ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ΤΟΜΕΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΚΠΑΙΔΕΥΣΗ ΜΟΝΟΣΤΡΩΜΑΤΙΚΩΝ ΔΙΚΤΥΩΝ ΣΕ ΠΕΠΕΡΑΣΜΕΝΟ ΑΡΙΘΜΟ ΒΗΜΑΤΩΝ και ΕΠΟΠΤΕΥΟΜΕΝΗ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΒΑΣΙΛΗ ΒΙΡΒΙΛΗ

Αθήνα, 22 Νοεμβρίου 1999

Επιβλέπων καθηγητής: Ν. Καλουπτσίδης

Ευχαριστίες

Μέσα από αυτές τις γραμμές θα ήθελα να εκφράσω τις θερμές ευχαριστίες μου σε όλους εκείνους που είτε άμεσα είτε έμμεσα βοήθησαν στην εκπόνηση αυτής της διδακτορικής διατριβής.

Στον σεβαστό και φίλο επιβλέποντα ερευνητή κ. Σ. Περαντώνη ερευνητή του ΕΚΕΦΕ 'ΔΗΜΟΚΡΙΤΟΣ', ο οποίος παρ' όλες τις πολυποίκιλες υποχρεώσεις του στάθηκε αναντικατάστατος βοηθός στα πλέον κρίσιμα σημεία της διατριβής, τόσο επιστημονικά και διοικητικά όσο θεωρητικά, πρακτικά αλλά και ηθικά.

Στον κ. Κ. Κοντοβασίλη που χωρίς την πολύτιμη συμβολή του σε όλο το φάσμα της γνώσης των ηλεκτρονικών υπολογιστών το τελικό αποτέλεσμα των περισσότερων προγραμμάτων και κειμένων, που χρειάστηκε να υλοποιήσω, θα ήταν πολύ λιγότερο κομψό από ότι είναι τώρα, και εγώ πολύ φτωχότερος σε γνώσεις.

Ιδιαίτερες ευχαριστίες αρμόζουν στον κ. Ν. Καλουπτσίδη, καθηγητή του Πανεπιστημίου Αθηνών, χωρίς τον οποίο δεν θα είχα ούτε την ευκαιρία ούτε τη δυνατότητα να εκπονήσω αυτή την εργασία.

Ιδιαιτέρως θα ήθελα να εκφράσω τις ευχαριστίες μου στον σεβαστό κ. Σ. Βαρουφάκη, αλλά και στο εργαστήριο Νευρωνικών δικτύων του Ινστιτούτου Πληροφορικής και Τηλεπικοινωνιών του ΕΚΕΦΕ 'ΔΗΜΟΚΡΙΤΟΣ', χωρίς την πολύπλευρη υποστήριξη του οποίου, ηθική, οικονομική και επιστημονική θα ήταν δύσκολη η ολοκλήρωση αυτής της διατριβής.

Θα ήταν μεγάλη παράλειψη αν δεν αναφερόταν η σημαντικότατη βοήθεια, τόσο από πλευράς οικονομικής όσο και από πλευράς υποδομής σε υπολογιστές και περιφερειακά, βιβλιοθήκης κλπ. του ερευνητικού κέντρου 'ΔΗΜΟΚΡΙΤΟΣ' σε όλα τα στάδια εκπόνησης αυτής της διατριβής.

Τέλος, αλλά όχι λιγότερο, θα ήθελα να εκφράσω τις ευχαριστίες μου στην οικογένειά μου και ιδιαίτερα στον πατέρα μου και τους θείους μου που στάθηκαν δίπλα μου στα χρόνια εκπόνησης αυτής της εργασίας. Χωρίς την συμπαράστασή τους θα ήταν πραγματικά αδύνατη η ολοκλήρωσή της.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο Διατριβής

Ο στόχος αυτής της διατριβής είναι η διερεύνηση τεχνικών αποτελεσματικής μάθησης στα Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ), αποσκοπώντας αφενός στην επιτυχή χρήση τους σε εφαρμογές που απαιτούν αποκρίσεις των συστημάτων σε πραγματικό χρόνο (real time) και, κυρίως, αφετέρου στην βελτίωση των ιδιοτήτων τους ώστε να παρουσιάζουν καλύτερες επιδόσεις σε προβλήματα αναγνώρισης προτύπων του πραγματικού κόσμου. Οι μελετώμενες και προτεινόμενες τεχνικές εξετάζονται σε δύο από τα σημαντικότερα προβλήματα στο χώρο των ΤΝΔ:

1. Μονοστρωματικά δίκτυα τύπου perceptron
2. Προεπεξεργασία δεδομένων

1.1.1 Μονοστρωματικά δίκτυα τύπου perceptron

Τα μονοστρωματικά δίκτυα είναι τα πιο απλά μοντέλα τεχνητών νευρωνικών δικτύων που υπάρχουν στον τομέα των δικτύων πρόσω τροφοδότησης (feed forward). Εξ' αιτίας των υπερβολικών προσδοκιών [1, 2] της πρώιμης εποχής των νευρωνικών δικτύων και της επίθεσης του Minsky [3] στο perceptron του Rosenblatt [4] το ενδιαφέρον για τα μονοστρωματικά δίκτυα ατόνησε. Το βασικό επιχείρημα της πολεμικής του Minsky ήταν ο εγγενής περιορισμός του perceptron στη λύση μόνο γραμμικά διαχωρίσιμων προβλημάτων.

Παρ' όλο που αυτό είναι αλήθεια το perceptron εξακολουθεί να είναι να είναι χρήσιμο από ερευνητικής σκοπιάς για αρκετούς λόγους:

- Αποτελεί το δομικό λίθο των πολυστρωματικών δικτύων [5, 6, 7] που είναι σήμερα η αιχμή τόσο της έρευνας όσο και της τεχνολογίας των τεχνητών νευρωνικών δικτύων σε επίπεδο εφαρμογών.
- Είναι το μοναδικό νευρωνικό δίκτυο που έχει απόδειξη σύγκλισης σε πεπερασμένο αρκιθμό βημάτων. Αυτό από θεωρητική άποψη αποτελεί θελκτική ιδιότητα. Είναι αναμενόμενη η επιθυμία να μπορέσουν να μελετηθούν οι πιθανές γενικεύσεις μιας τέτοιας ιδιότητας σε άλλου τύπου δίκτυα.
- Αποδεικνύεται ότι το perceptron δεν έχει τοπικά ελάχιστα. Παρ' όλα αυτά παρουσιάζει παρόμοια συμπεριφορά με άλλα νευρωνικά δίκτυα όσον αφορά τις δυσκολίες που έχει κατά την εκπαίδευσή του, πράγμα που κάνει εξαιρετικά πολύτιμη την μελέτη του.
- Μπορεί να χρησιμοποιηθεί και σε μη γραμμικά διαχωρίσιμα προβλήματα, με κατάλληλη γραμμικοποίηση των εισόδων του.

- Εφ' όσον το perceptron δεν χρησιμοποιεί τη σιγμοειδή, δεν παρουσιάζει το φαινόμενο της 'μυωπίας' που παρουσιάζουν άλλα νευρωνικά δίκτυα. Είναι δηλαδή το ίδιο ευαίσθητο σε όλα τα λάθος ταξινομημένα πρότυπα εισόδου.

- Εξ' αιτίας της απεμπλοκής του από τη σιγμοειδή το perceptron δύναται να λειτουργήσει σε ακατέργαστα, μη κανονικοποιημένα, δεδομένα.

Ακόμα όμως και για το perceptron δεν έχουν υπερνικηθεί όλες οι θεωρητικές και πρακτικές δυσκολίες. Πιο συγκεκριμένα:

- Το perceptron παρουσιάζει καλύτερους χρόνους εκμάθησης, όταν χρησιμοποιούνται ευριστικές παράμετροι όπως όροι ορμής (momentum) και μεταβλητού ρυθμού εκμάθησης (learning rate). Είναι περιττό να πούμε ότι το θεώρημα σύγκλισης παύει να ισχύει με την επιβολή τέτοιων ευριστικών μεθόδων και παραμέτρων στην διαδικασία της εκμάθησης.

- Το perceptron παρ' όλο που έχει θεώρημα σύγκλισης στην περίπτωση που το πρόβλημα είναι γραμμικά διαχωρίσιμο, δεν παρέχει κανένα κριτήριο τερματισμού στην περίπτωση που το πρόβλημα δεν είναι επιλύσιμο. Αυτή η έλλειψη αποτελεί σημαντική τροχοπέδη για τον σχεδιασμό αποτελεσματικών αλγορίθμων εκπαίδευσης πολυστρωματικών δικτύων που να βασίζονται στο perceptron [8, 9].

- Το θεώρημα σύγκλισης του perceptron αναφέρεται στον κανόνα για την on line ενημέρωση των βαρών του. Παρ' όλο που οι on line αλγόριθμοι συνεχίζουν την νευροβιολογική παράδοση των νευρωνικών δικτύων οι off line (batch) εκδόσεις των αλγορίθμων έχουν συνήθως πληρέστερο μαθηματικό υπόβαθρο. Αυτό κάνει πιο εύκολη την χρησιμοποίηση δοκιμασμένων αναλυτικών μαθηματικών εργαλείων που προέρχονται από τον χώρο της βελτιστοποίησης.

1.1.2 Προεπεξεργασία δεδομένων

Οι υπάρχουσες μέθοδοι προεπεξεργασίας δεδομένων αποτυγχάνουν να αντιμετωπίσουν τις ολοένα αυξανόμενες ανάγκες της επιστημονικής κοινότητας. Ο λόγος είναι ότι με την αύξηση της υπολογιστικής δύναμης των σύγχρονων υπολογιστών είναι δυνατόν να έχουμε πρόσβαση σε όλο και μεγαλύτερα προβλήματα που μέχρι πρότινος ήταν απροσπέλαστα από υπολογιστικής απόψεως.

Δυστυχώς όμως ενώ η αύξηση της υπολογιστικής ισχύος είναι γραμμική, η συνδυαστική έκρηξη που προκύπτει από την αύξηση του πλήθους των διαστάσεων ενός προβλήματος είναι εκθετική. Αυτό σημαίνει επίσης ότι η δυνατότητα επισκόπησης σε αυτού του τύπου τα προβλήματα είναι αδύνατη. Η προεπεξεργασία των δεδομένων έχει σαν στόχο τη μείωση της διάστασης του χώρου του προβλήματος με τέτοιο τρόπο ώστε να μην μειώνεται όμως η υπάρχουσα πληροφορία.

Μέχρι πρόσφατα δεν είχε συλληφθεί η ιδέα της χρησιμοποιήσεως των νευρωνικών δικτύων για μια τέτοια εργασία, με αποτέλεσμα όλες οι μέθοδοι να έχουν στατιστική υφή. Με την χρησιμοποίηση των νευρωνικών δικτύων μπορούμε να πετύχουμε την απαραίτητη μείωση των δεδομένων, με παράλληλο στόχο όμως τη διατήρηση ή την αύξηση της γενικευτικής ικανότητας, ποσότητας που αποτελεί ένα καλό μέτρο για την ποσότητα και ποιότητα της πληροφορίας που εμπεριέχει το τελικό σύνολο δεδομένων.

