου για κάθε νευρώνα.

Η δημοφιλέστερη τεχνική για τον αναλυτικό υπολογισμό των μερικών παραγώγων της $E(\mathbf{w})$ είναι ο *αλγόριθμος οπισθοδρομικής διάδοσης του σφάλματος (backpropagation)*. Η πρώτη προσέγγιση της τεχνικής αυτής οφείλεται στον Werbos (1974), ενώ η μορφή που έγινε ευρέως γνωστή και καθιερώθηκε στο χώρο των ΤΝΔ οφείλεται στους Rumelhart *et al.* (1986). Η αναλυτική περιγραφή του αλγόριθμου οπισθοδρομικής διάδοσης του σφάλματος υπάρχει σε όλα τα βιβλία σχετικά με ΤΝΔ και δε θα μας απασχολήσει εδώ. Ωστόσο, αξίζει να αναφερθούν οι προσπάθειες των Saarinen (1992) και Rojas (1993) για μια απλή και συνάμα γενική περιγραφή του αλγόριθμου, πέρα από τη συνηθισμένη που συναντιέται στη βιβλιογραφία και να τονιστεί η δυνατότητα παράλληλης υλοποίησής του με τεχνολογία VLSI η οποία αποτελεί πόλο έλξης για τις εφαρμογές ΤΝΔ.

Χρησιμοποιώντας τον αλγόριθμο οπισθοδρομικής διάδοσης του σφάλματος μπορούμε να υπολογίσουμε το σύνολο των q μερικών παραγώγων της συνάρτησης σφάλματος του δικτύου ως προς τα στοιχεία του διανύσματος \mathbf{w} για ένα πρότυπο p (βλέπε Rumelhart *et al.* 1986). Επαναλαμβάνοντας τους υπολογισμούς για όλα τα πρότυπα $p \in [1, P]$, όπου P είναι το πλήθος των προτύπων που απαρτίζουν το σύνολο των παραδειγμάτων προς μάθηση, καταλήγουμε σε ένα $q \times P$ πίνακα μερικών παραγώγων, που χρησιμοποιείται για την περίπτωση μάθησης ανά ομάδα προτύπων εισόδου.

Για την περίπτωση μάθησης ανά πρότυπο p εισόδου χρησιμοποιείται μια στιγμιαία προσέγγιση, 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 μάθησης (Saarinen *et al.* 1992).

3. Η Εσσιανή της συνάρτησης σφάλματος και οι προσεγγίσεις της

Η πληροφορία των δευτέρων παραγώγων της συνάρτησης σφάλματος ως προς τα βάρη που περιέχεται στην Εσσιανή $\nabla^2 E(\mathbf{w}_k)$ έχει μεγάλη θεωρητική και πρακτική αξία. Για παράδειγμα, επιτρέπει την εφαρμογή σύνθετων αλγόριθμων τύπου Newton για την εκπαίδευση ΤΝΔ, όπως αυτοί που πρότεινε ο Watrous (1987), εμφανίζεται σε αρκετές τεχνικές πρόβλεψης και ενίσχυσης της γενίκευσης των ΤΝΔ, όπως αυτές των MacKay (1991), Moody (1992), Le Cun *et al.* (1990), Hassibi & Stork (1993) και βοηθά στη μελέτη και τη σύγκριση της συμπεριφοράς των αλγόριθμων μάθησης, χρησιμοποιώντας τα ιδιοδιανύσματα των ακραίων ιδιοτιμών της Εσσιανής, όπως πρότειναν οι Androurlakis *et al.* (1997).

Μια τεχνική αναλυτικού υπολογισμού της Εσσιανής δεν εφαρμόζει απλά τον κανόνα παραγωγίσης των πεπλεγμένων συναρτήσεων. Ασφαλώς και οι δύο καταλήγουν στο ίδιο αποτέλεσμα αλλά όπως και στην περίπτωση του αλγόριθμου οπισθοδρομικής διάδοσης του σφάλματος για τον υπολογισμό του πίνακα των μερικών παραγώγων, η σημασία της τεχνικής του αναλυτικού υπολογισμού της Εσσιανής είναι ότι επιτρέπει την οργάνωση των δεδομένων με τρόπο ώστε να αποφεύγονται περιττοί υπολογισμοί, καθώς ο υπολογισμός της Εσσιανής περιλαμβάνει επαναλαμβανόμενο υπολογισμό των ίδιων όρων. Διάφορες τεχνικές έχουν προταθεί για να αντιμετωπιστεί η περιπλοκότητα του αναλυτικού υπολογισμού δευτέρων παραγώγων για ένα ΤΝΔ με q βάρη, που είναι ανάλογη του τετραγώνου του πλήθους των βαρών $O(q^2)$ (Werbos *et al.* 1992, Bishop 1992, Buntine & Weigend 1994).

Η τεχνική των Buntine και Weigend είναι η γενικότερη καθώς εφαρμόζεται σε ΤΝΔ οποιασδήποτε αρχιτεκτονικής και συνάρτησης ενεργοποίησης. Ωστόσο, οι υπολογισμοί επιβαρύνονται και από την αντιστροφή της Εσσιανής που είναι απαραίτητη σε πολλούς αλγόριθμους μάθησης και πρέπει να επαναλαμβάνεται σε κάθε επανάληψη του αλγόριθμου. Απαραίτητη προϋπόθεση είναι η Εσσιανή να έχει μη μηδενική ορίζουσα ώστε να μπορεί να γίνει η αντιστροφή της. Δυστυχώς παρατηρείται ότι κατά τη διαδικασία εκπαίδευσης η Εσσιανή είναι ως επί το πλείστον μη αντιστρέψιμη. Επιπλέον, πειραματικά αποτελέσματα δείχνουν πως έχει μεγάλο συντελεστή αστάθειας, εισάγοντας πρόσθετες υπολογιστικές δυσχέρειες (Saarinen *et al.* 1992). Τα υπολογιστικά προβλήματα συνδυάζονται και με προβλήματα αποθήκευσης για ΤΝΔ με αρκετές εκατοντάδες βάρη καθώς ο χώρος αποθήκευσης που απαιτείται για ένα δίκτυο με q βάρη είναι $q \times q$. Για αυτούς τους λόγους έχουν προταθεί διάφορες τεχνικές για την προσέγγιση πληροφοριών που σχετίζονται με την συμπεριφορά της επιφάνειας σφάλματος σε κάθε επανάληψη και περιέχονται στην Εσσιανή. Οι τεχνικές προσέγγισης που ενσωματώνονται στους αλγόριθμους μάθησης λαμβάνουν ποικίλες μορφές:

- *Αλγόριθμοι μάθησης βασισμένοι σε προσεγγίσεις της Εσσιανής.* Για μεγάλα ΤΝΔ ο αναλυτικός υπολογισμός της Εσσιανής έχει απαγορευτικό υπολογιστικό κόστος. Οι Becker & Le Cun (1988) πρότειναν την απλή προσέγγιση της Εσσιανής από τη διαγώνιο της και τη χρησιμοποίησαν σε έναν αλγόριθμο μάθησης τύπου Newton. Ο Fahlman (1988) χρησιμοποίησε στον αλγόριθμο Quickprop πεπερασμένες διαφορές για να προσεγγίσει τη διαγώνιο (βλέπε Vrahatis *et al.* 2000b, για μια βελτίωση και απόδειξη σύγκλισης), ενώ οι El-

Jaroudi & Makhoul (1990) προσέγγισαν τμήματα της Εσσιανής θεωρώντας ότι τα βάρη που συνδέουν διαφορετικούς νευρώνες δεν επιδρούν στην Εσσιανή. Μια νέα προσέγγιση σε αλγόριθμους μάθησης που χρησιμοποιούν τη δεύτερη παράγωγο προτείνει την ελάττωση της διάστασης του προβλήματος μετασχηματίζοντας την Εσσιανή σε έναν πίνακα ελαττωμένης διάστασης. Σε αυτήν την *ελαττωμένη Εσσιανή* οι δεύτερες παράγωγοι προσεγγίζονται με πεπερασμένες διαφορές και ο απαιτούμενος χώρος για την αποθήκευσή της είναι της τάξης $(q-1) \times (q-1)$. Ο αλγόριθμος χρησιμοποιεί μόνο το πρόσημο των παραγώγων, δεν απαιτεί αντιστροφές πινάκων και εμφανίζει βελτιωμένη συμπεριφορά σε περιπτώσεις που η Εσσιανή έχει μηδενική ορίζουσα ή μεγάλο συντελεστή αστάθειας. Σε κάθε επανάληψη προσαρμόζονται τα $(q-1)$ βάρη ενώ το q βάρος, που απομένει, υπολογίζεται από τις τιμές των άλλων (Magoulas *et al.* 1997a).

