Εισαγωγή στα Τεχνητά Νευρωνικά Δίκτυα

Μ.Ν. Βραχάτης[1], Β.Π. Πλαγιανάκος[1], και Γ.Δ. Μαγουλάς[2]

[1] Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών, Πάτρα 26500, Ελλάδα.
[2] Department of Information Systems and Computing, Brunel University, Uxbridge UB8 3PH, U.K.
e-mail: vrahatis@math.upatras.gr, vpp@math.upatras.gr, George.Magoulas@brunel.ac.uk

1 Εισαγωγή

Ένα *Τεχνητό Νευρωνικό Δίκτυο* (ΤΝΔ) είναι 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). Αλλιώς, ένα ΤΝΔ είναι μια αρχιτεκτονική που εκτελεί αριθμητικούς υπολογισμούς χρησιμοποιώντας δομή μαζικού παραλληλισμού (massively parallel structure) ή παράλληλης κατανεμημένης εργασίας (parallel distributed processing) και έχει την έμφυτη ιδιότητα να αποθηκεύει εμπειρική γνώση και να την έχει διαθέσιμη για χρήση στο μέλλον. Τα ΤΝΔ παρέχουν ένα εναλλακτικό αλγοριθμικό μοντέλο σύμφωνα με το οποίο οι υπολογισμοί γίνονται παράλληλα και μαζικά και η εκπαίδευση αντικαθιστά την ανάπτυξη προγράμματος. 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* [21].

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_ι 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 κοιτάξουμε 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 [11].

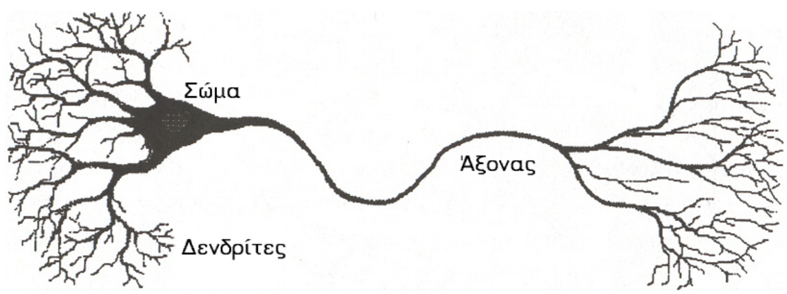
Και πάλι ο ανθρώπινος εγκέφαλος υπερισχύει. Υπάρχουν δύο λόγοι για αυτό: η δυνατότητα *μάθησης* και η *παράλληλη επεξεργασία* των δεδομένων. Η ικανότητα του ανθρώπινου εγκεφάλου να φτιάχνει από πολύ μικρή ηλικία και διαρκώς να διορθώνει κανόνες (αυτό που συνήθως ονομάζουμε εμπειρία), καθώς και η δυνατότητά 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 Νευροφυσιολογίας 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.

2 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 [64]. Οι νευρώνες αποτελούνται από 3 βασικά τμήματα, όπως φαίνεται και στο Σχήμα 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.



Σχήμα 1: Απλουστευμένο μοντέλο ενός βιολογικού νευρώνα

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 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.

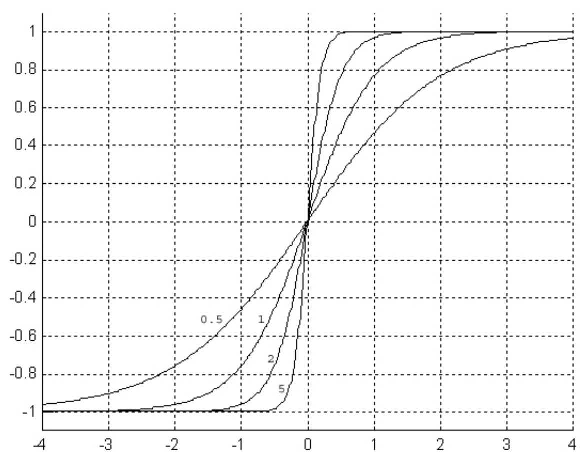
Η εμπειρία και η γνώση που αποκτά το ΤΝΔ αποθηκεύεται στα βάρη του. Η προσαρμογή των τιμών των βαρών έτσι ώστε το ΤΝΔ να έχει την επιθυμητή απόκριση ονομάζεται *εκπαίδευση* και γίνεται με τη βοήθεια κάποιου *αλγόριθμου εκπαίδευσης*.

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:
  $f_1(x) = \begin{cases} 1, & x \geqslant \theta \\ 0, & x < \theta \end{cases}$
- Η διπολική συνάρτηση με κατώφλι θ:
  $f_2(x) = \begin{cases} 1, & x \geqslant \theta \\ -1, & x < \theta \end{cases}$
- 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 $\lambda_2 = 1$. Στο Σχήμα 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 (δυαδικά, διπολικά, πραγματικοί αριθμοί κτλ.) και είναι κρίσιμη για την εκπαίδευση. Στα Σχήματα 3 και 4 απεικονίζονται ένας βιολογικός και ένας τεχνητός νευρώνας και δίνεται ο τύπος υπολογισμού της ενεργοποίησης του τεχνητού νευρώνα.

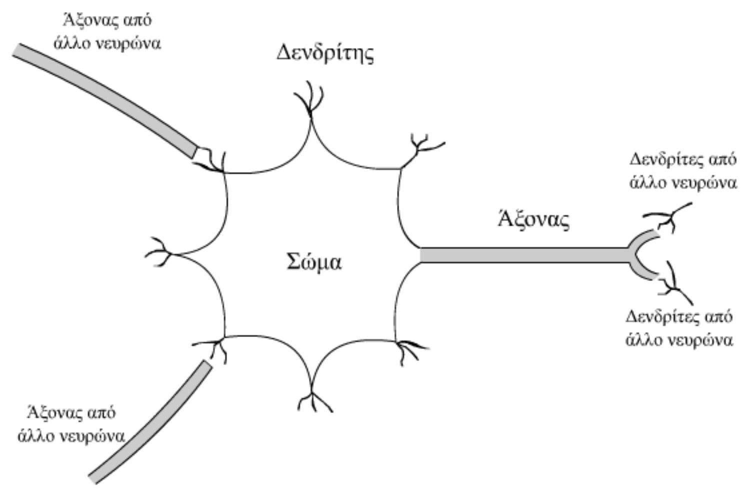


Σχήμα 2: Η υπερβολική εφαπτομένη για διάφορες τιμές της παραμέτρου λ

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 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



Σχήμα 3: Μοντέλο ενός βιολογικού νευρώνα

κάποιου νευρώνα του ιδίου ή προηγούμενου στρώματος, τότε θα λέμε ότι το ΤΝΔ είναι πρόσθιας τροφοδότησης (feed-forward). 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 στρωμάτων, 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 μελλοντικές τιμές θέλουμε να προβλέψουμε [55].