Η βασική ιδέα της μεθόδου συνίσταται στην εξαγωγή κατευθύνσεων, ορθογωνίων μεταξύ τους, τέτοιων που να μεγιστοποιούν την απόκριση (ευαισθησία) του νευρωνικού δικτύου που έχει εκπαιδευτεί με τα αρχικά δεδομένα. Κρατώντας τις πρώτες (σημαντικότερες) διευθύνσεις μπορούμε να επιτύχουμε μια πολύ μεγάλη μείωση του αριθμού των διαστάσεων χωρίς όμως αντίστοιχη μείωση της γενικευτικής ικανότητας. Όπως θα δείξουμε και παρακάτω η τεχνική καταλήγει στο μαθηματικό ισοδύναμο της εφαρμογής της δημοφιλούς μεθόδους PCA στον εφαπτόμενο χώρο των παραγώγων (αποκρίσεων) του νευρωνικού δικτύου.

Με την προτεινόμενη τεχνική πετυχαίνουμε να συνδυάσουμε, μ' εξαιρετικά αποτελέσματα τη μέθοδο PCA με τα νευρωνικά δίκτυα και να αξιοποιήσουμε την a-priori πληροφορία που υπάρχει κατά τη διάρκεια της εποπτευόμενης μάθησης. Αξίζει να σημειωθεί ότι οι μέχρι τώρα προσπάθειες εμπλοκής της μεθόδου

PCA στην τεχνολογία των νευρωνικών δικτύων είχαν σαν στόχο περισσότερο την υλοποίησή της [10, 11, 12] και όχι τη χρησιμοποίησή της.

1.2 Ανάλυση περιεχομένων των κεφαλαίων

Στο κεφάλαιο 2 ορίζεται το πρόβλημα της γραμμικής διαχωρισιμότητας και δίδονται μερικά παραδείγματα του δυϊκού χώρου, τα οποία προσφέρονται για οπτική επισκόπηση. Με την νεοαποκτηθείσα διαίσθηση σχεδιάζουμε και προτείνουμε έναν αλγόριθμο επιτρεπτών διευθύνσεων (feasible direction). Αν θεωρήσουμε επιπλέον την απαίτηση ο αλγόριθμος να ακολουθεί τη βέλτιστη επιτρεπτή διεύθυνση τότε αποδεικνύεται ότι ο προτεινόμενος αλγόριθμος **FLF** τερματίζει σε πεπερασμένο αριθμό βημάτων. Στην περίπτωση που το πρόβλημα είναι γραμμικά διαχωρίσιμο ο αλγόριθμος τερματίζει όταν βρει τη λύση. Όταν το πρόβλημα δεν είναι επιλύσιμο ο αλγόριθμος πάλι τερματίζει σε πεπερασμένα βήματα.

Στο κεφάλαιο 3 επανέρχεται το πρόβλημα της εύρεσης της βέλτιστης επιτρεπτής διεύθυνσης η λύση του οποίου θεωρήθηκε δεδομένη στο κεφάλαιο 2. Αποδεικνύεται ότι πρόκειται για ένα πρόβλημα τετραγωνικού προγραμματισμού. Αφού μελετούνται οι ιδιότητες του και γίνεται η απαιτούμενη βιβλιογραφική αναφορά, συγκρίνονται οι ήδη υπάρχοντες αλγόριθμοι και εξάγονται ομοιότητες και διαφορές. Όπως είναι εύκολο να δει κανείς η διεθνής βιβλιογραφία θεωρεί τα δύο διαφορετικά προβλήματα σαν ένα, πράγμα που κάνει πολύ δύσκολο να σχεδιαστούν νέοι και αποδοτικοί αλγόριθμοι. Τέλος προτείνονται δύο διαφορετικοί αλγόριθμοι, που επειδή έχουν μια διαφορά 2 χρόνων στη σύλληψη και υλοποίησή τους αναφέρονται όχι ως εναλλακτικές επιλογές αλλά ως φυσική μετεξέλιξη ο ένας του άλλου. Ο δεύτερος αλγόριθμος μάλιστα αποδεικνύεται ότι είναι πεπερασμένων βημάτων πράγμα που κάνει όλη τη διαδικασία σύγκλισης πεπερασμένη.

Στο κεφάλαιο 4 δοκιμάζεται ο προτεινόμενος αλγόριθμος τόσο σε τεχνητά προβλήματα όσο και σε πραγματικά προβλήματα. Συγκρίνεται με κλασσικούς αλγόριθμος όπως η Simplex και το perceptron σε όρους ταχύτητας, απόδοσης, και γενικευτικής ικανότητας, καθώς και με μερικούς όχι τόσο κλασσικούς όπως αυτός του Bobrowski. Από τις συγκρίσεις αυτές εξάγονται συμπεράσματα τα οποία φυσικά συζητούνται εκτενώς.

Το κεφάλαιο 5 είναι αφιερωμένο στην προεπεξεργασία δεδομένων. Η προτεινόμενη μέθοδος αναπτύσσεται έτσι ώστε το μαθηματικό της υπόβαθρο να ταιριάζει με αυτό της PCA, ενώ δείχνεται καθαρά ότι αποτελεί και γενίκευση της μεθόδου του Ruck. Στο ίδιο κεφάλαιο παρουσιάζονται και τα πειραματικά αποτελέσματα όπου δείχνουν μια συντριπτική μείωση των εισόδων του προβλήματος με ταυτόχρονη αύξηση της γενικευτικής ικανότητας του δικτύου. Τέλος εξετάζεται η πιθανή χρήση της μεθόδου για εφαρμογές οπτικοποίησης είτε δεδομένων είτε νευρωνικών δικτύων.

Επειδή κατά τη διάρκεια της διατριβής ήταν απαραίτητο να γίνουν πολλαπλές και εκτεταμένες εξομοιώσεις αναπτύχθηκε μια πλατφόρμα λογισμικού (billnet) που βοήθησε τα μέγιστα σε αυτόν τον σκοπό και η οποία περιγράφεται στο παράρτημα Α΄.

1.3 Συμβολή, πρωτοτυπία και δημοσιεύσεις

Όσον αφορά στην εκπαίδευση των μονοστρωματικών δικτύων η συμβολή και η πρωτοτυπία της παρουσιαζόμενης διατριβής έγκειται στα ακόλουθα σημεία:

- Ο προτεινόμενος αλγόριθμος **FLF** έχει απόδειξη σύγκλισης σε πεπερασμένα βήματα χρησιμοποιώντας τον κανόνα του off line perceptron. Σε περίπτωση που το πρόβλημα είναι γραμμικά διαχωρίσιμο, τότε ο αλγόριθμος βρίσκει μια αποδεκτή λύση. Σε αντίθετη περίπτωση, ο αλγόριθμος τερματίζει σε μια συνήθως 'καλή' μη αποδεκτή λύση, όχι όμως απαραίτητα και την καλύτερη δυνατή.

- Ο προτεινόμενος αλγόριθμος (**FLF**) δεν έχει καθόλου ευριστικές παραμέτρους.

- Είναι πολύ γρήγορος. Παρ' όλο που υπάρχουν ισχυρές ενδείξεις για πολυωνυμική σύγκλιση δεν υπάρχει ακόμα αυστηρή απόδειξη.

- Είναι αυτοκλιμακούμενος. Αντίθετα με τη συνήθη πρακτική των υπαρχόντων νευρωνικών δικτύων να απαιτούν κανονικοποιημένες εισόδους για να αποφύγουν προβλήματα αποκοπής από τη σιγμοειδή (overshooting, saturation), ο προτεινόμενος αλγόριθμος μπορεί να χρησιμοποιήσει 'ακατέργαστα στοιχεία' (raw data), χωρίς μείωση στην ταχύτητα εκμάθησης ή στη γενικευτική ικανότητα.

- Αναλύει το μεγάλο γραμμικό πρόβλημα γραμμικού προγραμματισμού σε πολλά μικρότερα, τετραγωνικού προγραμματισμού αναζητώντας μια φυσική διεύθυνση κίνησης.

- Το πρόβλημα τετραγωνικού προγραμματισμού που βρίσκει την βέλτιστη επιτρεπτή διεύθυνση πρώτη φορά διατυπώνεται για τη λύση του γραμμικού προβλήματος. Το αν αξίζει να λύσει κανείς μυριάδες τετραγωνικά προβλήματα για μπορέσει να λύσει το γραμμικό συζητείται στο κεφάλαιο των πειραματικών αποτελεσμάτων (κεφ. 4) όπου και γίνεται εκτενής σύγκριση και ανάλυση όλων των μεθόδων.

- Η λύση που προτείνεται για το τετραγωνικό υποπρόβλημα είναι κι' αυτή πρωτότυπη.

- Η πειραματική απόδειξη, σε γραμμικά και όχι διαχωρίσιμα προβλήματα, του ότι χρησιμοποιώντας έναν σχετικά αποτελεσματικό αλγόριθμο τετραγωνικού προγραμματισμού (αλγόριθμος διπλής αναζήτησης) για την εύρεση της βέλτιστης διεύθυνσης είναι δυνατόν να εκπαιδεύσει κανείς μονοστρωματικά δίκτυα τύπου perceptron πολύ αποδοτικότερα απ' ότι με άλλες τεχνικές.

- Η δυνατότητα του αλγόριθμου να μην αυξάνει τον αριθμό των λάθος ταξινομημένων προτύπων είναι πραγματικά πολύτιμη. Χρησιμοποιώντας αυτή την ιδιότητα σαν δομικό λίθο είναι δυνατόν να στηρίξει κανείς το οικοδόμημα ενός διστρωματικού δικτύου. Επίσης, με ελάχιστες μετατροπές είναι δυνατόν ο αλγόριθμος να μετασχηματιστεί έτσι ώστε να εντοπίζει τα πραγματικά τοπικά ελάχιστα της γραμμικής διαχωρισιμότητας στην περίπτωση των μη γραμμικά διαχωρίσιμων προβλημάτων.

  Αυτή η ιδιότητα είναι τόσο σημαντική, πέρα από όρους ταχύτητας και απόδοσης, που μια άλλη πιο βιολογική διατύπωση θα ήταν ίσως καταλληλότερη. Συγκεκριμένα θα μπορούσε να πει κανείς ότι *Το σύστημα αυτό, έτσι όπως εκπαιδεύεται από τον προτεινόμενο αλγόριθμο, έχει την δυνατότητα της αποθήκευσης νέας γνώσης χωρίς όμως να 'ξεχνάει' την παλιά.*

Τα συμπεράσματα και οι προτεινόμενες τεχνικές έχουν ήδη δημοσιευθεί [13, 14, 15, 16], ενώ παραμένουν ανοικτά ακόμα τα ακόλουθα ζητήματα:

- Απόδειξη ότι η διαδικασία σύγκλισης έχει πολυωνυμική πολυπλοκότητα.

- Εφαρμογή των παραπάνω γνώσεων και ιδιοτήτων σε πολυστρωματικά δίκτυα.

Όσον αφορά τώρα την προεπεξεργασία δεδομένων η πρωτότυπη χρήση νευρωνικών δικτύων για εξαγωγή σημαντικών κατευθύνσεων μπορεί να έχει τις εξής εφαρμογές:

- Οι προκύπτουσες κατευθύνσεις είναι σημαντικές όχι μόνο για νευρωνικούς αλλά και για στατιστικούς ταξινομητές, πράγμα που σημαίνει ότι υπάρχει κάποιος βαθύτερος λόγος που εγγυάται την καλή απόδοση της μεθόδου.