- *Αλγόριθμοι μάθησης χωρίς απευθείας υπολογισμό της Εσσιανής.* Αυτή η περίπτωση αφορά συνήθως αλγόριθμους μάθησης που βασίζονται σε συζυγείς κλίσεις και χρησιμοποιούν δεύτερες παραγώγους κατά την ευθύγραμμη ανίχνευση. Σε αυτές τις περιπτώσεις δεν είναι απαραίτητος ο υπολογισμός ολόκληρης της Εσσιανής αλλά ο υπολογισμός του γινομένου της Εσσιανής με ένα διάνυσμα. Πεπερασμένες διαφορές και ακριβής υπολογισμός του γινομένου έχουν προταθεί για αυτές τις περιπτώσεις (Møller 1993, Pearlmutter 1994). Η τεχνική που πρότεινε ο Pearlmutter βρίσκει εφαρμογή και σε άλλους αλγόριθμους μάθησης όπως αυτούς που πρότειναν οι Ackley *et al.* (1985), ο Almeida (1987) και ο Pineda (1987). Ενώ η προσέγγιση με πεπερασμένες διαφορές χρησιμοποιήθηκε και σε αλγόριθμους μάθησης που βασίζονται στη μέθοδο μέγιστης κλίσης για την εκτίμηση των ιδιοδιανυσμάτων και ιδιοτιμών της Εσσιανής με σκοπό τη μελέτη της τοπικής σύγκλισης αυτών των αλγόριθμων (Le Cun *et al.* 1991, Pearlmutter 1992) και την εκτίμηση του μεγαλύτερου δυνατού ρυθμού μάθησης για κάθε επανάληψη της διαδικασίας εκπαίδευσης (Le Cun *et al.* 1993). Στην τελευταία περίπτωση ο αλγόριθμος μάθησης επιβαρύνεται, για μερικές επαναλήψεις στην αρχή της εκπαίδευσης, με αρκετούς επιπλέον υπολογισμούς της συνάρτησης σφάλματος.

Τέλος πρέπει να αναφερθεί πως οι αλγόριθμοι μάθησης που χρησιμοποιούν την πληροφορία της Εσσιανής ή τις προσεγγίσεις της μπορούν να χρησιμοποιηθούν μόνο για μάθηση ανά ομάδα προτύπων, εξαιτίας του ισχυρού υπολογιστικού κόστους κάθε επανάληψης.

4. Αλγόριθμοι μάθησης μέγιστης κλίσης

Οι μέθοδοι μέγιστης κλίσης αποτελούν την πιο δημοφιλή κατηγορία αλγόριθμων μάθησης. Η εφαρμογή τους είναι απλή και ταυτόχρονα αποτελεσματική είτε για *μάθηση ανά ομάδα προτύπων εισόδου* (ντετερμινιστική προσέγγιση) είτε για *μάθηση ανά πρότυπο εισόδου* (στοχαστική προσέγγιση). Η κατεύθυνση ανίχνευσης των μεθόδων μέγιστης κλίσης θεωρείται τοπικά ως η κατεύθυνση της ταχύτερης μείωσης του σφάλματος μάθησης.

Όπως είναι γνωστό η ελαχιστοποίηση της συνάρτησης σφάλματος $E(\mathbf{w})$ με q παραμέτρους απαιτεί μια ακολουθία $\{\mathbf{w}_k\}_{k=0}^{\infty}$, όπου k δηλώνει επαναλήψεις, η οποία συγκλίνει σε ένα σημείο \mathbf{w}^* που ελαχιστοποιεί το σφάλμα μάθησης $E(\mathbf{w})$. Η παρακάτω επαναληπτική σχέση, που προτάθηκε από το Goldstein το 1962, χρησιμοποιείται ευρέως για την προσαρμογή των βαρών ενός ΤΝΔ:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \alpha_k \mathbf{d}_k = \mathbf{w}_k - \alpha_k \nabla E(\mathbf{w}_k), \quad (3)$$

όπου το \mathbf{d} δηλώνει μια φραγμένη απεικόνιση, ορισμένη στην περιοχή $S = \{\mathbf{w} : E(\mathbf{w}) \leq E(\mathbf{w}_0)\}$, που ικανοποιεί την ανισότητα $\nabla E(\mathbf{w})^T \mathbf{d}(\mathbf{w}) \geq 0$ έτσι ώστε για ένα $\varepsilon > 0$, υπάρχει $\delta > 0$ για το οποίο:

$$\nabla E(\mathbf{w})^T \mathbf{d}(\mathbf{w}) < \delta \Rightarrow \|\nabla E(\mathbf{w})\| < \varepsilon.$$

Επιπλέον, η ακολουθία $\{\mathbf{w}_k\}_{k=0}^{\infty}$ συγκλίνει σε ένα τοπικό ελάχιστο \mathbf{w}^* της E όταν ο ρυθμός μάθησης ικανοποιεί τη σχέση:

$$\sup \|\mathbf{H}(\mathbf{w})\| \leq \frac{1}{\alpha} < \infty$$

σε μια φραγμένη περιοχή όπου ισχύει η σχέση $E(\mathbf{w}) \leq E(\mathbf{w}_0)$ και \mathbf{H} δηλώνει την Εσσιανή της E ως προς το διάνυσμα \mathbf{w} (Goldstein, 1962).

Κάνοντας μια μικρή ιστορική αναδρομή στο πρόβλημα της επιλογής του ρυθμού μάθησης πρέπει να αναφερθεί ότι πρώτος ο Goldstein (1965) πρότεινε μια σχέση που βασίζεται στην Εσσιανή και μπορεί να χρησιμοποιηθεί για τον καθορισμό του α . Ο ίδιος επίσης απέδειξε τη σύγκλιση της μεθόδου όταν $\mathbf{d}(\mathbf{w}) \equiv -\nabla E(\mathbf{w})$, με την προϋπόθεση ότι $E \in C^2$ (δηλαδή είναι δύο φορές συνεχώς παραγωγίσιμη) και έδωσε μια εκτίμηση του ρυθμού σύγκλισης της για την περίπτωση που είναι γνωστό ένα φράγμα της νόρμας της \mathbf{H} στην περιοχή $S(\mathbf{w}_0)$. Αυτά τα αποτελέσματα επεκτάθηκαν το 1967 για την περίπτωση $\mathbf{d}(\mathbf{w}) \equiv -\mathbf{B}_k^{-1} \nabla E(\mathbf{w})$, όπου \mathbf{B}_k είναι μια προσέγγιση της Εσσιανής στο \mathbf{w}^* , χρησιμοποιώντας μόνο τιμές της συνάρτησης και του διανύσματος των παραγώγων της, ενώ παρουσιάστηκε και ένα αποτέλεσμα για τον ολικό ρυθμό σύγκλισης της μεθόδου, ο οποίος είναι *υπερ-γραμμικός* (Goldstein & Price, 1967).

Ακολουθώντας μια διαφορετική προσέγγιση που δε χρησιμοποιεί την Εσσιανή αλλά τη σταθερά *Lipschitz*, ο Armijo πρότεινε το 1966 την πρώτη μέθοδο μέγιστης κλίσης που επέτρεπε μεταβλητό a και απέδειξε τη σύγκλιση της υπό λιγότερο αυστηρές προϋποθέσεις από ότι ο Goldstein (Armijo 1966). Μια βελτίωση της μεθόδου αυτής για την εκπαίδευση νευρωνικών δικτύων πρότειναν οι Magoulas *et al.* (1997b).

Στο χώρο της εκπαίδευσης των ΤΝΔ η επιλογή του κατάλληλου ρυθμού μάθησης έχει σχετιστεί με τις ιδιοτιμές της Εσσιανής (LeCun *et al.* 1993), ο υπολογισμός των οποίων είναι επίπονος ακόμα και όταν χρησιμοποιείται μια προσέγγιση της Εσσιανής (Becker & Le Cun 1988, LeCun *et al.* 1993). Έτσι στις εφαρμογές οι χρήστες συνήθως επιλέγουν αυθαίρετα το ρυθμό μάθησης $0 < \alpha < 1$, καθώς σχετικά μικρές τιμές βοηθούν στο να διατηρηθεί η ευστάθεια του αλγόριθμου μάθησης. Το ζήτημα της εύρεσης ενός κατάλληλου ρυθμού μάθησης ώστε η ακολουθία $\{\mathbf{w}_k\}_{k=0}^{\infty}$ να συγκλίνει σε ένα ελάχιστο \mathbf{w}^* της E αποτελεί σημαντικό αντικείμενο έρευνας στα ΤΝΔ.

5. Αλγόριθμοι μάθησης με τοπική σύγκλιση

Η μάθηση με αλγόριθμους τοπικής σύγκλισης χρησιμοποιεί την πληροφορία της Εσσιανής, ή των προσεγγίσεών της. Ειδικότερα, η Εσσιανή χρησιμοποιείται στην περιγραφή των μεθόδων

συζυγών κλίσεων, παίζει βασικό ρόλο στη μέθοδο Newton, ενώ οι προσεγγίσεις της χρησιμοποιούνται στις μεθόδους μεταβλητής μετρικής.