Στο υπόλοιπο της παρούσας Διατριβής θα ασχοληθούμε μόνο με πλήρως διασυνδεδεμένα πρόσθιας τροφοδότησης ΤΝΔ, όπως αυτό που φαίνεται στο Σχήμα 5. Συμπερασματικά, ένα ΤΝΔ χαρακτηρίζεται από: (α) την αρχιτεκτονική του, (β) τις συναρτήσεις ενεργοποίησης που χρησιμοποιεί, και (γ) τον αλγόριθμο εκπαίδευσης.

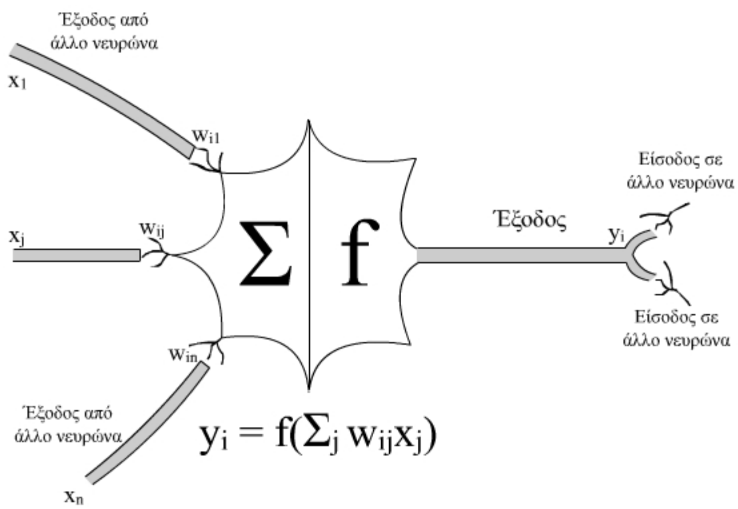
3 Ιστορική Αναδρομή

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 [4, 20].

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 τις 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 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 [44], 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 [22] 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 [58]. 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 [76] παρουσίασαν ένα αλγόριθμο εκπαίδευσης γραμμικών ΤΝΔ περίπου των ιδίων δυνατοτήτων και δομής με το perceptron. Ο αλγόριθμός τους χρησιμοποιείται



Σχήμα 4: Μοντέλο ενός τεχνητού νευρώνα

ακόμα και σήμερα.

Δυστυχώς, αυτά τα ΤΝΔ είχαν κάποιους σημαντικούς περιορισμούς, όπως έδειξαν οι Martin Minsky και Seymour Papert [45] το 1969. Τα προβλήματα αυτά μπορούσαν να λυθούν 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, που για πολλά χρόνια 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 [60, 61]. 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 εκατοντάδες αλγόριθμοι εκπαίδευσης και αρχιτεκτονικές ΤΝΔ και έχουν γραφτεί χιλιάδες 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.

4 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.

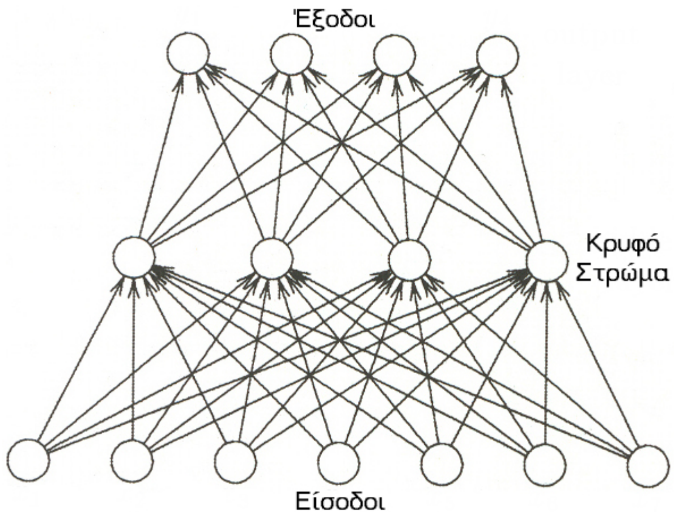
Αεροπλοΐα. Δημιουργία αυτόματων πιλότων και προγραμμάτων προσομοίωσης πτήσης, συστήματα ελέγχου πτήσης, ανίχνευση ελαττωμάτων σε τμήματα των αεροπλάνων.

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

Γεωλογία. Ανάλυση πιθανότητας ύπαρξης πετρελαίου σε γεωλογικούς σχηματισμούς, ανάλυση πετρωμάτων σε ορυχεία.

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

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



Σχήμα 5: Ένα πολυστρωματικό πλήρως διασυνδεδεμένο πρόσθιας τροφοδότησης ΤΝΔ

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.

Στρατός. 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.

5 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 [21].

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 έξοδός του να ταυτίζεται με την επιθυμητή έξοδο. 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. Παραδείγματος χάρη, 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.

6 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 [10]. Όμως, 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.

6.1 Η μορφολογία του χώρου των βαρών

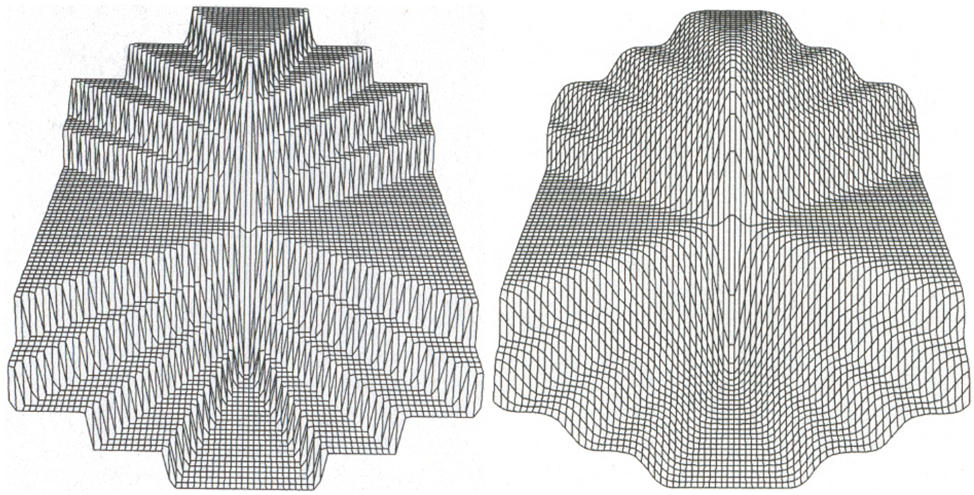
Η εκπαίδευση ΤΝΔ είναι ένα πολύ δύσκολο πρόβλημα κυρίως λόγω της μορφής του χώρου των βαρών. Διαισθητικά, ο χώρος των βαρών παρουσιάζει πολλαπλά τοπικά ελάχιστα διότι είναι η σύνθεση των μη γραμμικών συναρτήσεων ενεργοποίησης (που έχουν ελάχιστα σε διαφορετικά σημεία), με αποτέλεσμα πολλές φορές η τελική συνάρτηση να μην είναι κυρτή [18]. 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 [27, 57], όπως φαίνεται και στο Σχήμα 6.