- Το πλήθος των σημαντικών διαστάσεων που προκύπτουν είναι συνήθως πολύ μικρότερο από αυτό του αρχικού προβλήματος, χωρίς όμως να υπάρχει και μείωση στη γενικευτική ικανότητα. Το γεγονός αυτό ανοίγει τον δρόμο για:

  1. Εφαρμογές γρήγορων υλοποιήσεων (fast knn) του αλγόριθμου του κοντινότερου γείτονα. Το βασικό πρόβλημα αυτών των μεθόδων έγκειται στο γεγονός ότι η απόδοσή τους μειώνεται εξαιρετικά σε πολυδιάστατα προβλήματα.
  2. Εφαρμογές οπτικοποίησης.

Τα συμπεράσματα και οι προτεινόμενες τεχνικές έχουν ήδη δημοσιευθεί [17, 18, 19, 20, 21, 22], ενώ παραμένουν ανοικτά ακόμα τα ακόλουθα ζητήματα:

- Μετατροπή του αλγορίθμου knn ώστε να είναι δυνατόν να δίνει ένα μέτρο της σημαντικότητας αντί για την πληροφορία κλάσης. Με αυτόν τον τρόπο η προτεινόμενη μέθοδος είναι δυνατόν να απεμπλακεί εντελώς από τη χρήση των νευρωνικών δικτύων.
- Πολύς λόγος γίνεται τελευταία για την ανάλυση σε ανεξάρτητους άξονες (ICA) αντί σε βασικούς (PCA). Η πρώτη φέρεται μάλιστα ως γενίκευση της δεύτερης. Είναι μάλλον θέμα ενός άλλου διδακτορικού η διερεύνηση της παραπάνω ιδιότητας.

Κεφάλαιο 2

Το γραμμικά διαχωρίσιμο πρόβλημα

Το πρόβλημα που θα μας απασχολήσει σ΄ αυτήν την ενότητα είναι η εύρεση μιας αποδεκτής (feasible) λύσης για ένα σύστημα γραμμικών ανισώσεων. Αν και η λύση στο σύστημα των γραμμικών εξισώσεων κοστίζει υπολογιστικά όσο και η αντιστροφή ενός πίνακα ($O(n^3)$), για το σύστημα των γραμμικών ανισώσεων δεν έχει προταθεί αλγόριθμος αποδεδειγμένα πολυωνυμικού χρόνου.

Το πρόβλημα αυτό είναι πάρα πολύ σημαντικό τόσο στην ταξινόμηση (classification) και την αναγνώριση προτύπων (pattern recognition) όσο και στη βελτιστοποίηση συναρτήσεων, διότι κάθε ανισότητα παριστάνει ένα δεσμό που πρέπει να ικανοποιηθεί προτού αναζητηθεί η βέλτιστη λύση μέσα στο χώρο των αποδεκτών λύσεων με κάποια άλλη μέθοδο. Στην περίπτωση που οι δεσμοί είναι μη γραμμικοί τότε μπορούμε με αύξηση της διάστασης να γραμμικοποιήσουμε το πρόβλημα. Αυτό παράγει όμως χώρους με στρεβλές κατανομές που σε συνδυασμό με την κατάρα της πολυδιαστικότητας (curse of dimensionality) παγιδεύει σε χαμηλή απόδοση αλγόριθμους που βασίζονται σε στατιστικά ή νευρωνικά μοντέλα (perceptron). Ειδικότερα, τα νευρωνικά δίκτυα υποφέρουν επιπλέον από ένα πλήθος ευριστικών παραμέτρων που αν η τιμή τους δεν επιλεγεί κατάλληλα, ουσιαστικά απαγορεύει την σύγκλιση σε λογικό χρόνο.

Η πρότασή μας βασίζεται σ' έναν αλγόριθμο που:

- Έχει απόδειξη σύγκλισης σε πεπερασμένα βήματα. Σε περίπτωση που το πρόβλημα είναι γραμμικά διαχωρίσιμο, τότε ο αλγόριθμος βρίσκει μια αποδεκτή λύση. Σε αντίθετη περίπτωση, ο αλγόριθμος τερματίζει σε μια συνήθως 'καλή' μη αποδεκτή λύση, όχι όμως απαραίτητα και την καλύτερη δυνατή.

- Δεν έχει καθόλου ευριστικές παραμέτρους.

- Στην πορεία της μάθησης ο αλγόριθμος δεν καταστρέφει τους ήδη ικανοποιημένους δεσμούς, με αποτέλεσμα να μην παρουσιάζονται ταλαντεύσεις και να μειώνει το συνολικό χρόνο εκπαίδευσης. Μια άλλη διατύπωση της παραπάνω πρότασης με μαθησιακούς όρους, είναι ότι το σύστημα δεν ξεχνάει όσα έχει ήδη μάθει σωστά.

- Είναι πολύ γρήγορος. Παρ΄ όλο που υπάρχουν ισχυρές ενδείξεις για πολυωνυμική σύγκλιση δεν υπάρχει ακόμα αυστηρή απόδειξη.

- Είναι αυτοκλιμακούμενος. Αντίθετα με τη συνήθη πρακτική των υπαρχόντων νευρωνικών δικτύων να απαιτούν κανονικοποιημένες εισόδους για να αποφύγουν προβλήματα αποκοπής από τη σιγμοειδή (overshooting, saturation), ο προτεινόμενος αλγόριθμος μπορεί να χρησιμοποιήσει 'ακατέργαστα στοιχεία' (raw data), χωρίς μείωση στην ταχύτητα εκμάθησης ή στη γενικευτική ικανότητα.

2.1 Μαθηματική διατύπωση του προβλήματος

2.1.1 Ο χώρος των βαρών

Το ζητούμενο είναι να βρούμε την εξίσωση ενός υπερεπιπέδου που διαχωρίζει πλήρως P διανύσματα (pattern) ($\boldsymbol{X}_p \quad p = 1, \ldots, P$) που ανήκουν σε δύο κλάσεις, έστω C_1, C_2. Το πρόβλημα πρέπει να λυθεί γενικά στις N διαστάσεις.

Θέλουμε, δηλαδή, ένα βρούμε ένα σύνολο μεταβλητών $\boldsymbol{w}, w_0$, τέτοιων ώστε

$$\boldsymbol{w}^\tau \boldsymbol{X}_p + w_0 \quad > \quad 0 \quad \forall p \in C_1 \tag{2.1}$$

$$\boldsymbol{w}^\tau \boldsymbol{X}_p + w_0 \quad < \quad 0 \quad \forall p \in C_2 \tag{2.2}$$

Οι παραπάνω εξισώσεις γράφονται σε πιο συμπαγή μορφή αν θεωρήσουμε ότι

- Το διάνυσμα $\boldsymbol{X}_p$ περιέχει μια επιπλέον διάσταση που η συντεταγμένη της όμως είναι πάντα 1 για όλα τα σημεία (threshold).
- Κατ' αναλογία, το διάνυσμα $\boldsymbol{W}$ αποτελείται από το $\boldsymbol{w}$ και περιέχει τον όρο w_0 σαν τελευταία συνιστώσα.
- Αντιστοιχούμε μια δυαδική τιμή T_p (0 ή 1) σε κάθε διάνυσμα $\boldsymbol{X}_p$. Η μεταβλητή T_p παίρνει τιμή 1 όταν το σημείο p ανήκει στην κλάση C_1 και 0 αν ανήκει στην κλάση C_2.
- Ορίζουμε το διάνυσμα $\boldsymbol{d}_p = (2T_p - 1)\boldsymbol{X}_p$

Έτσι το αρχικό μας πρόβλημα τώρα γίνεται

$$\theta(\boldsymbol{W}^\tau \boldsymbol{d}_p) = 1 \quad p = 1, \ldots, P \tag{2.3}$$

όπου θ είναι η γνωστή συνάρτηση βήματος.

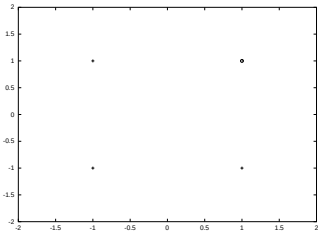
Στον χώρο των βαρών των $N + 1$ διαστάσεων το διάνυσμα $\boldsymbol{W}$ αντιπροσωπεύει ένα σημείο. Κάθε διάνυσμα εισόδου $\boldsymbol{X}_p$ αποτελεί ένα υπερεπίπεδο που διέρχεται από την αρχή των αξόνων και χωρίζει το χώρο των βαρών σε 2 υποχώρους. Τα υπερεπίπεδα που αντιστοιχούν στα διανύσματα $\boldsymbol{d}_p$ (και στα $\boldsymbol{X}_p$) θρυμματίζουν το χώρο σε πολλαπλά κυρτά (convex) πολύτοπα των οποίων τα κοινά όρια είναι τμήματα των υπερεπιπέδων. Για κάθε $\boldsymbol{W}$ που δίδεται, κάθε υπερεπίπεδο ταξινομείται σαν 'σωστό' (BR, Bit Right) αν η ποσότητα $O_p = \theta(\boldsymbol{W}^\tau \boldsymbol{X}_p)$ είναι ίση με την τιμή στόχο T_p για το συγκεκριμένο διάνυσμα p, και 'λάθος' (BW, Bit Wrong) αν το O_p είναι διαφορετικό από το T_p.

Μια χρήσιμη παρατήρηση είναι ότι το διάνυσμα $\boldsymbol{d}_p = (2T_p - 1)\boldsymbol{X}_p$ δείχνει πάντοτε προς την πλευρά του υπερεπιπέδου που αντιστοιχεί στη 'σωστή' ταξινόμηση του διανύσματος $\boldsymbol{X}_p$. Τα διανύσματα που ταξινομούνται σαν BR χαρακτηρίζονται επίσης και από την θετική τιμή της ποσότητας $Z_p = \boldsymbol{W}^\tau \boldsymbol{d}_p$.

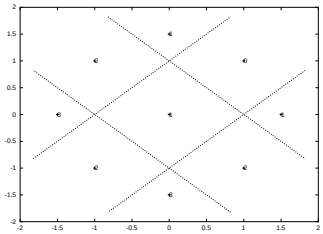
Για να γίνει κατανοητή η προηγούμενη δήλωση ας θεωρήσουμε ότι το γινόμενο $\boldsymbol{W}^\tau \boldsymbol{X}_p > 0$. Σ' αυτή την περίπτωση το T_p πρέπει να ισούται με 1 αφού θεωρήσαμε το $\boldsymbol{X}_p$ σαν BR. Έτσι, $2T_p - 1 = 1 > 0$ και $Z_p > 0$. Στην αντίθετη περίπτωση, αν ισχύει ότι $\boldsymbol{W}^\tau \boldsymbol{X}_p < 0$, το T_p πρέπει να είναι ισούται με 0. Αλλά τότε $2T_p - 1 = -1 < 0$ και άρα το Z_p είναι πάλι θετικό.

Με παρόμοια συλλογιστική μπορούμε να δείξουμε ότι στην περίπτωση που ένα διάνυσμα έχει ταξινομηθεί σαν BW από το $\boldsymbol{W}$, χαρακτηρίζεται από την αρνητική τιμή του εσωτερικού γινομένου Z_p. Από αυτό το σημείο και έπειτα θα αναφερόμαστε στα διανύσματα εισόδου σαν $\boldsymbol{d}_p$ αντί για τα αρχικά $\boldsymbol{X}_p$ αφού όπως είδαμε διατηρούμε όλη την απαιτούμενη πληροφορία του αρχικού προβλήματος στην πιο συμπαγή μορφή της 2.3.