Θεωρώντας ότι η συνάρτηση σφάλματος σε ένα σημείο \mathbf{w} προσεγγίζεται τοπικά από μια τετραγωνική συνάρτηση και αναπτύσσοντάς τη γύρω από την k εκτίμηση ενός τοπικού ελάχιστου \mathbf{w}_k , λαμβάνουμε την παρακάτω σχέση ανάμεσα στις συναρτησιακές τιμές των δύο σημείων:

$$E(\mathbf{w}) - E(\mathbf{w}_k) = \nabla E(\mathbf{w}_k)^T (\mathbf{w} - \mathbf{w}_k) + \frac{1}{2} (\mathbf{w} - \mathbf{w}_k)^T \nabla^2 E(\mathbf{w}_k) (\mathbf{w} - \mathbf{w}_k), \quad (4)$$

και η οποία έχει ένα μοναδικό ελάχιστο στο σημείο

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \nabla^2 E(\mathbf{w}_k)^{-1} \nabla E(\mathbf{w}_k), \quad (5)$$

τότε και μόνο τότε εάν η Εσσιανή $\nabla^2 E(\mathbf{w}_k)$ είναι θετικά ορισμένη. Η (5) δίνει σε κάθε επανάληψη του αλγόριθμου μάθησης μια νέα εκτίμηση \mathbf{w}_{k+1} των βέλτιστων βαρών βασισμένη σε μια προηγούμενη εκτίμησή τους \mathbf{w}_k σύμφωνα με τη μέθοδο Newton. Δυστυχώς, παρά τον τετραγωνικό ρυθμό σύγκλισης προς το ελάχιστο που χαρακτηρίζει τη μέθοδο Newton, η χρήση της στους αλγόριθμους μάθησης περιορίζεται σημαντικά από την πολυπλοκότητα των πράξεων υπολογισμού και αντιστροφής της Εσσιανής καθώς και από την τοπική της σύγκλιση. Το ζήτημα περιπλέκεται περισσότερο καθώς στην εκπαίδευση ΤΝΔ δεν παρέχεται καμία εγγύηση ότι η Εσσιανή είναι πάντοτε αντιστρέψιμη. Επιπλέον έχει βρεθεί πως έχει μεγάλο συντελεστή αστάθειας (Saarinen *et al.* 1992). Ο Battiti (1992) προτείνει διάφορες τροποποιήσεις στη μέθοδο Newton που διευκολύνουν τη χρήση της για την εκπαίδευση δικτύων πρόσθιας τροφοδότησης.

Σχετικά με την τοπική σύγκλιση των μεθόδων μεταβλητής μετρικής αναπτύχθηκαν διάφορες ολοκληρωμένες θεωρίες τη δεκαετία του 1970. Τα σπουδαιότερα αποτελέσματα οφείλονται στους Broyden *et al.* (1973) και Dennis & Moré (1974). Ένα τυπικό αποτέλεσμα τοπικής σύγκλισης ισχύει στην περιοχή ενός ελαχίστου της συνάρτησης σφάλματος και θεωρεί ότι \mathbf{w}^* είναι ένα ελάχιστο στο οποίο η Εσσιανή είναι θετικά ορισμένη. Έτσι εάν \mathbf{w}_0 είναι στην περιοχή του \mathbf{w}^* και μια προσέγγιση της Εσσιανής \mathbf{B}_0 είναι ικανοποιητικά κοντά στην πραγματική Εσσιανή $\nabla^2 E(\mathbf{w}^*)$, τότε οι επαναλήψεις των αλγόριθμων συγκλίνουν στο \mathbf{w}^* υπερ-γραμμικά.

6. Αλγόριθμοι μάθησης με ευρεία σύγκλιση

Οι παραπάνω αλγόριθμοι μάθησης δε συγκλίνουν πάντα σε ένα τοπικό ελάχιστο \mathbf{w}^* όταν η αρχική τιμή του διανύσματος βαρών \mathbf{w}_0 βρίσκεται μακριά από τη γειτονιά του τοπικού ελαχίστου. Αντίθετα ένας αλγόριθμος που έχει την *ιδιότητα της ευρείας σύγκλισης* (globally convergent algorithm) συγκλίνει σε ένα ελάχιστο ξεκινώντας από οποιαδήποτε αρχική συνθήκη (Dennis & Schnabel 1983, Byrd *et al.* 1987). Αυτή η συμπεριφορά διευκολύνει ιδιαίτερα τη διαδικασία εκπαίδευσης ΤΝΔ καθώς τις περισσότερες φορές η μάθηση 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 μάθησης.

Η ευρεία σύγκλιση επιτυγχάνεται εφαρμόζοντας μεθόδους *μη ακριβούς ευθύγραμμης ανίχνευσης* (Brown & Saad 1990, 1994, Eisenstat & Walker 1994) και μεθόδους *ασφαλούς περιοχής* (Powell

1975, Sorensen 1982, Moré 1983, Dennis & Schnabel 1983, Schultz *et al.* 1985, Møller 1993). Στη συνέχεια, η ανάλυση της σύγκλισης των αλγόριθμων μάθησης εστιάζεται στη χρήση μεθόδων μη ακριβούς ευθύγραμμης 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 οποιοδήποτε αλγόριθμο που επαναληπτικά προσαρμόζει τα βάρη ακολουθώντας κατευθύνσεις μείωσης της συνάρτησης σφάλματος εξασφαλίζει στον αλγόριθμο την ιδιότητα της ευρείας σύγκλισης (Magoulas *et al.* 2000b, Vrahatis *et al.* 2000a).

7. Βασικές αρχές σύγκλισης

Οι ιδιότητες σύγκλισης των μεθόδων ευθύγραμμης ανίχνευσης μπορούν να μελετηθούν μετρώντας την ποιότητα της κατεύθυνσης ανίχνευσης, όπως αυτή ορίζεται από τη γωνία που σχηματίζεται μεταξύ της κατεύθυνσης μέγιστης κλίσης $-\nabla E(\mathbf{w}_k)$ και της κατεύθυνσης ανίχνευσης, δηλαδή

$$\cos \theta_k \equiv \frac{-\nabla E(\mathbf{w}_k)^T \mathbf{d}_k}{\|\nabla E(\mathbf{w}_k)\| \|\mathbf{d}_k\|} \quad (6)$$

και λαμβάνοντας υπόψη το μήκος του βήματος.

Το μήκος του βήματος καθορίζεται από μια επανάληψη της μεθόδου ευθύγραμμης ανίχνευσης. Μια στρατηγική που προτείνεται δέχεται ως βήμα ένα θετικό αριθμό που ικανοποιεί τις παρακάτω συνθήκες:

$$\begin{aligned} E(\mathbf{w}_k + \alpha_k \mathbf{d}_k) &\leq E(\mathbf{w}_k) + \sigma_1 \alpha_k \nabla E(\mathbf{w}_k)^T \mathbf{d}_k, \\ \nabla E(\mathbf{w}_k + \alpha_k \mathbf{d}_k)^T \mathbf{d}_k &\geq \sigma_2 \nabla E(\mathbf{w}_k)^T \mathbf{d}_k, \end{aligned} \quad (7)$$

όπου $0 < \sigma_1 < \sigma_2 < 1$. Οι δύο ανισότητες είναι γνωστές ως *συνθήκες Wolfe*. Η πρώτη ανισότητα εξασφαλίζει ότι η συνάρτηση σφάλματος μειώνεται επαρκώς σε κάθε επανάληψη του αλγόριθμου μάθησης, ενώ η δεύτερη εμποδίζει το ρυθμό μάθησης 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 της κλίσης να είναι μικρότερη από 90° (Wolfe 1969, 1970).

Εύκολα μπορεί να αποδειχθεί ότι εάν η \mathbf{d}_k είναι μια κατεύθυνση μείωσης του αλγόριθμου μάθησης και η συνάρτηση σφάλματος είναι παραγωγίσιμη και από κάτω φραγμένη κατά μήκος της ακτίνας $\{\mathbf{w}_k + \alpha \mathbf{d}_k \mid \alpha > 0\}$, τότε πάντοτε υπάρχουν μήκη βημάτων που ικανοποιούν τις συνθήκες (7) (Wolfe 1969, 1971).

Το παρακάτω θεώρημα 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.

Θεώρημα 1 (Zoutendijk 1970, Wolfe 1969, 1970). *Υποθέτουμε ότι η συνάρτηση E είναι κάτω φραγμένη στο \mathfrak{R}^q . Επίσης ότι για ένα αρχικό διάνυσμα βαρών $\mathbf{w}_0 \in \mathfrak{R}^q$ και για κάθε \mathbf{w} σε μια περιοχή που περιέχει το αρχικό διάνυσμα βαρών, $S = \{\mathbf{w} : E(\mathbf{w}) \leq E(\mathbf{w}_0)\}$, η E είναι συνεχώς παραγωγίσιμη στο $S(\mathbf{w}_0)$ και Lipschitz συνεχής, δηλαδή υπάρχει μια σταθερά $L > 0$ τέτοια ώστε*

$$\|\nabla E(\mathbf{w}_1) - \nabla E(\mathbf{w}_2)\| \leq L \|\mathbf{w}_1 - \mathbf{w}_2\|, \quad \forall \mathbf{w}_1, \mathbf{w}_2 \in \mathfrak{R}^q. \quad (8)$$

Τότε για κάθε επανάληψη της μορφής (1) και για οποιαδήποτε κατεύθυνση μείωσης \mathbf{d}_k όπου α_k ικανοποιεί τις συνθήκες του Wolfe (7) ισχύει ότι

$$\sum_{k \geq 1} \cos^2 \theta_k \|\nabla E(\mathbf{w}_k)\|^2 < \infty. \quad (9)$$

Σκοπός της ανισότητας (8) είναι να τεθεί θεωρητικά ένα άνω φράγμα στο βαθμό της μη γραμμικότητας της συνάρτησης σφάλματος και να διασφαλιστεί ότι οι πρώτες παράγωγοι είναι συνεχείς. Η ανισότητα (9) ονομάζεται η *ανισότητα του Zoutendijk* και είναι χρήσιμη για την απόδειξη ευρείας σύγκλισης πολλών αλγόριθμων μάθησης. Έτσι αν οι επαναλήψεις (1) είναι τέτοιες ώστε ισχύει

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

για όλα τα k . Τότε συμπεραίνουμε απευθείας από την (9) ότι

$$\lim_{k \rightarrow \infty} \|\nabla E(\mathbf{w}_k)\| = 0. \quad (10)$$

Δηλαδή, εάν η κατεύθυνση ανίχνευσης δεν είναι ορθογώνια στην κλίση $-\nabla E(\mathbf{w}_k)$, τότε η ακολουθία των κλίσεων συγκλίνει στο μηδέν.