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 7). Στην περίπτωση αυτή δημιουργούνται στενές περιοχές ανεπιθύμητων τοπικών ελαχίστων που μπορούν να παγιδεύσουν τις μεθόδους εκπαίδευσης [27], όπως φαίνεται και στο Σχήμα 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 9.

6.2 Η αρχικοποίηση των βαρών

Σε αυτή την Ενότητα θα δούμε 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



Σχήμα 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 $\lambda = 0.1$

τιμές αυτές μπορεί να είναι από την ομοιόμορφη κατανομή στο διάστημα $(-1, 1)$ ή κάποιο μικρότερο. Εναλλακτικά, μπορούν να χρησιμοποιηθούν κάποιες εμπειρικές τεχνικές για την επιλογή των αρχικών βαρών, έτσι ώστε να βοηθηθεί η διαδικασία της εκπαίδευσης [15, 62, 83].

Μια από τις πιο γνωστές τεχνικές είναι αυτή που πρότειναν οι Nguyen και Widrow [37, 47]. 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:

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

όπου $\mathcal{N}$ είναι ο αριθμός των νευρώνων εισόδου και $\mathcal{M}$ είναι ο αριθμός των κρυφών νευρώνων. Στη συνέχεια επιλέγονται τα βάρη $w^r = (w_{11}^r, \cdots, w_{nm}^r, \cdots, w_{\mathcal{N}\mathcal{M}}^r)$ 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:

$$w_{nm}^0 = \frac{\rho w_{nm}^r}{\|w^r\|}. \tag{2}$$

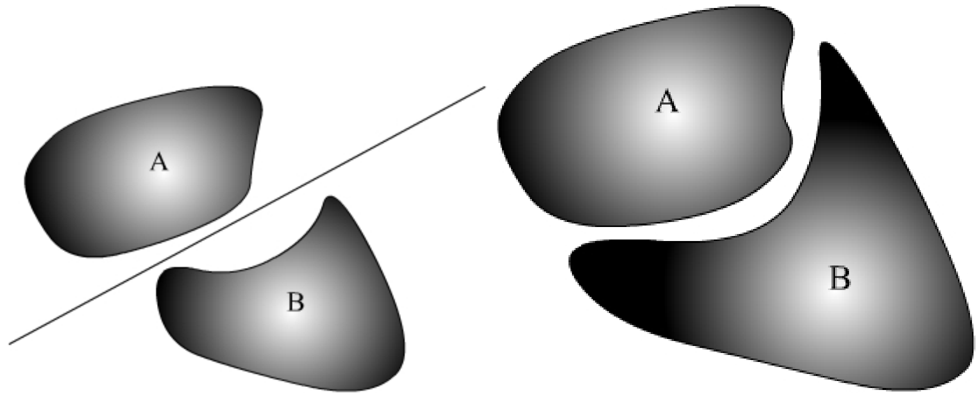
Αυτή η διαδικασία έχει σαν σκοπό να κατανείμει τα αρχικά βάρη έτσι ώστε να είναι πιο πιθανό κάθε πρότυπο εισόδου να προκαλεί αποδοτική εκπαίδευση των κρυφών νευρώνων, με αποτέλεσμα να επιταχύνεται η διαδικασία της εκπαίδευσης [47].

6.3 Κατηγορίες Μεθόδων Εκπαίδευσης

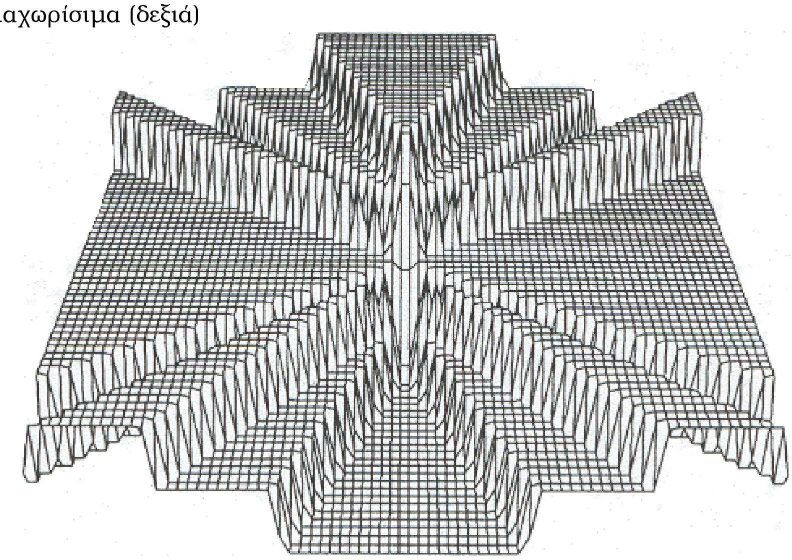
Οι μέθοδοι εκπαίδευσης ΤΝΔ μπορούν να χωριστούν σε δύο βασικές κατηγορίες: (α) μεθόδους εκπαίδευσης με επίβλεψη (supervised learning), και (β) μεθόδους εκπαίδευσης χωρίς επίβλεψη (unsupervised learning). Στην πρώτη περίπτωση είναι αναγκαία η παρουσία ενός «δασκάλου», ενώ στη δεύτερη το ΤΝΔ πρέπει να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

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



Σχήμα 7: Δύο σύνολα προτύπων εκπαίδευσης A και B. Τα σύνολα είναι γραμμικώς διαχωρίσιμα (αριστερά). Τα σύνολα δεν είναι γραμμικώς διαχωρίσιμα (δεξιά)