Η διάσταση του προβλήματος αυξάνεται κατά ένα βαθμό ελευθερίας αλλά επειδή έχουμε χρησιμοποιήσει ομογενείς συντεταγμένες μπορούμε να το σκεφτόμαστε σαν πρόβλημα N διαστάσεων, ειδικά στον δυϊκό χώρο των βαρών όπου η μόνη παραμόρφωση είναι στην κλίμακα. Αργότερα, θα δούμε πως μπορούμε να χρησιμοποιήσουμε την επιπλέον μεταβλητή για να αποφύγουμε τον διαχωρισμό του χώρου σε δύο διαφορετικές μη γειτονικές περιοχές.



AND		
Inputs		Target
-1	-1	0
-1	1	0
1	-1	0
1	1	1



Σχήμα 2.1: Το εικονιζόμενο πρόβλημα είναι το and. Μπορεί κανείς να παρατηρήσει την αντιστοιχία σημείων σε γραμμές στο δυϊκό χώρο.

2.1.2 Συναρτήσεις κόστους

Υπάρχουν 2 συναρτήσεις κόστους που σχετίζονται με το πρόβλημα που μελετούμε:

- Η συνάρτηση κόστους του perceptron η οποία ορίζεται

$$E = \sum_{p=1}^{P} (T_p - O_p) \boldsymbol{W}^\tau \boldsymbol{X}_p = - \sum_{p=BW} \boldsymbol{W}^\tau \boldsymbol{d}_p \tag{2.4}$$

  όπου το δεύτερο άθροισμα τρέχει πάνω στα διανύσματα που είναι ταξινομημένα σαν BW από το $\boldsymbol{W}$. Σύμφωνα με την ανάλυση που έγινε στο κεφάλαιο 2.1.1, η συνάρτηση κόστους είναι θετική. Είναι επίσης γραμμική ανά περιοχή (piecewise linear) στον R^{N+1} και έχει σταθερή παράγωγο $\boldsymbol{\Delta W}$ σε κάθε πολύτοπο που δίνεται από τον τύπο

$$-\boldsymbol{\Delta W} = \sum_{p=1}^{P} (T_p - O_p) \boldsymbol{X}_p = - \sum_{p=BW} \boldsymbol{d}_p \tag{2.5}$$

  Αν επιχειρήσουμε να ακολουθήσουμε αυτήν την κατεύθυνση, μ' αυτήν την συνάρτηση κόστους (gradient descent) έχουμε ουσιαστικά την offline (batch) ποικιλία του αλγόριθμου του Rosenblatt για το perceptron.

- Η τετραγωνική συνάρτηση κόστους

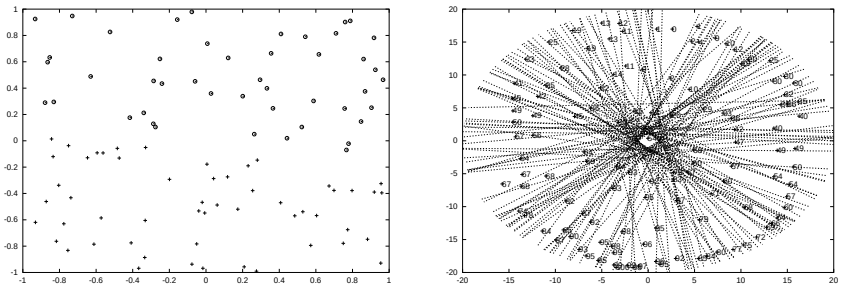
$$E_{SE} = \sum_{p=1}^{P} \left(\theta(\boldsymbol{W}^\tau \boldsymbol{d}_p) - 1\right)^2 \tag{2.6}$$

  η οποία μετράει τον αριθμό των λάθος ταξινομημένων διανυσμάτων BW από το τρέχον $\boldsymbol{W}$ και προφανώς έχει σταθερή τιμή για κάθε πολύτοπο.

2.1.3 Παραδείγματα

and Η απεικόνιση του λογικού και στο χώρο των βαρών (βλ. σχήμα 2.1) είναι το πιο απλό (με πιθανή εξαίρεση το λογικό ή) γραμμικά διαχωρίσιμο πρόβλημα το οποίο μπορούμε να χρησιμοποιήσουμε για να επιδείξουμε μερικές από τις ιδέες του προηγούμενου κεφαλαίου.

Μερικές χρήσιμες παρατηρήσεις:



Σχήμα 2.2: Εδώ το πρόβλημα απαρτίζεται από 100 σημεία. Είναι σαφώς πιο δύσκολο αν και η κατανομή των σημείων στο επίπεδο είναι ομογενής.

- Τα σημεία $\boldsymbol{d}_p$ έχουν γίνει ευθείες και αποτελούν πλέον μια διάταξη (arrangement), χωρίζοντας τον χώρο σε κυρτές (convex) πολυγωνικές υπερπεριοχές (polytopes). Όλα τα γνωστά θεωρήματα για τις διατάξεις ισχύουν [23].
- Μέσα σε κάθε πολύτοπο ο αριθμός των BW είναι σταθερός, και το πολύτοπο χαρακτηρίζεται απ' αυτόν. Είναι προφανές ότι τα γειτονικά πολύτοπα διαφέρουν κατά ένα BW.
- Μέσα σε κάθε πολύτοπο το διάνυσμα κατεύθυνσης $\boldsymbol{\Delta W}$ του perceptron παραμένει σταθερό.
- Η ζητούμενη λύση δεν είναι μοναδική. Αποδεκτή λύση είναι οποιοδήποτε $\boldsymbol{W}$ ανήκει στο πολύτοπο που έχει μηδενικό αριθμό (0) BW, γεγονός που είναι απολύτως λογικό αφού το αρχικό πρόβλημα είναι πρόβλημα συστήματος ανισώσεων.
- Το διάνυσμα $\boldsymbol{d}_p$ έτσι όπως ορίστηκε δείχνει πάντα από την 'σωστή' (BR) πλευρά του υπερεπιπέδου p και είναι κάθετο ως προς αυτό. Δηλαδή, αν διασχίσουμε τα όρια του πολυτόπου ομόρροπα προς το αντίστοιχό $\boldsymbol{d}_p$, τότε μειώνουμε των αριθμό των BW. Αν όμως περάσουμε στο γειτονικό πολύτοπο αντίρροπα προς το $\boldsymbol{d}_p$, τότε αυξάνουμε τον αριθμό των BW.

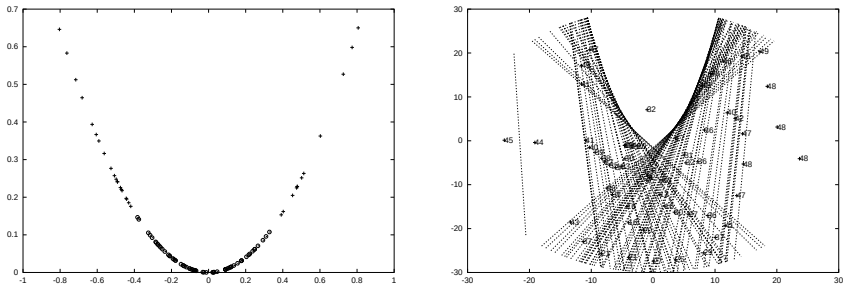
Ομογενής κατανομή Το επόμενο πρόβλημα (σχήμα 2.2) είναι σαφώς πιο πολύπλοκο αν και όχι ουσιαστικά διαφορετικό από το λογικό και. Πρόκειται για μια ομογενή κατανομή 100 σημείων στις δύο διαστάσεις που αποτελείται από δύο γραμμικά διαχωρίσιμες κλάσεις.

Η λύση του προβλήματος βρίσκεται ψηλά στην άκρη του σχήματος. Όπως μπορεί να δει κανείς, αυτή την φορά το πολύτοπο που λύνει το πρόβλημα εκτείνεται στο άπειρο, είναι δηλαδή ανοικτό (unbounded). Στην πραγματικότητα όλα τα πολύτοπα είναι ανοικτά και χωρίς όρια, τουλάχιστον ως προς την διεύθυνση του w_0. Δεν πρέπει να ξεχνάμε ότι ένα πρόβλημα N διαστάσεων εικονίζεται στον R^{N+1} στο δυϊκό χώρο των βαρών. Αν αναλογιστεί όμως κανείς ότι το μηδενικό διάνυσμα $\boldsymbol{W}^\tau = [\boldsymbol{w}^\tau, w_0] = \boldsymbol{0}^\tau$ αποτελεί, έστω τετριμμένη, λύση του συστήματος 2.7

$$\boldsymbol{W}^\tau \boldsymbol{d}_p \geq 0 \quad p = 1, \ldots, P \tag{2.7}$$

βλέπει ότι από την αρχή των αξόνων διέρχονται όλα τα υπερεπίπεδα που αποτελούν τα σύνορα των πολυτόπων κάνοντας το πρόβλημα να καταρρέει γύρω από το $\boldsymbol{0}$.

Αυτό που απεικονίζουμε στην πραγματικότητα σ' όλα τα παραδείγματα είναι η προβολή της διάταξης για ένα σταθερό w_0. Άξιο προσοχής είναι επίσης ότι το σχήμα είναι συμμετρικό ως προς τον άξονα του w_0, και διατηρεί την ομοιότητα του σε θετικούς μετασχηματισμούς κλίμακας. Αν δηλαδή το $\boldsymbol{W}$ είναι λύση του προβλήματος τότε και το $\lambda\boldsymbol{W}, \lambda > 0$ αποτελεί λύση. Αφού όπως είδαμε η λύση του ανισοτικού συστήματος είναι στενά συνδεδεμένη με την γεωμετρική έννοια της διάταξης, είναι προφανές ότι και η διάταξη των



Σχήμα 2.3: Το πρόβλημα αυτό έχει προέλθει από την γραμμικοποίηση της εικονιζόμενης παραβολής. Το ζητούμενο είναι να διαχωρίσει τα σημεία που διαχωρίζονται από την οριζόντια γραμμή.

υπερεπιπέδων πρέπει να έχει μείνει όμοια. Επίσης βλέπουμε ότι για οσοδήποτε μεγάλη τιμή του λ, το $\boldsymbol{W}$ παραμένει στο πολύτοπο εφ' όσον είναι λύση. Συνεπώς, το πολύτοπο είναι ανοικτό. Το συμπέρασμα αυτό ισχύει για όλα τα πολύτοπα και όχι μόνο το μηδενικό πολύτοπο. Δηλαδή ο μετασχηματισμός $\boldsymbol{W} \to \lambda \boldsymbol{W}$ δεν αλλάζει την τρέχουσα ταξινόμηση.

Εάν, αντίθετα, επιλέξουμε $\lambda < 0$ τότε ο αριθμός των BW γίνεται συμπληρωματικός του συνολικού αριθμού P των διανυσμάτων που έχουμε προς ταξινόμηση. Αυτό γίνεται εύκολα αντιληπτό αν σκεφτούμε ότι αλλάζοντας με αυτόν το τρόπο το πρόσημο του $\boldsymbol{W}$, ουσιαστικά μετατρέπουμε την τρέχουσα ταξινόμηση των BR σε BW και αντίστροφα. Δηλαδή, ενώ το σχήμα της διάταξης είναι συμμετρικό ως προς την αρχή των αξόνων είναι και συμπληρωματικό ως προς την ταξινόμηση.