Σύμφωνα με αυτό το αποτέλεσμα ένας αλγόριθμος μάθησης που βασίζεται στη μέθοδο μέγιστης κλίσης έχει την ιδιότητα της ευρείας σύγκλισης αν χρησιμοποιεί ευθύγραμμη ανίχνευση που ικανοποιεί τις συνθήκες του Wolfe για τον καθορισμό του ρυθμού μάθησης. Σε αυτή την περίπτωση έχουμε $\cos \theta_k = 1$ για όλα τα k και ισχύει η (10). Ας σημειωθεί πως για μεθόδους ευθύγραμμης ανίχνευσης της μορφής (1), το όριο (10) αποτελεί το καλύτερο αποτέλεσμα ευρείας σύγκλισης που μπορεί να επιτευχθεί.

Για την ευρεία σύγκλιση αλγόριθμων που βασίζονται σε μεθόδους Newton και μεταβλητής μετρικής μπορεί να χρησιμοποιηθεί πάλι η (9) για να πάρουμε το αποτέλεσμα (10). Για το σκοπό αυτό υποθέτουμε ότι ο πίνακας \mathbf{B}_k σε κάθε επανάληψη είναι θετικά ορισμένος (αυτό χρειάζεται για να κινηθούμε προς κατεύθυνση μείωσης) και ότι ο συντελεστής αστάθειάς του είναι φραγμένος για όλα τα k , δηλαδή:

$$\|\mathbf{B}_k\| \|\mathbf{B}_k^{-1}\| \leq \Delta,$$

για μια σταθερά $\Delta > 0$. Εάν επιπλέον η ευθύγραμμη ανίχνευση ικανοποιεί τις συνθήκες του Wolfe τότε από την (6) έχουμε ότι $\cos \theta_k \geq 1/\Delta$. Για αλγόριθμους που βασίζονται σε μεθόδους συζυγών κλίσεων δεν είναι δυνατό να δειχθεί ότι ισχύει το όριο (10) παρά μόνο η παρακάτω ασθενέστερη σχέση:

$$\liminf_{k \rightarrow \infty} \|\nabla E(\mathbf{w}_k)\| = 0. \quad (11)$$

Το συμπέρασμα που προκύπτει από τα παραπάνω είναι ότι για την ανάπτυξη αλγόριθμων μάθησης με καλές ιδιότητες σύγκλισης απαιτείται να εξασφαλίζεται ότι η κατεύθυνση ανίχνευσης δε γίνεται ορθογώνια με το διάνυσμα της κλίσης, ή ότι παρεμβάλλονται, μεθοδικά, επαναλήψεις της μεθόδου μέγιστης κλίσης με το α_k να ικανοποιεί τις συνθήκες του Wolfe.

Στην πράξη μια τεχνική που μπορεί να χρησιμοποιηθεί για να εξασφαλιστεί η σύγκλιση είναι να ελέγχεται η γωνία μεταξύ της κατεύθυνση ανίχνευσης και του διανύσματος της κλίσης και στην περίπτωση που είναι μικρότερη από μια προκαθορισμένη σταθερά τότε να αλλάζουμε την κατεύθυνση ανίχνευσης προς την κατεύθυνση της μέγιστης κλίσης. Επίσης έχει προταθεί η πληροφορία της γωνίας να λαμβάνεται υπόψη κατά την προσαρμογή των βαρών (Swanston *et al.* 1994, Hsin *et al.* 1995).

Ωστόσο ο υπολογισμός και ο έλεγχος της γωνίας σε κάθε επανάληψη αυξάνει την πολυπλοκότητα του αλγόριθμου χωρίς να αυξάνει απαραίτητα την ταχύτητα σύγκλισης σε ένα ελάχιστο της μη κυρτής συνάρτησης σφάλματος. Στο σημείο αυτό πρέπει να σημειωθεί πως η ευρεία σύγκλιση δεν είναι το μόνο επιθυμητό χαρακτηριστικό ενός αλγόριθμου μάθησης. Η ταχύτητα της μάθησης αποτελεί συχνά σημαντικότερο χαρακτηριστικό.

Για τη μελέτη της ταχύτητας σύγκλισης ενός αλγόριθμου είναι χρήσιμο ένα κλασικό αποτέλεσμα των Dennis & Moré (1974). Σύμφωνα με τους Dennis και Moré η επαναληπτική σχέση (1) συγκλίνει με ταχύτητα υπερ-γραμμική (superlinear) μόνο όταν

$$\alpha_k \mathbf{d}_k = \mathbf{d}_k^N + O\left(\|\mathbf{d}_k^N\|\right), \quad (12)$$

όπου \mathbf{d}_k^N είναι η κατεύθυνση Newton. Επομένως για να επιτύχουμε γρήγορη μάθηση είναι απαραίτητο να προσεγγίσουμε ασυμπτωτικά την κατεύθυνση Newton. Ο έλεγχος της γωνίας μπορεί να εμποδίσει τη γρήγορη μάθηση, π.χ. ένας αλγόριθμος μάθησης που βασίζεται στη μέθοδο BFGS μπορεί να δημιουργήσει ασταθείς προσεγγίσεις \mathbf{B}_k της Εσσιανής, δηλαδή πίνακες με μεγάλο συντελεστή αστάθειας. Σε τέτοιες περιπτώσεις δε μπορεί να καθοριστεί εάν αυτή η συμπεριφορά είναι επιθυμητή, ή εάν οι πίνακες \mathbf{B}_k προσεγγίζουν ικανοποιητικά μια Εσσιανή με μεγάλο συντελεστή αστάθειας (ill-conditioned). Για να αποφανθούμε θα έπρεπε να γνωρίζουμε το ίδιο το πρόβλημα που προσπαθούμε να επιλύσουμε. Έτσι στις εφαρμογές προτιμούμε όταν εκπαιδεύουμε ΤΝΔ με τη μέθοδο BFGS να αφήνουμε τους πίνακες \mathbf{B}_k να εξελίσσονται ελεύθερα γιατί τότε επιταχύνεται η μάθηση.

8. Πρακτική θεώρηση της ευρείας σύγκλισης αλγορίθμων μάθησης

Τα προηγούμενα θεωρητικά αποτελέσματα είναι χρήσιμα για την κατανόηση και τη μελέτη της συμπεριφοράς των αλγορίθμων μάθησης. Ωστόσο οι αλγόριθμοι αυτοί, 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:

- (i) *Το κόστος υπολογισμού των τιμών της συνάρτησης σφάλματος και των παραγώγων της.* Στις εφαρμογές το υπολογιστικό κόστος αποτελεί το βασικό κριτήριο επιλογής του αλγόριθμου μάθησης, καθώς σε πολλές περιπτώσεις είναι προτιμότερες μερικές ακόμα επαναλήψεις ενός αλγόριθμου που βασίζεται στη μέθοδο μέγιστης κλίσης από τη χρήση περίπλοκων αλγορίθμων τοπικής σύγκλισης.
- (ii) *Οι μη ακριβείς τιμές της συνάρτησης σφάλματος.* Είναι γνωστό πως οι αριθμητικοί υπολογισμοί υπόκεινται σε σφάλματα ακρίβειας (Wilkinson 1963). Οι αριθμητικές πράξεις που απαιτούνται στις εξομοιώσεις των αλγορίθμων μάθησης επηρεάζουν την ακρίβεια των τιμών της συνάρτησης σφάλματος (Wray & Green 1995). Επιπλέον, τα χαρακτηριστικά των μη γραμμικών νευρώνων εμποδίζουν τον ακριβή υπολογισμό των συναρτησιακών τιμών του σφάλματος και οδηγούν σε κορεσμό, τόσο στις εξομοιώσεις όσο και στις υλοποιήσεις των ΤΝΔ (Holt & Hwang 1993).
- (iii) *Τα πολλαπλά ελάχιστα της συνάρτησης σφάλματος.* Η συνάρτηση σφάλματος είναι μη 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 πολλαπλά «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 μάθησης 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 νευρώνων δεν επαρκεί για τη συγκεκριμένη εφαρμογή ή όταν ο αλγόριθμος εκκινεί από ακατάλληλα αρχικά βάρη και εμποδίζει το ΤΝΔ από το να μάθει πλήρως όλα τα πρότυπα.

9. Ευρεία σύγκλιση με στρατηγικές οπισθοδρόμησης

Οι αλγόριθμοι μάθησης με στρατηγικές ευθύγραμμης ανίχνευσης που ικανοποιούν τις συνθήκες του Wolfe ελαττώνουν το πλήθος των υπολογισμών της συνάρτησης σφάλματος και των παραγώγων της σε σχέση με τους αλγόριθμους τοπικής σύγκλισης που χρησιμοποιούν τεχνικές ακριβούς ευθύγραμμης ανίχνευσης.