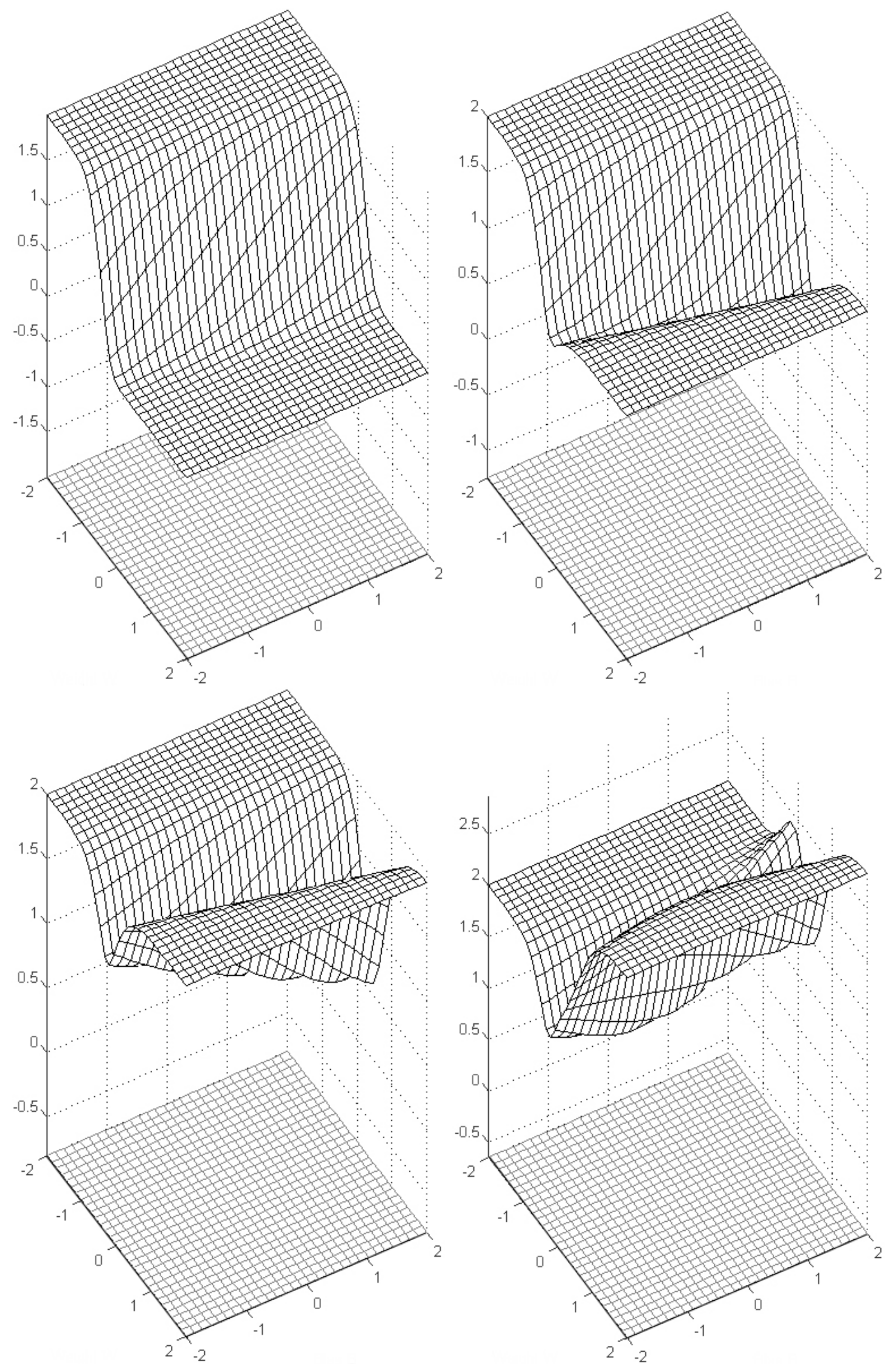


Σχήμα 8: Παράδειγμα γραφικής παράστασης του χώρου των βαρών ενός ΤΝΔ με ένα μόνο νευρώνα, όταν το σύνολο εκπαίδευσης δεν είναι γραμμικώς διαχωρίσιμο

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 [31, 32], James Anderson [3] και Stephen Grossberg [19].

Τέλος, ένας άλλος τρόπος εκπαίδευσης, που όμως χρησιμοποιείται σπανιότερα, είναι η εκπαίδευση με ενίσχυση (reinforcement learning) [7]. 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. Παραδείγματος χάριν, 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



Σχήμα 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 (επάνω αριστερά, επάνω δεξιά, κάτω αριστερά, κάτω δεξιά)

επόμενο Κεφάλαιο θα μελετήσουμε το θεωρητικό υπόβαθρο πάνω στο οποίο στηρίζονται αυτές οι μέθοδοι.

6.4 Παράλληλη εκπαίδευση ΤΝΔ

Η παράλληλη επεξεργασία, δηλαδή η μέθοδος όπου η παράλληλη επίλυση μικρών υποπροβλημάτων έχει σαν αποτέλεσμα την γρήγορη επίλυση ενός μεγάλου και δύσκολου προβλήματος, έχει αναπτυχθεί πολύ τα τελευταία χρόνια [35]. Τα ΤΝΔ σαν αρχιτεκτονικές που εκτελούν αριθμητικούς υπολογισμούς χρησιμοποιώντας δομή παράλληλης κατανεμημένης επεξεργασίας μπορούν να υλοποιηθούν και να εκτελεστούν από παράλληλες υπολογιστικές μηχανές.

Γενικά, η εκπαίδευση ενός ΤΝΔ με τη βοήθεια κάποιας μεθόδου της κλάσης της οπισθοδρομικής διάδοσης του σφάλματος [60] 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 [12]. 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.