Γραμμικοποιημένα προβλήματα Όπως έχει ήδη ειπωθεί αρκετά εμφατικά μέχρι τώρα, είναι δυνατόν μελετώντας προβλήματα γραμμικού προγραμματισμού να αποκτήσουμε πολύτιμη διαίσθηση για την λύση μη γραμμικών. Μια άλλη προσέγγιση είναι να μετατρέψουμε το αρχικά μη γραμμικά διαχωρίσιμο πρόβλημα σε γραμμικά διαχωρίσιμο και να λύσουμε ένα ουσιαστικά γνωστό πρόβλημα.

Ένα τέτοιο παράδειγμα εμφανίζεται στο σχήμα 2.3. Το αρχικό πρόβλημα αποτελείται από μια κατανομή σημείων που απλώνονται κατά το μήκος μιας παραβολής με εξίσωση $y = x^2$. Το ζητούμενο είναι να διαχωρίσουμε με μια ευθεία όλα τα σημεία που έχουν $y > y_0$.

Είναι προφανές ότι μόνο με δεδομένα τις τιμές των x είναι αδύνατο να βρεθεί γραμμική διαχώριση. Στην πραγματικότητα στην λύση ανήκουν τα x εκείνα για τα οποία $x < -x_0$ ή $x > x_0$. Απαιτούνται δηλαδή 2 ευθείες και μια πύλη λογικού ή. Αν όμως γραμμικοποιήσουμε το πρόβλημα αυξάνοντας τον αριθμό των διαστάσεων και θεωρήσουμε ότι οι είσοδοι του προβλήματος είναι το επαυξημένο διάνυσμα $[x, x^2] = [x, y]$ τότε είναι προφανές ότι η διαχώριση μπορεί να είναι γραμμική. Η γραμμικοποίηση μπορεί να περιλαμβάνει και ανώτερους όρους ή και γινόμενα όρων σε πιο πολύπλοκα προβλήματα. Ένα ενδιαφέρον πρόβλημα που το μελέτησαν πρώτοι οι Telfer and Casasent [24] είναι αυτό του κύκλου ($[x, y] \to [x, y, x^2, y^2, xy]$), διότι εξαιτίας της προφανούς του γεωμετρικής του ερμηνείας μπορεί να χρησιμοποιηθεί για της εξαγωγή σμηνών (clustering) [25, 26].

Η διαδικασία αυτή έχει μερικές ενδιαφέρουσες παρενέργειες οι οποίες όμως γίνονται ορατές μόνο στο χώρο των βαρών, όπως αυτός εικονίζεται στη δεξιά γραφική παράσταση του σχήματος 2.3. Η κατανομή δεν είναι πλέον ομογενής εφ' όσον φαίνεται να υπάρχουν μερικά επιμηκυμένα πολύτοπα καθώς και πολλά μικρότερα.

Μια διαδικασία εκμάθησης στοχαστικής στη φύση της, όπως είναι το perceptron, είναι δεδομένο ότι θα έχει μειωμένη απόδοση σε μη ομογενείς κατανομές. Ακόμα χειρότερα, οι ανομοιογένειες, και πιο συγκεκριμένα το γεγονός ότι το $\boldsymbol{\Delta W}$ τις περισσότερες φορές δεν είναι αντιπροσωπευτικό της κίνησης που πρέπει να γίνει και σίγουρα δεν περιέχει καμμιά πληροφορία για το βέλτιστο μέτρο που πρέπει να έχει η προτεινόμενη μετατόπιση οδηγεί σε μια σειρά προβλημάτων χαρακτηριστικών για τα νευρωνικά δίκτυα. Δηλαδή:

- Την εισαγωγή των ευριστικών παραμέτρων ρυθμού εκμάθήσης (learning rate), και της ορμής (momentum) σε μια προσπάθεια πρόχειρου υπολογισμού του μέτρου και της 'σωστής' συνιστώσας του $\boldsymbol{\Delta W}$.

- Την παρατηρούμενη ταλάντωση (oscillation) του τεχνητού νευρωνικού δικτύου στην περιοχή του πολύτοπου στόχου, αν λ.χ. ο ρυθμός εκμάθησης είναι μεγάλος.

- Την πολλές φορές υπερβολικά αργή εκμάθηση σε συγκεκριμένες περιοχές του χώρου (flat minima). Αυτό συμβαίνει όταν το διάνυσμα $\boldsymbol{W}$ βρίσκεται σε κάποια από τα 'επιμηκυμένα' πολύτοπα και ο ρυθμός εκπαίδευσης είναι μικρός. Σε περισσότερες διαστάσεις η προτεινόμενη διεύθυνση από το $\boldsymbol{\Delta W}$ έχει πολύ μικρότερη συνιστώσα προς την 'σωστή' διεύθυνση εξόδου από το πολύτοπο από αυτή που εικονίζεται στο σχήμα 2.3.

Αυτά τα μειονεκτήματα εκμηδενίζουν το ουσιαστικότερο πλεονέκτημα που έχει το perceptron σε σχέση με άλλα νευρωνικά δίκτυα, την απόδειξη σύγκλισης σε πεπερασμένα βήματα. Οι Volper and Hampson απόδειξαν ότι για την τετραγωνική γραμμικοποίηση, παίρνοντας δηλαδή τους πρώτους, τους δεύτερους όρους και τα διπλάσια γινόμενα μόνο, η σύγκλιση του perceptron απαιτεί N^8 επαναλήψεις. Χωρίς να υπάρχει αυστηρή απόδειξη εικάζουμε ότι όσο πιο ανομοιογενείς είναι οι κατανομές, όσο δηλαδή πιο παραμορφωτικές είναι οι συναρτήσεις που χρησιμοποιούνται για την γραμμικοποίηση, τόσο χειροτερεύει η συμπεριφορά του perceptron.

Στα διστρωματικά δίκτυα άλλωστε, το κάτω επίπεδο νευρώνων αποτελεί έναν μη γραμμικό μετασχηματισμό που πρέπει να κάνει το πρόβλημα γραμμικά επιλύσιμο, ώστε να μπορέσει όντως να επιλυθεί από τους νευρώνες του δεύτερου επιπέδου. Είναι φανερή λοιπόν η σημασία που έχει η κατανόηση των γραμμικών προβλημάτων για την λύση άλλων πιο πολύπλοκων προβλημάτων. Πράγματι εφ' όσον απ' ότι φαίνεται μπορούμε να αναπαράγουμε τα βασικά χαρακτηριστικά προβλήματα που συναντούνται στην εκπαίδευση των τεχνητών νευρωνικών δικτύων είναι σαφές ότι το απλό μας μοντέλο περιέχει αρκετή πληροφορία ώστε να κάνει την συμπεριφορά του μη τετριμμένη, και τα συμπεράσματα που εξάγονται από την μελέτη του σημαντικά.

2.1.4 Εκφυλισμένα προβλήματα

Για λόγους που θα γίνουν κατανοητοί αργότερα, θα απαιτήσουμε τα αρχικά διανύσματα εισόδου να είναι γραμμικά ανεξάρτητα ανά $N+1$. Δηλαδή η λύση του συστήματος $\boldsymbol{W}^\tau \boldsymbol{d}_p = 0$ με $N+1$ διανύσματα $\boldsymbol{d}_p$ έχει μόνο μια λύση, την μηδενική $\boldsymbol{W} = 0$. Με όρους γραμμικής άλγεβρας θα λέγαμε ότι ο πίνακας που έχει γραμμές τα $N+1$ $\boldsymbol{d}_p$ διανύσματα είναι τάξης $N+1$ ή πλήρους τάξης. Στην περίπτωση που αυτό δεν ισχύει στο πραγματικό πρόβλημα που έχουμε να λύσουμε, μπορούμε να προσθέσουμε μικρές τυχαίες ποσότητες στις συνιστώσες των διανυσμάτων εκπαίδευσης όπως προτείνεται και από τον Strang [27]. Είναι αξιοσημείωτο ότι τον ίδιο περιορισμό έχει και η μέθοδος simplex και τον αίρει με το ίδιο τέχνασμα.

Παρ' όλα αυτά όμως υπάρχει μια τελική ανωμαλία που δεν έχουμε καταφέρει να αναιρέσουμε. Όλα τα διανύσματα περνούν από την αρχή των αξόνων. Για να λυθεί αυτό το πρόβλημα έχουν προταθεί διαφορετικές τεχνικές. Μια αρκετά δημοφιλής είναι να λυθεί την προβολή του αρχικού προβλήματος σ' ένα επίπεδο που το w_0 είναι σταθερό. Στον εναπομείναντα υπόχωρο των N διαστάσεων, οποιαδήποτε N διανύσματα δεν έχουν κοινά σημεία. Η μέθοδος αυτή έχει το μειονέκτημα ότι χωρίζει το αρχικό πρόβλημα σε 2 ανεξάρτητους, μη επικοινωνούντες υποχώρους ($w_0 < 0$ και $w_0 > 0$), οι οποίοι πρέπει να ερευνηθούν σε ξεχωριστά στάδια για να βρεθεί η λύση.

Μια άλλη τεχνική, την οποία θα υιοθετήσουμε, υπαγορεύει την ελαφρά τροποποίηση του προβλήματος σε

$$\theta(\boldsymbol{W}^\tau \boldsymbol{d}_p - \epsilon_p) = 1, p = 1, 2, \ldots P \tag{2.8}$$

από την εξίσωση 2.3, όπου τα ϵ_p είναι θετικοί, όχι απαραίτητα μικροί, τυχαίοι αριθμοί. Οποιαδήποτε λύση της εξίσωσης 2.8 είναι επίσης λύση και της 2.3. Για να συνεχίσουμε να είμαστε συνεπείς πρέπει να αλλάξουμε ελαφρά και τους ορισμούς των BW και BR χαρακτηρισμών.

- Αν το διάνυσμα $\boldsymbol{d}_p$ ταξινομείται σαν BR από το $\boldsymbol{W}$, τότε $\boldsymbol{W}^\tau \boldsymbol{d}_p - \epsilon_p > 0$.
- Αν το διάνυσμα $\boldsymbol{d}_p$ ταξινομείται σαν BW από το $\boldsymbol{W}$, τότε $\boldsymbol{W}^\tau \boldsymbol{d}_p - \epsilon_p < 0$.

Αυτό που έχουμε κάνει ουσιαστικά είναι να σκληρύνουμε τις απαιτήσεις του προβλήματος. Μια εύλογη ερώτηση είναι πως γνωρίζουμε ότι το νέο δυσκολότερο πρόβλημα είναι επιλύσιμο, και αν όντως είναι, πώς γνωρίζουμε ότι είναι 'ισοδύναμο' με το αρχικό. Η απάντηση και στα δύο ερωτήματα έχει να κάνει με το γεγονός ότι το πρόβλημα παραμένει όμοιο σε μετασχηματισμούς κλίμακας. Τα εισερχόμενα ϵ_p αλλάζουν την δομή και την γεωμετρία του προβλήματος μόνο κοντά στην αρχή των αξόνων, μετατοπίζοντας το τυχόν υπερεπίπεδο που αντιστοιχεί στο $\boldsymbol{d}_p$ κατά ϵ_p και δημιουργώντας κλειστά καινούρια πολύτοπα. Μακριά όμως από την αρχή των αξόνων το πρόβλημα παραμένει αναλλοίωτο αφού για οποιοδήποτε ϵ_p, αν το γινόμενο $\boldsymbol{W}^\tau \boldsymbol{d}_p > 0$ υπάρχει $\lambda > 0$ τέτοιο ώστε $\lambda \boldsymbol{W}^\tau \boldsymbol{d}_p - \epsilon_p > 0$.