Στην πράξη είναι δυνατό ο αριθμός των απαιτούμενων υπολογισμών να μειωθεί περαιτέρω χρησιμοποιώντας μια διαφορετική στρατηγική ευθύγραμμης ανίχνευσης που ονομάζεται *οπισθοδρόμηση* (backtracking) και τις τροποποιήσεις της. Η στρατηγική οπισθοδρόμησης εκμεταλλεύεται τον καλύτερο ρυθμό μάθησης που μπορεί να βρεθεί στο διάστημα $(0, \alpha_{\max}]$, όπου α_{\max} είναι ο μέγιστος ρυθμός μάθησης της εκάστοτε επανάληψης. Η ανίχνευση ξεκινά από το μέγιστο ρυθμό και εκτελώντας διαδοχικές ελαττώσεις του προχωρά προς το ελάχιστο έως ότου επιτευχθεί ικανοποιητική μείωση της συνάρτησης σφάλματος. Για κριτήριο ικανοποιητικής μείωσης μπορούμε να χρησιμοποιούμε την πρώτη από τις συνθήκες του Wolfe (7) που εξασφαλίζει ότι η συνάρτηση σφάλματος μειώνεται επαρκώς σε κάθε επανάληψη του αλγόριθμου μάθησης. Η παράμετρος $\sigma_1 \in (0,1)$ της πρώτης ανισότητας του Wolfe μπορεί να λαμβάνει μικρές τιμές σε επίπεδες επιφάνειες σφάλματος, ώστε να περιορίζεται ο αριθμός των συναρτησιακών υπολογισμών αφού οι διαφορές των σφαλμάτων για τις διαδοχικές τιμές του ρυθμού είναι μηδαμινές, και μεγάλες τιμές σε απότομες επιφάνειες, ώστε να εξασφαλίζεται η μεγαλύτερη δυνατή μείωση της συνάρτησης σφάλματος. Η τιμή $\sigma_1 = 0.5$, χωρίς να είναι πάντοτε η βέλτιστη, αποδείχθηκε στην πράξη πως κατά μέσο όρο περιορίζει τον αριθμό των συναρτησιακών υπολογισμών και χρησιμοποιήθηκε στις παρακάτω εξομοιώσεις.

Οι διαδοχικές ελαττώσεις του ρυθμού μάθησης συντελούνται με διάφορους μηχανισμούς, όπως η διαίρεση με ένα σταθερό συντελεστή ή με μια εκτίμηση του βέλτιστου συντελεστή για κάθε επανάληψη. Πρέπει να σημειωθεί πως η επιλογή του διαιρέτη δεν είναι κρίσιμη για την ευρεία σύγκλιση. Για παράδειγμα, ο διαιρέτης θα μπορούσε να είναι σταθερός και ίσος με 2. Ένας διαιρέτης 10, ή 20, είναι πιθανόν καταλληλότερος σε περιπτώσεις που απαιτείται ο ρυθμός να κρατείται χαμηλός για πολλές διαδοχικές επαναλήψεις. Βέβαια τυχόν υπερβολική μείωση του ρυθμού μάθησης μπορεί τελικά να αυξήσει το υπολογιστικό κόστος, επιβαρύνοντας τον αλγόριθμο με επιπλέον επαναλήψεις.

Για το λόγο αυτό ο αλγόριθμος μπορεί να ενσωματώνει και μηχανισμούς άνω και κάτω φράγματος που εξασφαλίζουν ότι ο ρυθμός μάθησης κινείται σε επίπεδα που δεν επιβραδύνουν επικίνδυνα την εκπαίδευση, ενώ παράλληλα μειώνουν το σφάλμα σε κάθε επανάληψη. Τα φράγματα αυτά έχουν το ίδιο θεωρητικό αποτέλεσμα με τη δεύτερη από τις συνθήκες (7) (Dennis & Schnabel 1983), εμποδίζοντας το ρυθμό μάθησης από το να γίνει πολύ αργός, χωρίς να απαιτούν επιπλέον υπολογισμούς παραγώγων της συνάρτησης σφάλματος.

Πρέπει να σημειωθεί πως η στρατηγική οπισθοδρόμησης με φράγματα στο ρυθμό μάθησης επιτρέπει να εγγυηθούμε την ευρεία σύγκλιση για τους αλγόριθμους μάθησης χωρίς τη χρήση ευθύγραμμης ανίχνευσης που ικανοποιεί τις συνθήκες του Wolfe (βλέπε το σχετικό αποτέλεσμα των Shultz *et al.* 1982). Έτσι στη πράξη με αυτή τη στρατηγική μπορούμε να εξασφαλίσουμε την ευστάθεια του αλγόριθμου μάθησης και τη *σθεναρότητά* του σε ταλαντώσεις.

10. Ανάλυση αλγορίθμων σε ένα κλασσικό παράδειγμα

Η ταξινόμηση των τεσσάρων προτύπων του Αποκλειστικού-ΕΙΤΕ (XOR) σε δύο κατηγορίες, την 0 και την 1, είναι ένα κλασσικό πρόβλημα για ΤΝΔ (Jacobs 1988, Kollias & Anastasiou 1989, van Ooyen 1992, van der Smagt 1994). Το ΤΝΔ έχει 2 γραμμικούς νευρώνες στην είσοδο, 2 μη γραμμικούς νευρώνες, που ακολουθούν τη λογιστική συνάρτηση (Haykin 1994), στο κρυμμένο επίπεδο και 1 ίδιου τύπου μη γραμμικό νευρώνα στην έξοδο. Αυτή η αρχιτεκτονική δηλώνεται για ευκολία 2-2-1 και περιλαμβάνει 9 βάρη.

Η μάθηση του XOR παρουσιάζει μεγάλη εξάρτηση από τις αρχικές τιμές των βαρών και τις μεταβολές του ρυθμού μάθησης, ενώ εμφανίζει πολλαπλά ελάχιστα. Το ζήτημα των πολλαπλών ελάχιστων σε αυτή την εφαρμογή έχει μελετηθεί αναλυτικά (Blum 1989, Lisboa & Perantonis 1991). Τα Σχήματα 1 και 2 παρουσιάζουν την επιφάνεια σφάλματος του XOR και τις ισοϋψείς της στο διδιάστατο υπόχωρο του \mathcal{R}^q που ορίζεται από τα ιδιοδιανύσματα που αντιστοιχούν στις ακραίες ιδιοτιμές της Εσσιανής, δηλαδή στη μεγαλύτερη και στη μικρότερη ιδιοτιμή της Εσσιανής της συνάρτησης σφάλματος, σε ένα επιθυμητό ελάχιστο \mathbf{w}^* το οποίο αντιστοιχίζεται στο σημείο (0, 0) των σχημάτων. Η προσέγγιση αυτή βασίζεται στο ότι η Εσσιανή είναι πραγματικός και συμμετρικός πίνακας και επομένως όλες οι ιδιοτιμές της είναι πραγματικές και τα ιδιοδιανύσματα που αντιστοιχούν στις διάφορες ιδιοτιμές είναι ορθογώνια. Σε αυτό το πλαίσιο οι κατευθύνσεις των αξόνων των ισοϋψών γραμμών δίνονται από τα ιδιοδιανύσματα της Εσσιανής, ενώ το μήκος των αξόνων είναι αντίστροφα ανάλογο της τετραγωνικής ρίζας των αντίστοιχων ιδιοτιμών.

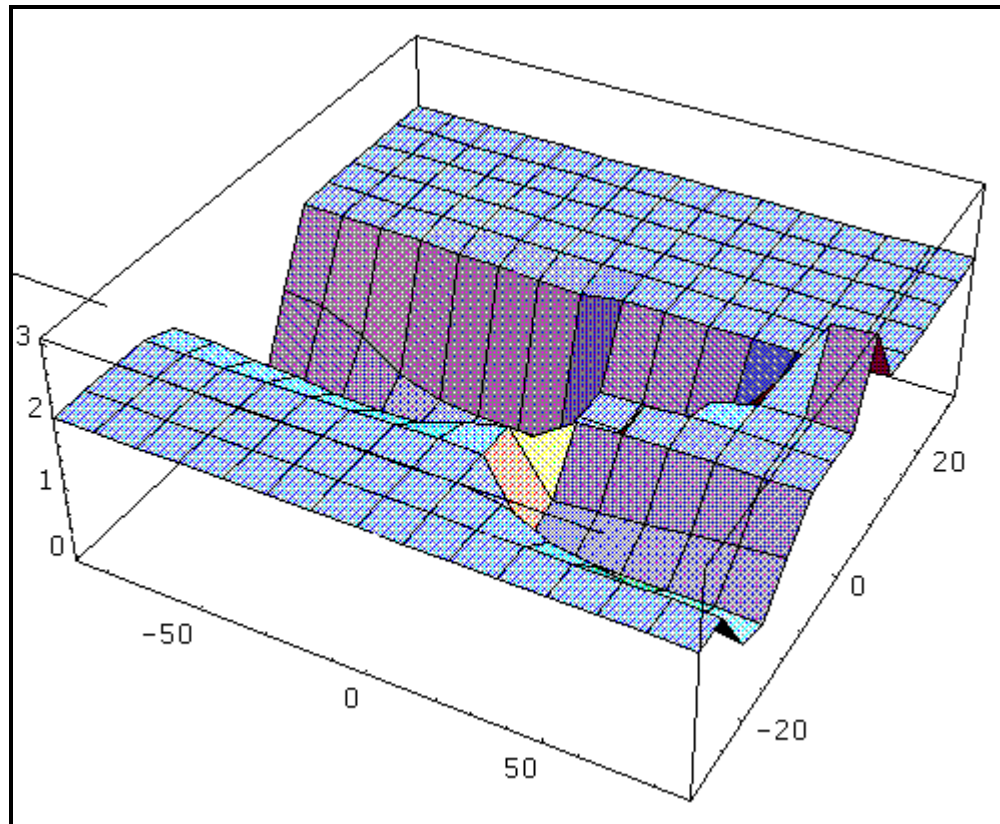
Οι επίπεδες περιοχές του Σχήματος 1 δηλώνουν σημεία με την ίδια συναρτησιακή τιμή. Η μορφή αυτή αναπαράστασης μιας γενικής επιφάνειας σφάλματος με $(q+1)$ διαστάσεις στις 3 διαστάσεις προτάθηκε από τους Androulakis *et al.* (1997). Στη συγκεκριμένη περίπτωση αυτή η προσέγγιση μας επιτρέπει να οπτικοποιήσουμε την επιφάνεια σφάλματος του 9-διάστατου προβλήματος XOR γύρω από ένα ελάχιστο και να μελετήσουμε τη συμπεριφορά διάφορων αλγορίθμων μάθησης γύρω από το ελάχιστο και την ευαισθησία τους στις αρχικές τιμές των βαρών.