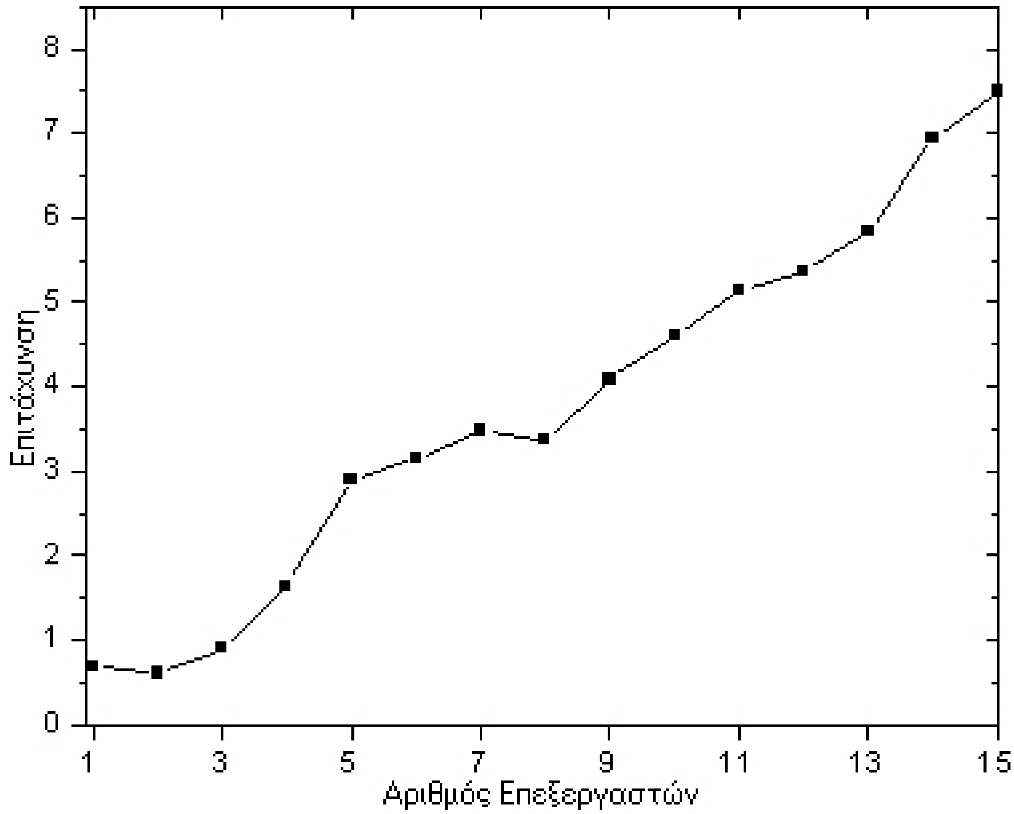
Η διαδικασία που θα περιγράψουμε βασίζεται στην υλοποίηση μιας Παράλληλης Εικονικής Μηχανής (ΠΕΜ) (Parallel Virtual Machine – PVM) [16], που αποτελείται από 15 προσωπικούς υπολογιστές και χρησιμοποιείται για την παράλληλη εκπαίδευση ΤΝΔ με πολύ μεγάλα σύνολα προτύπων εκπαίδευσης. Στην εργασία [51] δίνουμε 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, 68]. Αξίζει να σημειωθεί ότι αφού υλοποιηθεί η ΠΕΜ μπορεί να χρησιμοποιηθεί και για την επίλυση άλλων υπολογιστικά δαπανηρών προβλημάτων (βλ. την εργασία [54] 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 [16]. 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 [16]. 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 [51] για μια αναλυτική παρουσίαση της μεθόδου σε ψευδοκώδικα).

Στο Σχήμα 10 παρουσιάζουμε την επιτάχυνση (speedup) της παράλληλης διαδικασίας εκπαίδευσης σε σχέση με τη σειριακή εκπαίδευση ανάλογα με το πλήθος των επεξεργαστών που χρησιμοποιούμε.



Σχήμα 10: Επιτάχυνση του χρόνου εκπαίδευσης ανάλογα με το πλήθος των επεξεργαστών

7 Θεωρητικό Υπόβαθρο των Μεθόδων Εκπαίδευσης ΤΝΔ

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 $w^* = (w_1^*, w_2^*, \ldots, w_n^*) \in \mathbb{R}^n$, 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:

$$E(w) = \frac{1}{2}\sum_{p=1}^{P}\sum_{j=1}^{N_L}\left(y_{j,p}^L - t_{j,p}\right)^2, \tag{3}$$

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 ελάχιστων 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 [29, 30, 59] 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 [8] ο Battiti παρουσιάζεται μια επισκόπηση τεχνικών ελαχιστοποίησης με χρήση παραγώγων πρώτης και δεύτερης τάξης που εφαρμόζονται για εκπαίδευση των ΤΝΔ με επίβλεψη.

Μια μεγάλη ποικιλία από μεθόδους που είναι δανεισμένες από την Αριθμητική Ανάλυση έχουν χρησιμοποιηθεί για την επιτάχυνση της εκπαίδευσης, χρησιμοποιώντας πληροφορίες από τις παραγώγους δεύτερης τάξης [8, 43, 46, 50, 66, 75]. Όμως, οι μέθοδοι δεύτερης τάξης είναι πολλές φορές υπολογιστικά ακριβές, όταν το ΤΝΔ έχει μερικές χιλιάδες βάρη [9]. Επιπρόσθετα, δεν είναι βέβαιο ότι το επιπλέον υπολογιστικό κόστος θα επιταχύνει την διαδικασία εκπαίδευσης, ειδικά σε μη κυρτές συναρτήσεις μακριά από το ελάχιστο [13, 48], όπως είναι η συναρτήσεις που δημιουργούνται κατά την εκπαίδευση ΤΝΔ [8]. 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) [60].

Η μέθοδος της οπισθοδρομικής διάδοσης του σφάλματος χαρακτηρίζεται από καλή απόδοση όταν οι αρχικές τιμές του διανύσματος βαρών είναι μακριά από το ελάχιστο, κάτι που ισχύει στις περισσότερες περιπτώσεις εκπαίδευσης των ΤΝΔ. Ωστόσο, η σύγκλισή 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.

Τέλος, στην περίπτωση εκπαίδευσης ανά πρότυπο εισόδου p [40, 52, 53] 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 [63].

8 Η Επιλογή του Ρυθμού Εκπαίδευσης

Κάνοντας μια μικρή ιστορική αναδρομή στο πρόβλημα της επιλογής του ρυθμού εκπαίδευσης πρέπει να αναφερθεί ότι πρώτος ο Goldstein [17] 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 σύγκλισή της υπό λιγότερο αυστηρές προϋποθέσεις [5]. Μια βελτίωση της μεθόδου αυτής για την εκπαίδευση νευρωνικών δικτύων προτάθηκε στην εργασία [42].

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 «στροφές» και ο «πυθμένας» είναι ελαφρά κεκλιμένος [28, 60]. Επίσης, αυτή η αντιμετώπιση του προβλήματος εισάγει δυσκολίες στην προσπάθεια δημιουργίας μεθόδων ευρείας σύγκλισης [34, 36]. 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 [5, 42, 73].

9 Αλγόριθμοι Εκπαίδευσης με Ευρεία Σύγκλιση

Όπως αναφέρουν οι Dennis και Schnabel [14, σελ. 5], ο όρος «αλγόριθμος ευρείας σύγκλισης»[1] χρησιμοποιείται για να δηλώσει ότι μια μέθοδο είναι κατασκευασμένη έτσι ώστε να συγκλίνει σε ένα *τοπικό* ελάχιστο μιας μη γραμμικής συνάρτησης, *από σχεδόν οποιοδήποτε αρχικό σημείο*. Επιπρόσθετα, ο Nocedal [48, 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 διαδικασία εκπαίδευσης ΤΝΔ καθώς τις περισσότερες φορές η εκπαίδευσης 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 οποιοδήποτε αλγόριθμο που επαναληπτικά προσαρμόζει τα βάρη ακολουθώντας κατευθύνσεις μείωσης της συνάρτησης σφάλματος εξασφαλίζει στον αλγόριθμο την ιδιότητα της ευρείας σύγκλισης [38, 39, 73].