2.2 Προηγούμενες εργασίες

2.2.1 Perceptron

Το 1958 ο Rosenblatt [4, 28] υλοποίησε (σε hardware) αυτό που ονόμασε perceptron. Το perceptron είναι η απλούστερη εκπαιδεύσιμη μηχανή και είναι ικανή να λύνει μόνο γραμμικά διαχωρίσιμα προβλήματα. Αυτός είναι και ο λόγος που δέχτηκε και αυστηρή κριτική από τον Minsky [3]. Με την έλευση της ανάστροφης διάδοσης (back propagation) στα τέλη της δεκαετίας του 80, αποδείχτηκε ότι μπορούν να λυθούν και μη γραμμικά προβλήματα και το ενδιαφέρον για τα νευρωνικά δίκτυα αναθερμάνθηκε και παραμένει ζωντανό μέχρι σήμερα.

Για το perceptron υπάρχει απόδειξη σύγκλισης σε πεπερασμένα βήματα. Συγκεκριμένα, το άνω όριο επαναλήψεων που μπορούν να γίνουν δίνεται από

$$Epochs < \frac{N}{D_{min}^2} \tag{2.9}$$

όπου D_{min}^2 σχετίζεται αναλογικά με την ελάχιστη διάσταση του πολυτόπου της περιοχής της λύσης ($BW = 0$).

Παρατηρούμε ότι όσο πιο δύσκολο είναι το πρόβλημα τόσο αυξάνεται η διάρκεια της εκπαίδευσης. Εξαιτίας της στατιστικής φύσης της καταγωγής του το perceptron δεν έχει αυστηρή προσέγγιση της πολυπλοκότητας που απαιτείται. Ακόμα, το perceptron αποδίδει καλύτερα σε όσο το δυνατόν πιο ομογενείς κατανομές, αφού οι ανομοιογένειες όπως αυτές που προκαλούνται από την γραμμικοποίηση ενός προβλήματος φτάνουν το θεώρημα σύγκλισης στα όρια του, όσον αφορά στην πρακτική του χρησιμότητα.

Τέλος, το perceptron υποφέρει από όλα τα κλασσικά προβλήματα των νευρωνικών δικτύων όπως ταλαντεύσεις, περιοχές χαμηλής κλίσης (flat minima), αλλά όχι και από τοπικά ελάχιστα όπως θα δείξουμε. Πράγματι, στην περίπτωση που το πρόβλημα είναι γραμμικά διαχωρίσιμο, η συνάρτηση κόστους του perceptron έχει ένα ελάχιστο, λόγω όμως των μεγάλων χρόνων σύγκλισης στη διεθνή βιβλιογραφία πολλά προβλήματα θεωρήθηκαν μη διαχωρίσιμα, ή με πολλά τοπικά ελάχιστα με εξέχον παράδειγμα το πρόβλημα (sonar.dat) που προσπάθησαν να λύσουν οι Sejnowski και Gorman [29, 30].

2.2.2 Simplex

Η μέθοδος Simplex προτάθηκε το 1947 από τον Dantzig και δημοσιεύτηκε το 1951 [31, 32]. Η Simplex αποτελεί γενικό εργαλείο γραμμικού προγραμματισμού και χρησιμοποιείται ευρύτατα σε πολλούς διαφορετικούς κλάδους. Το τμήμα της Simplex που αναφέρεται διεθνώς στην βιβλιογραφία ως Φάση I είναι το τμήμα της Simplex που μπορεί να λύσει το πρόβλημα της γραμμικής διαχωρισιμότητας όπως το έχουμε θέσει.

Για την Simplex δεν είχε αποδειχτεί πολυωνυμική σύγκλιση, και για πολλά χρόνια η επιστημονική κοινότητα έλπιζε ότι μια απόδειξη θα μπορούσε να βρεθεί, μια και η Simplex έδειχνε ότι χρειαζόταν $O(P+N)^3)$

επαναλήψεις για τα περισσότερα προβλήματα. Τελικά, οι Klee και Minty [33] επινόησαν ως αντιπαράδειγμα έναν κομμένο κύβο στον οποίο η Simplex αναγκάστηκε να επισκεφθεί όλες τις πιθανές λύσεις. Έτσι αποδείχτηκε ότι η Simplex είναι τελικά μια μέθοδος εκθετικής πολυπλοκότητας. Το γιατί πετυχαίνει πολυωνυμική συμπεριφορά στο συντριπτικό ποσοστό των προβλημάτων είναι κάτι που παραμένει μυστήριο και κρυμμένο στο κόσμο της πολυδιάστατης γεωμετρίας.

Το μεγάλο μειονέκτημα της Simplex είναι ότι δεν εκτελεί τη Φάση I στις N διαστάσεις αλλά στις $P+N$. Το γεγονός αυτό είναι κάπως υποβαθμισμένο στην διεθνή βιβλιογραφία διότι η Simplex σχεδιάστηκε και επινοήθηκε για προβλήματα όπου το P είναι συγκρίσιμο με το N και όχι για προβλήματα μεγάλης κλίμακας με πολλούς δεσμούς όπου $P >> N$.

2.2.3 Bobrowski και Niemiro

Το 1984 οι Bobrowski και Niemiro [34] δημοσίευσαν μια εργασία τους σχετικά με την ικανότητα ενός προτεινόμενου αλγόριθμου να αναγνωρίζει την περίπτωση της μη γραμμικής διαχωρισιμότητας. Ο αλγόριθμος του Bobrowski ανήκει στην ευρύτερη κατηγορία των αλγορίθμων ενεργών δεσμών (active sets) ή στην κατηγορία των επιτρεπτών διευθύνσεων (feasible direction).

Η γενική ιδέα του αλγόριθμου είναι, όπως θα δούμε και στη δική μας περίπτωση, η χρησιμοποίηση του κανόνα του perceptron για τη μείωση του αριθμού των λάθος ταξινομημένων διανυσμάτων. Ο Bobrowski απέδειξε ότι, όπως και η Simplex, η μέθοδος του τερματίζει σε πεπερασμένο αριθμό βημάτων. Ο Bobrowski όμως αντίθετα από εμάς αποφεύγει να λύσει ένα δύσκολο τετραγωνικό πρόβλημα με αποτέλεσμα να μην μπορεί να ακολουθήσει τη βέλτιστη επιτρεπτή διεύθυνση. Αντίθετα ο αλγόριθμος του Bobrowski κινείται πάνω στις κορυφές των πολυτόπων ακολουθώντας τις ακμές τους.

Ίσως αυτή η περιγραφή να θυμίζει έντονα την Simplex. Πράγματι αν αγνοήσουμε τις διαφορές που έχει στον φορμαλισμό το πρόγραμμα γραμμικού προγραμματισμού, με το πρόβλημα της γραμμικής διαχωρισιμότητας, τότε μπορούμε να δούμε ότι η μέθοδος του Bobrowski είναι ισοδύναμη με την Simplex στη φάση 2, με την έννοια ότι παράγει την ίδια ακολουθία σημείων προς την τελική λύση.

Η παραπάνω πρόταση δεν είναι μειωτική για την ερευνητική συνεισφορά του Bobrowski. Είναι φανερό ότι τον Bobrowski δεν τον απασχολούσε η φάση 2 του αλγορίθμου, δηλαδή η βελτιστοποίηση μιας γραμμικής συνάρτησης εντός επιτρεπτής περιοχής. Το πρόβλημα που προσπαθούσε να λύσει ο Bobrowski ήταν η εύρεση ενός σημείου που να ανήκει στην επιτρεπτή περιοχή, δηλαδή να ικανοποιεί όλες τις απαραίτητες γραμμικές ανισώσεις (φάση 1).

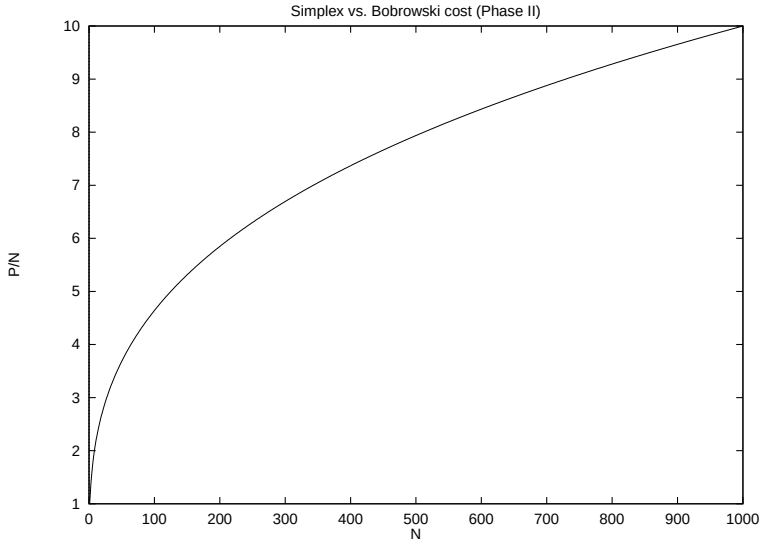
Για να το καταφέρει αυτό χρησιμοποίησε σαν διάνυσμα οδηγό, το διάνυσμα που προκύπτει από τον κανόνα ανανέωσης του off line perceptron. Κάτω απ' αυτό το πρίσμα η μέθοδος του Bobrowski μπορεί να ιδωθεί και σαν μία μέθοδος που χρησιμοποιεί συναρτήσεις τιμωρούς (penalty functions) όπου η τιμωρία ορίζεται για κάθε λάθος ταξινομημένο διάνυσμα. Έτσι πέτυχε να έχει έναν αλγόριθμο με κόστος $O(N^4 + P)$, αποφεύγοντας, την ομολογουμένως άκομψη, τεχνική της Simplex που με την εισαγωγή βοηθητικών μεταβλητών αυξάνει, υπερβολικά για προβλήματα γραμμικής διαχωρισιμότητας, τη διάσταση του προβλήματος.

Επειδή ακριβώς η συνάρτηση κόστους της Simplex, με την εισαγωγή βοηθητικών μεταβλητών, διαφέρει από αυτή του Bobrowski είναι αδύνατο να συγκρίνουμε σε θεωρητική βάση τους δύο αλγόριθμους στο πρόβλημα της γραμμικής διαχωρισιμότητας (Φάση I, για την Simplex). Αντίθετα κάτι τέτοιο είναι δυνατό για προβλήματα της δεύτερης φάσης, σε προβλήματα δηλαδή βελτιστοποίησης γραμμικών συναρτήσεων κάτω από γραμμικούς δεσμούς.