Στη συνέχεια θα συγκρίνουμε την απόδοση πέντε διαδεδομένων και πολυ γνωστών αλγορίθμων μάθησης στα ίδια 64000 αρχικά διανύσματα βαρών γύρω από ένα τυχαίο ελάχιστο. Πιο συγκεκριμένα οι αλγόριθμοι ελέγχονται στα σημεία:

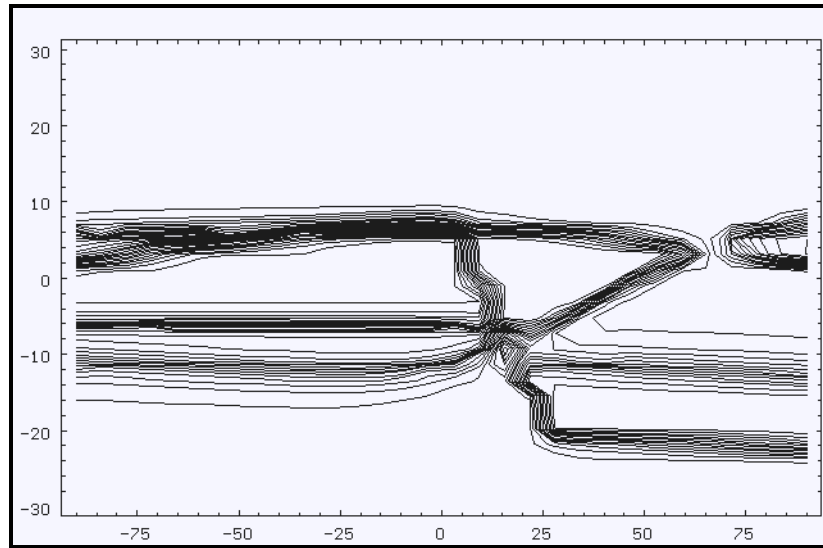
$$\mathbf{w} = \mathbf{w}^* + (c_1 e^{\max} + c_2 e^{\min}),$$

παίρνοντας κατάλληλο πλέγμα για τις παραμέτρους $c_1 \in [-50,50]$, $c_2 \in [-90,90]$, το οποίο καθορίζει τις συντεταγμένες των εικονοστοιχείων (οθόνη VGA, 320x200, 256 χρώματα) και χρησιμοποιώντας τα ιδιοδιανύσματα e^{\max}, e^{\min} που αντιστοιχούν στη μέγιστη και ελάχιστη ιδιοτιμή της Εσσιανής στο ελάχιστο \mathbf{w}^* .

Οι αλγόριθμοι των οποίων αποτελέσματα σύγκλισης παρουσιάζονται είναι οι εξής: μέγιστης κλίσης (SD), Fletcher-Reeves (FR), Polak-Ribière (PR), Davidon-Fletcher-Powell (DFP) και Broyden-Fletcher-Goldfarb-Shanno (BFGS), οι οποίοι μπορούν να βρεθούν στη βιβλιογραφία αναφορικά με την βελτιστοποίηση χωρίς περιορισμούς (Dennis & Schnabel 1983, Luenberger 1969, Polak (1997), Ortega & Rheinboldt 1970). Σε όλες τις περιπτώσεις η κλίση υπολογίστηκε με την τεχνική της οπισθοδρομικής διάδοσης του σφάλματος και χρησιμοποιήθηκε η στρατηγική της οπισθοδρόμησης για τον καθορισμό του ρυθμού μάθησης σε κάθε επανάληψη. Η εκπαίδευση θεωρείται επιτυχής όταν το σφάλμα μάθησης γίνεται $E \leq 0.04$. Τα αποτελέσματα συνοψίζονται στον Πίνακα 1 όπου χρησιμοποιούνται τα εξής σύμβολα: μ είναι η μέση τιμή του αριθμού των υπολογισμών της συνάρτησης σφάλματος, σ είναι η τυπική απόκλιση και Min/Max είναι ο ελάχιστος/μέγιστος αριθμός συναρτησιακών υπολογισμών που εκτελέστηκαν. Τέλος, % είναι το ποσοστό επιτυχίας του αλγόριθμου μάθησης που ταυτίζεται με το ποσοστό σύγκλισης σε «επιθυμητά» ελάχιστα, δηλαδή σε βάρη τα οποία δίνουν τιμές σφάλματος $E \leq 0.04$.



Σχήμα 1: Η επιφάνεια σφάλματος του προβλήματος Αποκλειστικού-ΕΙΤΕ (XOR), γύρω από ένα επιθυμητό ελάχιστο, απεικονισμένη στο διδιάστατο υπόχωρο που ορίζεται από τα ακραία ιδιοδιανύσματα της Εσσιανής. Τα ιδιοδιανύσματα υπολογίζονται στο συγκεκριμένο ελάχιστο, το οποίο στο σχήμα αντιστοιχίζεται στην αρχή των αξόνων.



Σχήμα 2: Ισοϋψείς γραμμές της επιφάνειας σφάλματος του προβλήματος Αποκλειστικού-ΕΙΤΕ (XOR) απεικονισμένες στο διδιάστατο υπόχωρο που ορίζεται από τα ακραία ιδιοδιανυσμάτα της Εσσιανής.

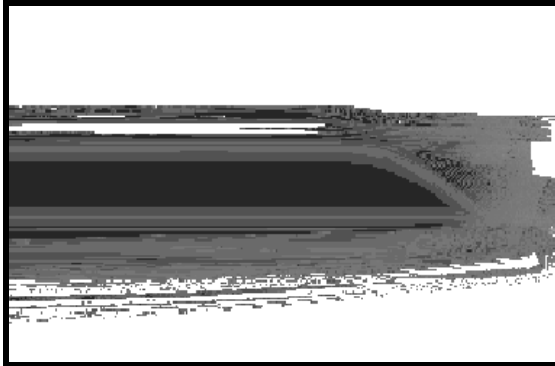
Πίνακας 1: Συγκριτικά αποτελέσματα για το πρόβλημα XOR.

Αλγόριθμος	μ	σ	Min/Max	%
SD	83	309.41	7/9856	47.7
FR	804	2357.01	20/24015	43.9
PR	270	152.05	20/2383	48.2
DFP	565	2252.78	20/25307	52.4
BFGS	175	136.04	20/1691	55.9

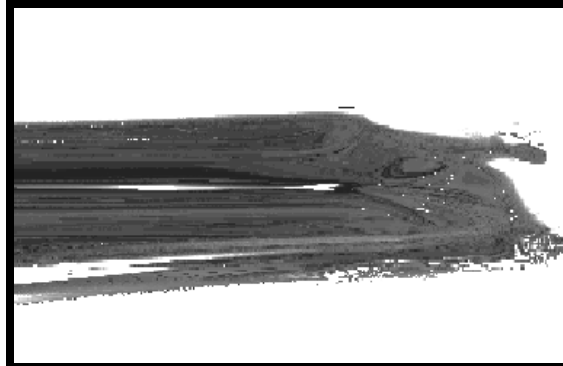
Τα αποτελέσματα του Πίνακα 1 δείχνουν ότι η μέθοδος της μέγιστης κλίσης απαιτεί τον μικρότερο αριθμό υπολογισμών της συνάρτησης σφάλματος και είναι η ταχύτερη στο μέσο όρο. Σημειώνοντας ότι το ποσοστό επιτυχίας των αλγόριθμων συνδέεται με το πρόβλημα της σύγκλισης σε ανεπιθύμητα ελάχιστα, που είναι ιδιαίτερα οξυμένο στο XOR, παρατηρούμε ότι η BFGS και η DFP εμφανίζουν τα καλύτερα ποσοστά επιτυχίας. Το ποσοστό επιτυχίας της μεθόδου μέγιστης κλίσης είναι αρκετά ικανοποιητικό, προσεγγίζοντας το ποσοστό της μεθόδου Polak-Ribière, που είναι υπολογιστικά πολυπλοκότερη.

Ενδιαφέρον παρουσιάζει η συμπεριφορά των δύο αλγόριθμων που βασίζονται σε συζυγείς κλίσεις. Η συμπεριφορά της μεθόδου Fletcher-Reeves επαληθεύει τις θεωρητικές αναφορές και υπολείπεται της μεθόδου Polak-Ribière. Πέρα από το γεγονός ότι η μέθοδος Polak-Ribière έχει καλύτερα ποσοστά επιτυχίας και αισθητά καλύτερη μέση συμπεριφορά από τη συγγενή της μέθοδο, εμφανίζει έναν από τους δύο χαμηλότερους μέγιστους αριθμούς υπολογισμών των τιμών της συνάρτησης σφάλματος, γεγονός που δίνει ιδιαίτερη βαρύτητα στη μέση συμπεριφορά της μεθόδου. Είναι επίσης ενδιαφέρον να παρατηρήσουμε στο Σχήμα 3 την ευαισθησία των αλγόριθμων από τα αρχικά βάρη, όπως φαίνεται από την κατανομή των σημείων στην περιοχή σύγκλισης απεικονισμένη στο διδιάστατο υπόχωρο του \mathcal{R}^q .