Οι ιδιότητες σύγκλισης των μεθόδων ευθύγραμμης (μονοδιάστατης) ανίχνευσης μπορούν να μελετηθούν μετρώντας την ποιότητα της κατεύθυνσης 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:

$$\cos\theta_k = \frac{-\nabla E(w^k)^\top d^k}{\|\nabla E(w^k)\|\,\|d^k\|}, \tag{4}$$

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

$$E(w^k + \alpha^k d^k) - E(w^k) \leqslant \sigma_1 \eta^k \nabla E(w^k)^\top d^k, \tag{5}$$

$$\nabla E(w^k + \eta^k d^k)^\top d^k \geqslant \sigma_2 \nabla E(w^k)^\top d^k, \tag{6}$$

όπου $0 < \sigma_1 < \sigma_2 < 1$. Οι δύο ανισότητες είναι γνωστές ως συνθήκες του Wolfe [78, 79]. 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 της κλίσης να είναι μικρότερη από $90°$ [78, 79].

Το θεώρημα των Wolfe και Zoutendijk [78, 79, 82] 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)

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

ανίχνευση που ικανοποιεί τις συνθήκες του Wolfe για τον καθορισμό του ρυθμού εκπαίδευσης. Στην εργασία [41] παρουσιάζουμε αυτό το θεώρημα και με τη βοήθειά του αποδεικνύουμε ευρεία σύγκλιση σε μια νέα κλάση μεθόδων εκπαίδευσης ΤΝΔ.

10 Βελτιστοποίηση μη Γραμμικών Συναρτήσεων ανά Κατεύθυνση

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

$$\nabla f(x^*) = \Theta^n = (0, 0, \ldots, 0). \tag{7}$$

Η Εξίσωση (7) 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:

$$\begin{gathered} \partial_1 f(x_1, x_2, \ldots, x_n) = 0, \\ \partial_2 f(x_1, x_2, \ldots, x_n) = 0, \\ \vdots \\ \partial_n f(x_1, x_2, \ldots, x_n) = 0, \end{gathered} \tag{8}$$

όπου $\partial_i f(x_1, \ldots, x_i, \ldots, x_n)$ είναι οι μερικές παράγωγοι 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 (8). Επίσης θα μελετήσουμε την μέθοδο του Powell και μια τροποποίησή της.

10.1 Μελέτη σύγκλισης της σύνθετης μη γραμμικής μεθόδου Jacobi

Η κλάση των μη γραμμικών μεθόδων *Jacobi* χρησιμοποιείται συχνά για την επίλυση του Συστήματος (8). Το κύριο χαρακτηριστικό τους είναι ότι είναι παράλληλοι αλγόριθμοι που μπορούν να υλοποιηθούν σε παράλληλους υπολογιστές [49]. Ξεκινώντας από ένα τυχαίο αρχικό διάνυσμα $x^0 \in \mathcal{D}$, στην k επανάληψη εκτελούμε μονοδιάστατη ελαχιστοποίηση της συνάρτησης:

$$f(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_n^k), \tag{9}$$

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

$$\partial_i f(x_1^k, \ldots, x_{i-1}^k, \hat{x}_i, x_{i+1}^k, \ldots, x_n^k) = 0. \tag{10}$$

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:

$$x_i^{k+1} = x_i^k + \tau_k(\hat{x}_i - x_i^k), \tag{11}$$

για κάποια παράμετρο χαλάρωσης τ_k. Συνεπώς, η αντικειμενική συνάρτηση (9) ελαχιστοποιείται μονοδιάστατα προς κάθε κατεύθυνση i.

Ανάλογα την χρησιμοποιούμενη μέθοδο μονοδιάστατης ελαχιστοποίησης, διάφορες σύνθετες μη γραμμικές μέθοδοι Jacobi μπορούν να κατασκευαστούν. Αξίζει να σημειώσουμε ότι ο αριθμός των επαναλήψεων της μονοδιάστατης ελαχιστοποίησης εξαρτάται από την ζητούμενη ακρίβεια. Έτσι ο 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 [49, 71].

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 1 *Έστω $f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}$ δύο φορές συνεχώς παραγωγίσιμη στην ανοικτή περιοχή $\mathcal{S}_0 \subset \mathcal{D}$ του σημείου $x^* \in \mathcal{D}$ για το οποίο $\nabla f(x^*) = \Theta^n$ και η Εσσιανή, $H(x^*)$ 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 $\{x^k\}_{k=0}^{\infty}$ 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 Εσσιανή είναι θετικά ορισμένη στο x^*. Η εύρεση του σημείου αυτού είναι ισοδύναμη με την επαναληπτική επίλυση του Συστήματος (8) με την εφαρμογή της μη γραμμικής μεθόδου Jacobi και μιας οποιασδήποτε μεθόδου μονοδιάστατης ελαχιστοποίησης.

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

$$H(x^*) = D(x^*) - L(x^*) - L^\top(x^*). \tag{12}$$

Αφού, $H(x^*)$ είναι συμμετρική και θετικά ορισμένη, τότε $D(x^*)$ είναι θετικά ορισμένος [70]. Επίσης, αφού η $H(x^*)$ έχει την ιδιότητα A^π, οι ιδιοτιμές του

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

είναι πραγματικές και $\rho(\Phi(x^*)) < 1$ [6] (όπου $\rho(A)$ είναι η φασματική ακτίνα του πίνακα A). Τότε 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 $\lim_{k\to\infty} x^k = x^*$ [49]. Το Θεώρημα αποδείχθηκε. □

Παρατήρηση 1 *Ο πίνακας A έχει την ιδιότητα A^π [6] εάν ο πίνακας $PAP^\top$, μπορεί να γραφτεί σε 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, $i = 1, \ldots, r$ είναι μη ιδιάζοντες. Μια αλγοριθμική διαδικασία για το μετασχηματισμό ενός συμμετρικού πίνακα στην παραπάνω τριδιαγώνια μορφή, βρίσκεται στο βιβλίο του Stewart [69, σελ. 335].