Αφού όπως είδαμε είναι ισοδύναμες σαν μέθοδοι, αυτό σημαίνει ότι θα κάνουν ακριβώς τον ίδιο αριθμό επαναλήψεων, αλλάζοντας με ακριβώς τον ίδιο τρόπο τον εσωτερικό τους πίνακα ενεργών δεσμών[1]. Όμως ενώ η Simplex απαιτεί $O(P^3 + NP)$ πράξεις για κάθε επανάληψη, η μέθοδος του Bobrowski χρειάζεται $O(N^4 + P)$. Αν θέσουμε $P = Nx$ και πάρουμε την διαφορά προκύπτει το πολυώνυμο:

$$N^2x^3 + (N-1)x - N^3 \tag{2.10}$$

[1] ταμπλό για την Simplex, λίστα για τον Bobrowski



Σχήμα 2.4: Η ρίζα του πολυώνυμου της διαφοράς των δύο μεθόδων. Αν ο λόγος του P/N είναι μικρότερος από την τιμή του x για το αντίστοιχο N, τότε η simplex είναι η πιο συμφέρουσα επιλογή.

Στην περίπτωση που το πολυώνυμο έχει θετική τιμή τότε η μέθοδος Simplex έχει μεγαλύτερο κόστος από αυτή του Bobrowski. Δεδομένου ότι γύρω από την πραγματική ρίζα το πολυώνυμο είναι αύξουσα συνάρτηση, υπολογίζοντας τη ρίζα μπορούμε να βρούμε το όριο που η μια επιλογή είναι συμφέρουσα σε σχέση με την άλλη. Αν υπολογίσουμε τη ρίζα του πολυωνύμου και κάνοντας τη γραφική παράσταση του x ως προς N μπορούμε να δούμε την περιοχή στην οποία είναι συμφέρουσα η μία μέθοδος έναντι της άλλης. Όπως βλέπουμε και από το σχήμα 2.4 το x αυξάνει σχεδόν με την κυβική ρίζα της διάστασης του προβλήματος. Έτσι σε προβλήματα των 1000 διαστάσεων, αν ο αριθμός των δεσμών είναι μέχρι 10000, η Simplex αποτελεί συμφέρουσα επιλογή, ακριβώς όπως θα περίμενε κανείς γνωρίζοντας τους στόχους των δύο αλγόριθμων.

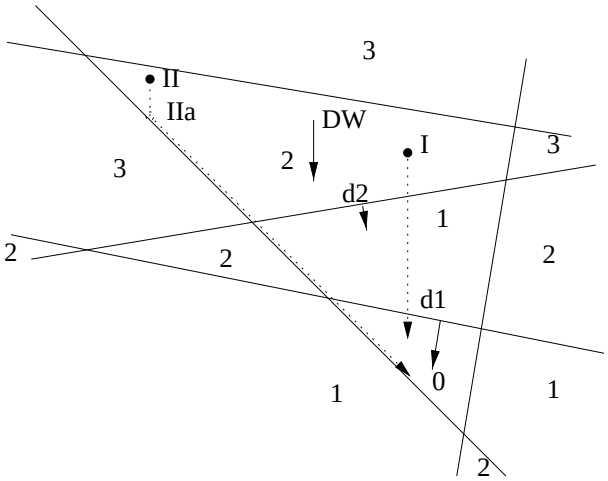
Να υπενθυμίσουμε ότι ακριβώς το ίδιο πλεονέκτημα παρουσιάζει η μέθοδος του Bobrowski έναντι στη Simplex και σε προβλήματα της πρώτης φάσης (ταξινόμηση). Τα πειραματικά αποτελέσματα δείχνουν ότι η μέθοδος του Bobrowski έχει σαφές πλεονέκτημα σε προβλήματα με πολλούς δεσμούς, αλλά εκεί η σύγκριση δεν είναι ούτε τόσο εύκολη ούτε τόσο ακριβής διότι οι αλγόριθμοι επιλέγουν διαφορετικό μονοπάτι επίλυσης του προβλήματος.

2.3 Μια καινούρια προσέγγιση

2.3.1 Προτεινόμενη Στρατηγική

Θα ξεκινήσουμε τη διαδικασία εκπαίδευσης από ένα τυχαίο διάνυσμα $\boldsymbol{W}$ το οποίο θ' ανήκει στο εσωτερικό κάποιου πολυτόπου. Ο αντικειμενικός μας στόχος είναι να ελαχιστοποιήσουμε την συνάρτηση κόστους E_{SE} της εξίσωσης 2.6, να μηδενίσουμε δηλαδή των αριθμό των BW. Με αυτόν τον σκοπό θα αναπτύξουμε μια στρατηγική η οποία θα μετακινεί το $\boldsymbol{W}$ σε πολύτοπα με μικρότερη E_{SE}. Για να επιτύχουμε τον στόχο αυτό θα μετακινούμε το $\boldsymbol{W}$ κατά μήκος διαδοχικών ερευνητικών διευθύνσεων (search directions), που χαρακτηρίζονται από το διάνυσμα $\boldsymbol{P}$.

Για να μπορέσουμε να υπολογίσουμε το διάνυσμα $\boldsymbol{P}$ θα χρησιμοποιήσουμε πληροφορία που προέρχεται από το διάνυσμα $\boldsymbol{\Delta W}$ που είναι η παράγωγος της συνάρτησης κόστους του perceptron E όπως περιγράφεται από την εξίσωση 2.4. Το κατά πόσο μια τέτοια δυϊκή στρατηγική είναι δόκιμη, δηλαδή το να χρησιμοποιή-



Σχήμα 2.5: Όταν ο αλγόριθμος ξεκινάει από την θέση I πέφτει στην κατάσταση επιταχυμένης κίνησης με την οποία καταλήγει κατευθείαν στην λύση. Στην περίπτωση που είναι στην θέση II, βλέπουμε ότι χρειάζεται μια επανάληψη επιπλέον για να βρει ένα πολύ καλό δρόμο που τον οδηγεί άμεσα στην λύση.

σουμε την παράγωγο μιας συνάρτησης για να ελαχιστοποιήσουμε μια άλλη, είναι το θέμα της συζήτησης ολόκληρου του κεφαλαίου από εδώ και στο εξής αφού θα μελετήσουμε τις ιδιότητες και τη σύγκλιση σε βάθος.

Για την πρώτη επανάληψη του αλγορίθμου διαλέγουμε $\boldsymbol{\Delta W} = \boldsymbol{P}$ και ανανεώνουμε το $\boldsymbol{W}$ σύμφωνα με τον κανόνα:

$$\boldsymbol{W}_{new} = \boldsymbol{W} + \eta \boldsymbol{P} \tag{2.11}$$

όπου το η είναι ο συντελεστής εκπαίδευσης. Όπως μπορεί να δει κανείς ο κανόνας εκπαίδευσης είναι ταυτόσημος με του perceptron. Στην πραγματικότητα όμως το διάνυσμα $\boldsymbol{P}$ δεν πρόκειται να παραμείνει ίδιο καθ' όλη την διάρκεια της εκπαίδευσης, και ο ρυθμός εκπαίδευσης η προκύπτει από την απαίτηση να διασχίσουμε όσο το δυνατόν περισσότερα BW, χωρίς όμως να διασχίσουμε BR. Ακολουθώντας αυτήν την στρατηγική μπορούμε να πετύχουμε την μέγιστη δυνατή μείωση της συνάρτησης κόστους E.

Σ' αυτό το σημείο θα πρέπει να θυμίσουμε ότι η ταξινόμηση για το τυχόν διάνυσμα εισόδου $\boldsymbol{d}_p$ αλλάζει πάνω στο σύνορο όπως αυτό γράφεται από το υπερεπίπεδο που είναι κάθετο σ' αυτό. Δηλαδή $\boldsymbol{W}_{new}^{\tau}\boldsymbol{d}_p - \epsilon_p = 0 \Rightarrow \eta_p = -(\boldsymbol{W}^{\tau}\boldsymbol{d}_p - \epsilon_p)/(\boldsymbol{P}^{\tau}\boldsymbol{d}_p)$. Τώρα μπορούμε να διακρίνουμε τις ακόλουθες περιπτώσεις ανάλογα με το αν συναντάμε BR ή BW:

1. Ας υποθέσουμε ότι υπάρχει ένα τουλάχιστον BR διάνυσμα κατά την διεύθυνση αναζήτησης $\boldsymbol{P}$. Έστω η_R το μικρότερο θετικό η_p που αντιστοιχεί στο κοντινότερο, κατά την διεύθυνση $\boldsymbol{P}$, BR υπερεπίπεδο.

   (α') Αν συναντούμε BW υπερεπίπεδο πριν φτάσουμε στο BR, δηλαδή υπάρχει υπερεπίπεδο p που να χαρακτηρίζεται σαν BW και $\eta_p < \eta_R$, θεωρούμε η_W το μεγαλύτερο από τα πιθανά (BW) η_p. Είναι τώρα σαφές ότι τη μέγιστη μείωση της E_{SE} πρέπει να επιλέξουμε ένα οποιοδήποτε η τέτοιο ώστε $\eta_W < \eta < \eta_R$. Στην πράξη πάντοτε διαλέγουμε $\eta = (\eta_W + \eta_R)/2$. Αυτός ο τύπος κίνησης θα καλείται από εδώ και στο εξής επιταχυμένη κίνηση ('fast moving') διότι προκαλεί συνήθως εντυπωσιακές πτώσεις στη συνάρτηση κόστους, παρ' όλο που υπάρχουν και περιπτώσεις που η κίνηση αυτή διορθώνει την ταξινόμηση ενός και μόνο διανύσματος. Ένα παράδειγμα επιταχυμένης κίνησης φαίνεται στο σχήμα 2.5 όταν το αρχικό διάνυσμα θέσης $\boldsymbol{W}$ είναι στη θέση I.

   (β') Αν τώρα δεν υπάρχει BW υπερεπίπεδο πριν από το BR, δηλαδή όλα τα η_p των BW υπερεπιπέδων είναι μεγαλύτερα από το η_R, τότε δεν μπορούμε να μειώσουμε των αριθμό των BW (E_{SE}). Αυτό

που μπορούμε να κάνουμε είναι να διατηρήσουμε την E_{SE} σταθερή και να κινηθούμε όσο πιο κοντά μπορούμε στο BR υπερεπίπεδο. Το υπερεπίπεδο προστίθεται σε μια λίστα ενεργών δεσμών η οποία επιτελεί διπλό στόχο. Προστατεύει τον δεσμό, διότι τώρα πρόκειται πλέον για δεσμό ($\boldsymbol{W}^\tau \boldsymbol{d}_p - \epsilon_p = 0$), από αριθμητικά λάθη και σφάλματα στρογγύλευσης, και συνεισφέρει στον υπολογισμό της νέας διεύθυνσης έρευνας $\boldsymbol{P}$, αφού το $\boldsymbol{P}$ πρέπει να είναι τουλάχιστον παράλληλο με το δεσμό για να μην τον παραβιάσει. Η διαδικασία αυτή κατά την οποία ερχόμαστε σε μηδενική απόσταση από το BR υπερεπίπεδο θα καλείται κοντινή προσέγγιση (moving near), και εικονίζεται στο σχήμα 2.5 όταν το αρχικό διάνυσμα θέσης $\boldsymbol{W}$ είναι στη θέση II.

Σ' αυτό το σημείο μπορούμε να αποκαλύψουμε, έστω μερικώς, γιατί επιμείναμε τόσο στο να καταργήσουμε τους εκφυλισμούς στο κεφάλαιο 2.1.4. Αν υποθέσουμε ότι περισσότερα από $N+1$ υπερεπίπεδα επιτρεπόταν να είχαν κοινά σημεία, οι διευθύνσεις έρευνας που υπολογίζουμε κάθε φορά θα μπορούσαν να τέμνουν παραπάνω από ένα υπερεπίπεδο στην ίδια απόσταση, έχοντας δηλαδή την ίδια τιμή για το η, κάνοντας έτσι την κοντινή προσέγγιση προβληματική και ασταθή.