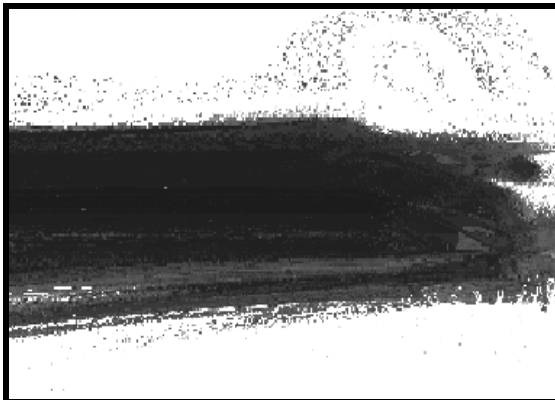
Σχήμα 3: Σύγκριση πέντε αλγόριθμων μάθησης στο πρόβλημα ταξινόμησης προτύπων Αποκλειστικού-ΕΙΤΕ (XOR). Κάθε σημείο των παρακάτω εικόνων αντιστοιχεί σε ένα αρχικό διάνυσμα βαρών και χρωματίζεται μαύρο ή άσπρο ανάλογα με το αν ο αλγόριθμος, με τη συγκεκριμένη αρχική τιμή, συγκλίνει σε επιθυμητό ελάχιστο, ή όχι. Τα σημεία ορίζονται στο πλέγμα $[-50, 50] \times [-90, 90]$ και το κέντρο της εικόνας αντιστοιχεί στο σημείο $(0,0)$.



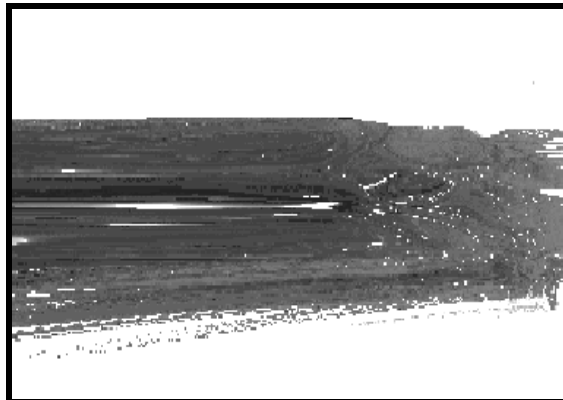
Περιοχή σύγκλισης του αλγόριθμου μέγιστης κλίσης.



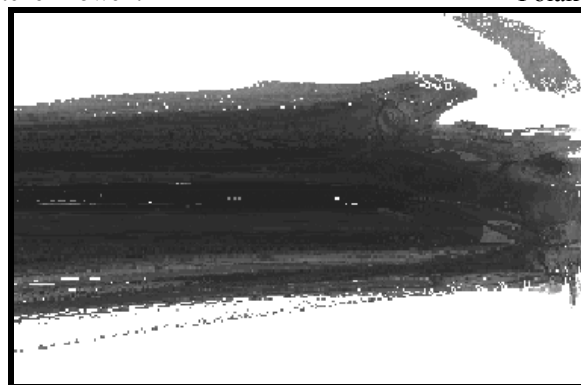
Περιοχή σύγκλισης του αλγόριθμου Fletcher-Reeves.



Περιοχή σύγκλισης του αλγόριθμου Davidon-Fletcher-Powell.



Περιοχή σύγκλισης του αλγόριθμου Polak-Ribière.



Περιοχή σύγκλισης του αλγόριθμου Broyden-Fletcher-Goldfarb-Shanno.

Αναφορές

1. Ackley, D., Hinton, G. & Sejnowski, T. (1985). A learning algorithm for Boltzmann machines, *Cognitive Science*, 9, 147-169.
2. Almeida, L. (1987). A learning rule for asynchronous perceptrons with feedback in a combinatorial environment, *Proceedings of the 1st IEEE Int. Conf. on Neural Networks*, vol. 2, San Diego, CA, 609-618.
3. Androulakis, G.S, Magoulas, G.D. & Vrahatis, M.N. (1997). Geometry of learning: visualizing the performance of neural network supervised training methods, *Nonlinear Analysis, Theory, Methods & Applications*, 30, No. 7, 4539-4544.
4. Armijo, L. (1966). Minimization of functions having Lipschitz continuous first partial derivatives, *Pacific J. of Math.*, 16, 1-3.
5. Battiti, R. (1989). Accelerated backpropagation learning: two optimization methods, *Complex Systems*, 3, 331-342.
6. Battiti, R. (1992). First- and second-order methods for learning: between steepest descent and Newton's method, *Neural Computation*, 4, 141-166.
7. Battiti, R. & Masulli, F. (1990). BFGS optimization for faster and automated supervised learning, in *Proceedings of the Inter. Conf. on Neural Networks*, Paris, France, 757-760.
8. Becker, S. & Le Cun, Y. (1988). Improving the convergence of the back-propagation learning with second order methods, in *Proceedings of the 1988 Connectionist Models Summer School* (Touretzky, D., Hinton, G. & Sejnowski, T., eds.), Morgan-Kaufman, San Mateo, CA, 29-37.
9. Beigi, H. & Li, C. (1993). Learning algorithms for neural networks based on quasi-Newton methods with self-scaling, *Trans. of the ASME, Journal of Dynamic Systems, Measurements, and Control*, 115, 38-43.
10. Bishop, C. (1992). Exact calculation of the Hessian matrix for the multi-layer perceptron, *Neural Computation*, 4, 494-501.
11. Blum, E. (1989). Approximation of Boolean functions by sigmoidal networks: Part I: XOR and other two-variable functions, *Neural Computation*, 1, 532-540.
12. Brown, P. & Saad, Y. (1990). Hybrid Krylov methods for nonlinear systems of equations, *SIAM J. Sci. Stat. Comput.*, 11, 450-481.
13. Broyden, G., Dennis, J. & Moré, J. (1973). On the local and superlinear convergence of quasi-Newton methods, *J. Inst. Math. Applies*, 12, 223-246.
14. Buntine, W. & Weigend, A. (1994). Computing second derivatives on feedforward networks: a review, *IEEE Trans. on Neural Networks*, 5, 480-488.
15. Byrd, R., Nocedal, J. & Yuan Y. (1987). Global convergence of a class of quasi-Newton methods on convex problems, *SIAM J. Numer. Anal.*, 24, 1171-1190.
16. Dennis, J. & Moré, J. (1974). A characterization of superlinear convergence and its application to quasi-Newton methods, *Math. Comp.*, 28, 549-560.
17. Dennis, J & Schnabel, R. (1983). *Numerical methods for unconstrained optimization and nonlinear equations*, Prentice-Hall, NJ.
18. Eisenstat, S. & Walker, H (1994). Globally convergent inexact Newton methods, *SIAM J. Optim.*, 4, 393-422.
19. El-Jaroudi, A. & Makhoul, J. (1990). A new error criterion for posterior probability estimation with neural nets, in *Proceedings of the Inter. J. Conf. on Neural Networks*, Vol. 3, San Diego, CA, 185-192.
20. Fahlman, S. (1988). Faster learning variations on back-propagation: an empirical study, in *Proceedings of the 1988 Connectionist Models Summer School* (Touretzky, D., Hinton, G. & Sejnowski, T., eds.), Morgan-Kaufman, San Mateo, CA, 38-51.

21. Goldstein, A. (1962). Cauchy's method of minimization, *Numerische Mathematik*, 4, 146-150.
22. Goldstein, A. (1965). On steepest descent, *SIAM J. Control*, 3, 147-151.
23. Goldstein, A. & Price, J. (1967). An effective algorithm for minimization, *Numerische Mathematik*, 10, 184-189.
24. Hassibi, B. & Stork, D. (1993). Second order derivatives for network pruning: Optimal brain surgeon, in *Advances in Neural Information Processing Systems 5* (Hanson, S., Cowan, J. & Giles, C., eds.), Morgan Kaufmann, San Mateo, CA, 164-171.
25. Haykin, S. (1994). *Neural Networks: a comprehensive foundation*, Macmillan College Publishing Company.
26. Holt, J. & Hwang, J. (1993). Finite precision error analysis of neural network hardware implementations, *IEEE Trans. on Comp.*, 42, 281-290.
27. Hsin, H.-C., Li, C.-C., Sun, M. & Scabassi, R. (1995). An adaptive training algorithm for back-propagation neural networks, *IEEE Trans. on System, Man and Cybernetics*, 25, 512-514.
28. Jacobs, R. (1988). Increased rates of convergence through learning rate adaptation, *Neural Networks*, 1, 295-307.
29. Johansson, E., Dowla, F., Goodman, D. (1990). Back-propagation learning for multi-layer feed-forward neural networks using the conjugate gradient method, *Technical report UCRL-JC-104850*, Lawrence Livermore National Laboratory, Livermore, CA.
30. Kollias, S. & Anastassiou, S. (1989). An adaptive least squares algorithm for the efficient training of artificial neural networks, *IEEE Trans. Circuits and Systems*, 36, 1092-1101.
31. Kramer, A. & Sangiovanni-Vincentelli, A. (1989). Efficient parallel learning algorithms for neural networks, in *Advances in Neural Information Processing Systems 1* (Touretzky, D., ed.), Morgan-Kaufmann, CA, 40-48.
32. Kung, S. (1993). *Digital neural networks*, PTR Prentice-Hall, Englewood Cliffs, NJ.
33. Becker, S. & Le Cun, Y. (1988). Improving the convergence of the back-propagation learning with second order methods, in *Proceedings of the 1988 Connectionist Models Summer School* (Touretzky, D., Hinton, G. & Sejnowski, T., eds.), Morgan-Kaufman, San Mateo, CA, 29-37.
34. Le Cun, Y., Denker, J. & Solla, S. (1990). Optimal brain damage, in *Advances in Neural Information Processing Systems 2* (Touretzky, D., ed.), Morgan-Kaufmann, CA, 598-605.
35. Le Cun, Y., Kanter, I. & Solla, S. (1991). Second order properties of error surfaces: learning time and generalization, in *Advances in Neural Information Processing Systems 3* (Lippmann, R., Moody, J. & Touretzky, D., ed.), Morgan-Kaufmann, CA, 918-924.
36. Le Cun, Y., Simard, P. & Pearlmutter, B. (1993). Automatic learning rate maximization by on-line estimation of the Hessian's eigenvectors, in *Advances in Neural Information Processing Systems 5* (Hanson, S., Cowan, J. & Giles, C., eds.), Morgan Kaufmann, San Mateo, CA, 156-163.
37. Lisboa, P. & Perantonis, S. (1991). Complete solution of the local minima in the XOR problem, *Network*, 2, 119-124.
38. Liu, D. & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization, *Math. Program.*, 45, 503-528.
39. Luenberger, D. (1969). *Optimization by vector space methods*, Wiley, NY.
40. MacKay, D. (1991). A practical Bayesian framework for back-prop networks, *Neural Computation*, 4, 448-472.
41. Magoulas, G.D, Vrahatis, M.N. & Androulakis, G.S. (1996). A new method in neural network supervised training with imprecision, in *Proceedings of the 3rd IEEE Inter. Conf. on Electr. Circuits and Syst.*, Rodos, Greece, 287-290.