10.2 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:

$$f(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \ldots, x_n^k), \tag{13}$$

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 (11). 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 *Έστω $f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}$ 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 Εσσιανή $H(x^*)$ είναι θετικά ορισμένη. Η εύρεση ενός τέτοιου σημείου ισοδυναμεί με την αντίστοιχη λύση $x^* \in \mathcal{D}$ του Συστήματος (8) 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 (12). Τώρα, αφού η $H(x^*)$ είναι συμμετρική και θετικά ορισμένη, τότε $D(x^*)$ είναι μη ιδιάζων [70]. Από το Θεώρημα του Ostrowski [70], έχουμε ότι $\rho\big(\Phi_\tau(x^*)\big) < 1$ για κάθε $\tau \in (0, 2)$ και συνεπώς, από το Θεώρημα της μεθόδου SOR [49], 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 $\lim_{k\to\infty} x^k = x^*$. Το Θεώρημα αποδείχθηκε. □

10.3 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 [56] 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, $i = 1, 2, \ldots, n$ είναι το αρχικό σύνολο των κατευθύνσεων ανίχνευσης, που είναι το σύνολο των διανυσμάτων της βάσης του χώρου e_i, $i = 1, 2, \ldots, n$:

1. Για $i = 1, 2, \ldots, n$ 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 $i = 1, 2, \ldots, n-1$, 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 γραμμικώς εξαρτημένες. Υπάρχουν πολλές διαδικασίες για την εξάλειψη 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:

$$f(x^0 + \lambda u) - f(x^0) = 0. \tag{14}$$

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:

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

Για την επίλυση της μονοδιάστατης εξίσωσης (14), χρησιμοποιούμε μια τροποποίηση της μεθόδου της διχοτόμησης (bisection) [72, 74]. Η μέθοδος αυτή δεν απαιτεί τις ακριβείς συναρτησιακές τιμές, αλλά μόνο τις σχετικές τιμές δύο σημείων, δηλαδή απλά συγκρίνει τις συναρτησιακές τιμές. Άλλα πλεονεκτήματα της μεθόδου είναι ότι μπορεί να υλοποιηθεί παράλληλα, συγκλίνει πάντα μέσα στο δοθέν διάστημα και είναι βέλτιστη, δηλαδή έχει ασυμπτωτικά τον καλύτερο ρυθμό σύγκλισης [65].

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 3 *[56, 81] 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 η λιγότερες επαναλήψεις, ανεξάρτητα από το αρχικό σημείο και την διαδοχή των κατευθύνσεων.*

Ορισμός 1 *Μια μέθοδος που ελαχιστοποιεί την f σύμφωνα με τις απαιτήσεις του Θεωρήματος 3, έχει την ιδιότητα του τετραγωνικού τερματισμού (quadratic termination).*

Θεώρημα 4 *[56, 81] Οι κατευθύνσεις που χρησιμοποιούνται από τη μέθοδο του Powell είναι συζυγείς.*

Θεώρημα 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 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 3 ικανοποιούνται. Το θεώρημα αποδείχθηκε. □

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.

11 Πρακτική Θεώρηση της Σύγκλισης των Αλγορίθμων Εκπαίδευσης

Τα θεωρητικά αποτελέσματα είναι χρήσιμα για την κατανόηση και τη μελέτη της συμπεριφοράς των αλγόριθμων 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 αλγόριθμων τοπικής σύγκλισης.

- *Οι μη ακριβείς τιμές της συνάρτησης σφάλματος.* Είναι γνωστό πως οι αριθμητικοί υπολογισμοί υπόκεινται σε σφάλματα ακρίβειας [77]. Οι αριθμητικές πράξεις που απαιτούνται στις προσομοιώσεις των αλγόριθμων εκπαίδευσης επηρεάζουν την ακρίβεια των τιμών της συνάρτησης σφάλματος [80]. 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 [24].

- *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 εκκινεί από ακατάλληλα αρχικά βάρη και εμποδίζει το ΤΝΔ από το να μάθει πλήρως όλα τα πρότυπα.

12 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 6 και 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 8, διατυπώνει τα παραπάνω θεωρητικά αποτελέσματα στην ορολογία των ΤΝΔ.

12.1 Θεωρήματα των Kolmogorov και Sprecher

Θεώρημα 6 *[33] Κάθε συνεχής συνάρτηση $f(x_1, \ldots, x_n)$ ορισμένη στο $I^n, n \geqslant 2$, 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 7 *[67] Για κάθε ακέραιο $n \geqslant 2$ υπάρχει μια πραγματική μονοτόνως αύξουσα συνάρτηση $\phi(x)$, $\phi : [0, 1] \to [0, 1]$, που εξαρτάται από το n και έχει την ακόλουθη ιδιότητα: για κάθε δεδομένο $\delta > 0$, υπάρχει ένας ρητός ϵ, $0 < \epsilon < \delta$, τέτοιος ώστε κάθε πραγματική συνεχής συνάρτηση n μεταβλητών, $f(x)$, που ορίζεται στο I^n, να μπορεί να γραφτεί:*

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

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

Θεώρημα 8 *[23] Έστω μια συνεχής συνάρτηση $f : I^n \to \mathbb{R}^m$, όπου $I = [0, 1]$, τότε η f μπορεί να προσεγγιστεί ακριβώς από ένα ΤΝΔ πρόσθιας 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:

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

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 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 [25] έδειξαν ότι μπορούμε να χρησιμοποιήσουμε οποιαδήποτε αύξουσα συνάρτηση ενεργοποίησης $h(x)$, όπου $0 \leqslant h(x) \leqslant 1$, για κάθε x και επίσης:

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

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 [26].

Αναφορές

[1] "The beowulf project", http://www.beowulf.org, last accessed 01/03/2002.

[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] J. Anderson, "A simple neural network generating an interactive memory", *Mathematical Biosciences*, vol. 14, 197–220, 1972.

[4] J.A. Anderson and E. Rosenfeld, *Neurocomputing: Foundations of research*, MIT Press, Cambridge, 1989.

[5] L. Armijo, "Minimization of functions having lipschitz–continuous first partial derivatives", *Pacific Journal of Mathematics*, vol. 16, 1–3, 1966.

[6] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, New York, 1996.

[7] 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.

[8] R. Battiti, “First- and second-order methods for learning: between steepest descent and newton’s method”, *Neural Computation*, vol. 4, 141–166, 1992.

[9] 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 Koufmann, San Mateo, 29–37, 1988.

[10] A.L. Blum and R. Rivest, “Training a 3–node neural network is np–complete”, *Neural Networks*, vol. 5, 117–127, 1992.

[11] P.S. Churchland, *Neurophilosophy: Toward a unified science of the mind/brain*, MIT Press, Cambridge, 1986.