42. Magoulas, G.D., Vrahatis, M.N., Grapsa, T.N. & Androulakis, G.S. (1997a). Neural network supervised training based on a dimension reducing method, in *Mathematics of Neural Networks, Models, Algorithms and Applications*, S.W. Ellacott, J.C. Mason & I.J. Anderson eds., Kluwer Academic Publishers, Boston, 1997, Chapter 41, 245-249.
43. Magoulas, G.D., Vrahatis, M.N. & Androulakis, G.S. (1997b). Effective back-propagation training with variable stepsize, *Neural Networks*, 10, 69-82.
44. Magoulas, G.D., Vrahatis, M.N. & Androulakis, G.S. (1999). Increasing the convergence rate of the error backpropagation algorithm by learning rate adaptation methods, *Neural Computation*, 11, No. 7, 1769-1796.
45. Magoulas, G.D., Karkanis, S.A., Karras, D.A. & Vrahatis, M.N. (2000a). Comparison study of textural descriptors for training neural network classifiers, *Inter. J. Computer Research*, accepted for publication.
46. Magoulas, G.D., Plagianakos, V.P., Androulakis, G.S. & Vrahatis, M.N. (2000b). A framework for the development of globally convergent adaptive learning rate algorithms, *Inter. J. Computer Research*, accepted for publication.
47. Magoulas, G.D. & Vrahatis, M.N. (2000). A class of adaptive learning rate algorithms derived by one-dimensional subminimization methods, *Neural, Parallel and Scientific Computations*, accepted for publication.
48. Møller, M. (1990). A scaled-conjugate gradient algorithm for fast supervised learning, *Technical report PB-339*, Computer Science Dept., University of Aarhus, Aarhus, Denmark.
49. Møller, M. (1993). A scaled conjugate gradient algorithm for fast supervised learning, *Neural Networks*, 6, 525-533.
50. Moody, J. (1992). The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems, in *Advances in Neural Information Processing Systems 4* (Moody, J., Hanson, S. & Lippmann, R., eds.), Morgan-Kaufmann, CA, 847-854.
51. Moré, J. (1983). Recent developments in algorithms and software for trust region methods, in *Mathematics Programming, The state of the art* (Bachem, A. Grottschel, M. & Korte, G., eds.), Springer-Verlag, Berlin, 256-287.
52. Ortega, J. & Rheinboldt, W. (1970). *Iterative solution of nonlinear equations in several variables*, Academic Press, NY.
53. Pearlmutter, B. (1992). Gradient descent: second order momentum and saturating error, in *Advances in Neural Information Processing Systems 4* (Moody, J., Hanson, S. & Lippmann, R., eds.), Morgan-Kaufmann, CA, 887-894.
54. Pearlmutter, B. (1994). Fast exact multiplication by the hessian, *Neural Computation*, 6, 147-160.
55. Pineda, F. (1987). Generalization of back-propagation to recurrent neural networks, *Physical Review Letters*, 59, 2229-2232.
56. Plagianakos, V.P., Magoulas, G.D. & Vrahatis, M.N. (1999a). Nonmonotone learning rules for backpropagation networks, in *Proceedings of the Sixth IEEE International Conference on Electronics, Circuits and Systems*, Pafos, Cyprus, 291-294.
57. Plagianakos, V.P., Vrahatis, M.N. & Magoulas, G.D. (1999b). Nonmonotone methods for backpropagation training with adaptive learning rate, in *Proceedings of the 1999 International Joint Conference on Neural Networks, (IJCNN'99)*, Washington DC, U.S.A., #2001 Session: 5.1.
58. Polak, E., (1997). *Optimization: algorithms and consistent approximations*, Springer, New York.
59. Powell, M. (1975). Convergence properties of a class of minimization algorithms, in *Nonlinear Programming 2* (Mangasarian, O., Meyer, R. & Robinson, S., eds.), Academic Press, NY, 1-27.

60. Riegler, A., Irvine, J. & Vogl, T. (1991). Rescaling of variables in back-propagation learning, *Neural Networks*, 4, 225-229.
61. Rojas, R. (1993). A graphical proof of the backpropagation learning algorithm, in *Parallel Computing Technologies* (Malyshekin, V., ed.), Obninsk, Russia.
62. Rumelhart, D., Hinton, G. & Williams, R. (1986). Learning internal representations by error back-propagation, in *Parallel distributed processing: explorations in the microstructure of cognition* (Rumelhart, D. & McClelland, J., eds.), MIT Press, Cambridge, MA.
63. Saارينen, S., Bramley, R. & Cybenko, G. (1992). Neural networks, back-propagation and automatic differentiation, in *Automatic differentiation of algorithms: theory, implementation and application* (Griewank, A. & Gorliss, G., eds.), SIAM, Philadelphia, PA, 31-42.
64. Schultz, G., Schnabel, R. & Byrd, R. (1985). A family of trust region based algorithms for unconstrained minimization with strong global convergence properties, *SIAM J. Numer. Anal.*, 22, 47-67.
65. Shultz, G., Schnabel, R. & Byrd, R. (1982). A family of trust region based algorithms for unconstrained minimization with strong global convergence properties, *Technical report CU-CS-216-82*, Computer Science Dept., University of Colorado.
66. Sorensen, D. (1982). Newton's method with a model trust region modification, *SIAM J. Numer. Anal.*, 19, 409-426.
67. Swanston, D., Bishop, J. & Mitchell, R. (1994). Simple adaptive momentum: new algorithm for training multilayer perceptrons, *Electronics Letters*, 30, 1498-1500.
68. Van der Smagt, P. (1994). Minimization methods for training neural networks, *Neural Networks*, 7, 1-11.
69. Van Ooyen, A. & Nienhuis, B. (1992). Improving the convergence of the backpropagation algorithm, *Neural Networks*, 5, 465-471.
70. Vrahatis, M.N., Androulakis, G.S., Lambrinos, J.N. & Magoulas, G.D. (2000a). A class of gradient unconstrained minimization algorithms with adaptive stepsize, *J. Comput. Appl. Math.*, 114, No. 2, 367-386.
71. Vrahatis, M.N., Magoulas, G.D. & Plagianakos, V.P. (2000b). Globally convergent modification of the Quickprop method, *Neural Processing Letters*, to appear vol. 12, No. 2, October 2000.
72. Watrous, R. (1987). Learning algorithms for connectionist networks: applied gradient methods of nonlinear optimization, in *Proceedings of the 1st IEEE Inter. Conf. on Neural Networks*, Vol. 2, San Diego, CA, 619-627.
73. Werbos, P. (1974). Beyond regression: new tools for prediction and analysis in the behavioral sciences, Ph.D. Thesis, Harvard University, Cambridge, MA.
74. Werbos, P. (1992). Neural networks and the human mind: new mathematics fits humanistic insight, *IEEE Inter. Conf. on Systems, Man and Cybernetics*, Vol. 1, Chicago, IL, 78-83.
75. Wilkinson, J. (1963). *Rounding errors in algebraic processes*, Prentice-Hall, Englewood Cliffs, NJ.
76. Wolfe, P. (1969). Convergence conditions for ascent methods, *SIAM Review*, 11, 226-235.
77. Wolfe, P. (1971). Convergence conditions for ascent methods. II: Some corrections, *SIAM Review*, 13, 185-188.
78. Wray, J. & Green, G. (1995). Neural Networks, approximation theory and finite precision computation, *Neural Networks*, 8, 31-37.
79. Zoutendijk, G. (1970). Nonlinear programming, computational methods, in *Integer and Nonlinear Programming* (Abadie, J., ed.), North-Holland, Amsterdam, 37-86.