[12] 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.

[13] J.E. Dennis, Jr. and J.J. Moré, “Quasi–newton methods, motivation and theory”, *SIAM Review*, vol. 19, 46–89, 1977.

[14] 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.

[15] 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.

[16] 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.

[17] A.A. Goldstein, “On steepest descent”, *SIAM Journal of Control*, vol. 3, 147–151, 1965.

[18] 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.

[19] S. Grossberg, “Some physiological and biochemical consequences of psychological postulates”, In the proceedings of the *National Academy of Sciences*, vol. 60, 758–765, 1968.

[20] M.T. Hagan, H.B. Demuth, and M. Beale, *Neural network design*, PWS Publishing Company, Boston, 1996.

[21] S. Haykin, *Neural networks: a comprehensive foundation*, Macmillan College Publishing Company, New York, 1994.

[22] D.O. Hebb, *The organization of behavior*, Wiley, New York, 1949.

[23] 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.

[24] J. Holt and J. Hwang, “Finite precision error analysis of neural network hardware implementations”, *IEEE Transactions on Computers*, vol. 42, 281–290, 1993.

[25] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators”, *Neural Networks*, vol. 2, 359–366, 1989.

[26] 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.

[27] 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.

[28] R.A. Jacobs, “Increased rates of convergence through learning rate adaptation”, *Neural Networks*, vol. 1, 295–307, 1988.

[29] 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.

[30] N.A. Katsanos, F. Roubani-Kalantzopoulou, E. Iliopoulou, V. Siokos I. Bassiotis, M.N. Vrahatis, and V.P. Plagianakos, "Lateral molecular interactions on heterogeneous surfaces experimentally measured", *Colloid Surfaces A*, 2001, in press.

[31] T. Kohonen, "Correlation matrix memories", *IEEE Transactions on Computers*, vol. 21, 353–359, 1972.

[32] T. Kohonen, *Self-organization and associative memory*, Springer–Verlag, Berlin, 1987.

[33] 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).

[34] C.M. Kuan and K. Hornik, "Convergence of learning algorithms with constant learning rates", *IEEE Transactions on Neural Networks*, vol. 2, 484–488, 1991.

[35] C. Leopold, *Parallel and Distributed Computing: A Survey of Models, Paradigms and Approaches*, John Wiley and Sons, 2000.

[36] 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.

[37] C.G. Looney, *Pattern recognition using neural networks*, Oxford University Press, New York, 1997.

[38] 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.

[39] 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.

[40] 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.

[41] G.D. Magoulas, V.P. Plagianakos, and M.N. Vrahatis, "On globally convergent algorithms with local learning rates", *IEEE Transactions on Neural Networks*, 2002, in press.

[42] G.D. Magoulas, M.N. Vrahatis, and G.S. Androulakis, "Effective back–propagation with variable stepsize", *Neural Networks*, vol. 10, 69–82, 1997.

[43] 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.

[44] W McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity", *Bulletin of Mathematical Biophysics*, vol. 5, 115–133, 1943.

[45] M. Minsky and S. Papert, *Perceptrons*, MIT Press, Cambridge, 1969.

[46] M.F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning", *Neural Networks*, vol. 6, 525–533, 1993.

[47] 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.

[48] J. Nocedal, "Theory of algorithms for unconstrained optimization", In *Acta Numerica*, 199–242, 1992.

[49] J.M. Ortega and W.C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York, 1970.

[50] 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.

[51] V.P. Plagianakos, G.D. Magoulas, N.K. Nousis, 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.

[52] 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.

[53] 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.

[54] V.P. Plagianakos, N.K. Nousis, 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.

[55] 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.

[56] 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.

[57] R.D. Reed and R.J. Marks II, *Neural smithing: Supervised learning in feedforward artificial neural networks*, MIT Press, Cambridge, 1999.

[58] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain", *Psychological Review*, vol. 65, 386–408, 1958.

[59] 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.

[60] 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.

[61] Δ.Ε. Ρυμεληαρτ ανδ Θ.Λ. ΜςΣλελλανδ, εδιτορς, *Παραλλελ διστριβυτεδ προςεσσινγ: Εξπλορατιονσ ιν τηε μιςροστρυςτυρε οφ ςογνιτιον*,ολ. 1, Σαμβριδγε, Μασσαςηυσεττς, 1986, MIT Πρεσς.

[62] 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.

[63] 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.

[64] 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.

[65] K. Sikorski, "Bisection is optimal", *Numerische Mathematik*, vol. 40, 111–117, 1982.

[66] P.P. Van der Smagt, "Minimisation methods for training feedforward neural networks", *Neural Networks*, vol. 7, 1–11, 1994.

[67] D.A. Sprecher, "On the structure of continuous functions of several variables", *Transactions of the American Mathematical Society*, vol. 115, 340–355, 1965.

[68] 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.

[69] G.W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[70] R. Varga, *Matrix Iterative Analysis, Second Edition*, Springer–Verlag, Berlin, 2000.

[71] R.G. Voigt, "Rates of convergence for a class of iterative procedures", *SIAM Journal of Numerical Analysis*, vol. 8, 127–134, 1971.

[72] 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.

[73] M.N. Vrahatis, G.S. Androulakis, J.N. Lambrinos, and G.D. Magoulas, "A class of gradient unconstrained minimisation algorithms with adaptive stepsize", *Journal of Computational and Applied Mathematics*, vol. 114, 367–386, 2000.

[74] 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.

[75] 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.

[76] B. Widrow and M.E. Hoff, "Adaptive switching circuits", In the proceedings of the *1960 IRE WESCON Convention Record*, New York, 96–104, 1960.

[77] J. Wilkinson, *Rounding errors in algebraic processes*, Prentice–Hall, Englewood Cliffs, NJ, 1963.

[78] P. Wolfe, "Convergence conditions for ascent methods", *SIAM Review*, vol. 11, 226–235, 1969.

[79] P. Wolfe, "Convergence conditions for ascent methods ii: Some corrections", *SIAM Review*, vol. 13, 185–188, 1971.

[80] J. Wray and G. Green, "Neural networks, approximation theory and finite precision computation", *Neural Networks*, vol. 8, 31–37, 1995.

[81] W.I. Zangwill, "Minimizing a function without calculating derivatives", *Computer Journal*, vol. 10, 293–296, 1967.

[82] G. Zoutendijk, "Nonlinear programming, computational methods", In *Integer and Nonlinear Programming*, edited by J. Abadie, North–Holland, Amsterdam, 37–86, 1970.

[83] M. Zurada, *Artificial Neural Systems*, West Publishing, St. Paul, 1992.