2. Αν τώρα υποθέσουμε ότι δεν υπάρχει BR μπροστά από την κατεύθυνση:

   (α') Αν υπάρχουν BW, το μόνο που έχουμε να κάνουμε είναι να διασχίσουμε όλα τα BW για να λύσουμε το πρόβλημα (Last Fast Moving). Έτσι επιλέγουμε το η μεγαλύτερο από όλα τα η_p.

   (β') Αν δεν υπάρχουν ούτε BW ούτε BR στην διεύθυνση έρευνας $\boldsymbol{P}$, τότε $\boldsymbol{P} = 0$ και το πρόβλημα έχει λυθεί οπότε και ο αλγόριθμος τερματίζει.

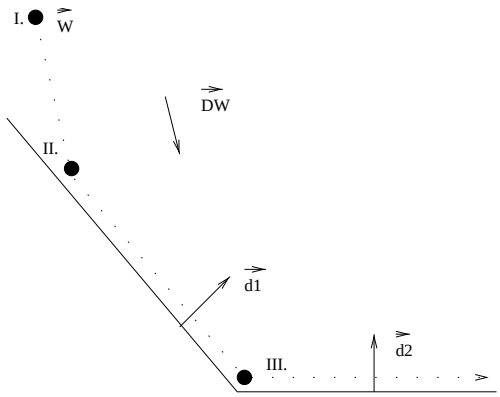
Η πρώτη επανάληψη του αλγορίθμου έχει τώρα τελειώσει. Η συνέχεια σκιαγραφείται ως εξής:

1. Αν στην προηγούμενη επανάληψη η κίνηση ήταν κοντινή προσέγγιση, τότε το διάνυσμα θέσης $\boldsymbol{W}$ βρίσκεται ακόμα μέσα στο πολύτοπο που ήταν προηγουμένως. Προφανώς η E_{SE} δεν έχει μειωθεί από την προηγούμενη επανάληψη. Σ' αυτή την περίπτωση επιλέγουμε μια νέα κατεύθυνση έρευνας, που ελπίζουμε ότι θα μας οδηγήσει σε BW, ώστε να μπορέσουμε να εφαρμόσουμε τον κανόνα της επιταχυμένης κίνησης κατά την διάρκεια αυτής της επανάληψης.

   Για να διαλέξουμε τη νέα διεύθυνση $\boldsymbol{P}$, χρησιμοποιούμε την συνάρτηση κόστους του perceptron E. Επιθυμούμε να βρούμε $\boldsymbol{W}_{new}$ μέσα στο τρέχον πολύτοπο τέτοιο, ώστε $E(\boldsymbol{W}_{new}) < E(\boldsymbol{W})$. Αφού στην πραγματικότητα θέλουμε το E να μειώνεται με τον ταχύτερο δυνατό ρυθμό, ουσιαστικά επιλέγουμε την διεύθυνση κίνησης της μέγιστης πτώσης (steepest descent) που ανήκει όμως στο τρέχον πολύτοπο (steepest feasible search direction).

   Αν η λίστα των ενεργών δεσμών περιλαμβάνει K υπερεπίπεδα, η επιθυμητή διεύθυνση $\boldsymbol{P}$ πρέπει να είναι επιτρεπτή ($\boldsymbol{P}^\tau \boldsymbol{d}_p \geq 0$) και να σχηματίζει την ελάχιστη γωνία ϕ με την παράγωγο $\boldsymbol{\Delta W}$ της E. Η εύρεση αυτής της διευθύνσεως είναι ένα πρόβλημα τετραγωνικού προγραμματισμού και το οποίο θα συζητηθεί εκτενέστατα στο κεφάλαιο 3. Σ' αυτό το σημείο θα θεωρήσουμε τη λύση αυτού του προβλήματος δεδομένη και θα πούμε απλώς ότι το τελικό διάνυσμα $\boldsymbol{P}$ είναι παράλληλο με κάποια υπερεπίπεδα της αρχικής λίστας ($\boldsymbol{P}^\tau \boldsymbol{d}_p = 0$) ενώ θετικό με κάποια άλλα ($\boldsymbol{P}^\tau \boldsymbol{d}_p > 0$), και έτσι πρέπει να επανυπολογιστεί η λίστα των ενεργών δεσμών. Από τη στιγμή που έχουμε βρει το $\boldsymbol{P}$, μπορούμε να προχωρήσουμε και να μετακινήσουμε το διάνυσμα θέσης μας $\boldsymbol{W}$ όπως περιγράφεται στην εξίσωση 2.11 με το η να καθορίζεται από τις περιπτώσεις 1 και 2 που ήδη έχουν περιγραφεί παραπάνω.

2. Αν τώρα δεχτούμε ότι η κίνηση στην προηγούμενη επανάληψη ήταν επιταχυμένη τότε το $\boldsymbol{W}$ βρίσκεται σ' ένα καινούριο πολύτοπο με χαμηλότερη E_{SE}. Πρέπει τώρα να βρούμε την μέγιστη επιτρεπτή πτώση για το συγκεκριμένο πολύτοπο. Η διαδικασία αυτή περιλαμβάνει επανυπολογισμό του $\boldsymbol{\Delta W}$ για το καινούριο πολύτοπο και επανεύρεση του $\boldsymbol{P}$ ώστε να απαλλαχθούμε και από τυχόν άχρηστους δεσμούς που είχαμε εισάγει στη λίστα σε προηγούμενα πολύτοπα. Για ακόμα μια φορά υπολογίζουμε το νέο $\boldsymbol{W}$ σύμφωνα με την εξίσωση 2.11 ενώ το η καθορίζεται από τις περιπτώσεις 1 και 2 που ήδη συζητήσαμε παραπάνω.



Σχήμα 2.6: Όπως κινείται ο αλγόριθμος είναι φανερό ότι πρέπει να εγκαταλείψει το ένα υπερεπίπεδο για να κινηθεί στο άλλο. Αν δεν το κάνει και προσπαθήσει να κινηθεί σεβόμενος και τους δύο δεσμούς τότε θα παγιδευτεί στην γωνία.

Μέχρι τώρα, στην περιγραφή του αλγόριθμου, αποφύγαμε επιμελώς να θίξουμε δύο μάλλον επώδυνα θέματα:

1. Είναι φανερό ότι με την διαδικασία αυτή κυριολεκτικά 'ελπίζουμε' ότι θα τύχει μια περίπτωση 'Fast Moving' για να επιτευχθεί η έξοδός μας από το πολύτοπο.

2. Είδαμε μόνο τον τρόπο με τον οποίον εισάγονται τα υπερεπίπεδα στη λίστα. Αν υποθέσουμε ότι δεν υπάρχει μηχανισμός για να βγαίνουν από την λίστα, τότε μόλις ο αριθμός των ενεργών δεσμών φτάσει να είναι $N+1$, τότε θα έχουμε $\boldsymbol{P} = 0$ [2] και ο αλγόριθμος θα τερματίσει πρόωρα. Ένα τέτοιο παράδειγμα φαίνεται στο σχήμα 2.6.

Για το πρώτο θέμα θα κάνουμε μια εκτενή συζήτηση στη επόμενη παράγραφο, όπου θα δείξουμε όχι μόνο ότι οι 'ελπίδες' μας είναι βάσιμες αλλά και ότι η διαδικασία τελειώνει σε πεπερασμένα βήματα (έξοδος από το πολύτοπο), ενώ υπάρχουν και ισχυρές ενδείξεις για πολυωνυμική πολυπλοκότητα.

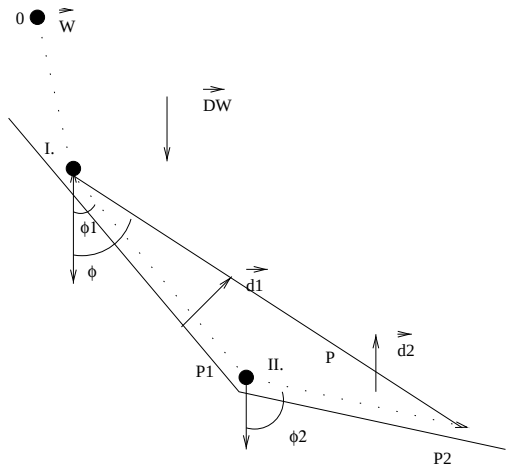
Το δεύτερο θέμα είναι πολύ μεγαλύτερο και σοβαρότερο. Υπάρχει πρόταση για την επίλυση του η οποία συζητείται στο κεφάλαιο 3. Το υποπρόβλημα αυτό πάντως είναι τετραγωνικού προγραμματισμού από την φύση του και το μηχανικό ανάλογο είναι η κίνηση υλικού σημείου που βρίσκεται σε γωνία πολλών διαστάσεων υπό την επίδραση της βαρύτητας ($\boldsymbol{\Delta W}$). Από εδώ και στο εξής θα υποθέσουμε πάντως ότι ο αλγόριθμος ακολουθεί την βέλτιστη διαδρομή, επιλέγοντας τους κατάλληλους κάθε φορά ενεργούς δεσμούς.

Μπορούμε όμως από τώρα να παρατηρήσουμε ότι αυτός ο αλγόριθμος παρουσιάζει έναν αβίαστο και φυσικό τρόπο να κινούμαστε κατά την διεύθυνση μέγιστη πτώσης στον χώρο βαρών του perceptron. Μ' αυτή την έννοια είναι στενά συνδεδεμένος με τον αλγόριθμο των Bobrowski και Niemiro [34], ο οποίος όμως επιλέγει την ακμή του πολυτόπου που οδηγεί σε μεγαλύτερη πτώση και όχι την μέγιστη επιτρεπτή (steepest feasible) όπως ο προτεινόμενος.

Ο προτεινόμενος αλγόριθμος ανήκει στην γενικότερη κλάση των αλγορίθμων επιτρεπτών διευθύνσεων (feasible direction), όπως και η Simplex του Dantzig. Η ερευνητική συνεισφορά της παρούσας εργασίας είναι ότι:

1. Αποδεικνύει τη σύγκλιση σε πεπερασμένα βήματα για αυτόν τον κανόνα του τροποποιημένου off line perceptron.

[2] Ένα διάνυσμα έχει τόσες συνιστώσες όσες και ο χώρος που ανήκει.



Σχήμα 2.7: Η γωνία ϕ μειώνεται διαδοχικά καθώς ο αλγόριθμος εξερευνά το πολύτοπο.

2. Αναλύει το μεγάλο γραμμικό πρόβλημα γραμμικού προγραμματισμού $P \times N$ σε πολλά μικρότερα, τετραγωνικού προγραμματισμού όμως, τάξης $K \times (N+1)$ όπου το K είναι ο αριθμός των ενεργών δεσμών μέσα στη λίστα και μπορεί να γίνει μέχρι $N+1$ το πολύ.

3. Το τετραγωνικό υποπρόβλημα που βρίσκει την βέλτιστη επιτρεπτή διεύθυνση $\boldsymbol{P}$ πρώτη φορά διατυπώνεται για τη λύση του γραμμικού προβλήματος. Το αν αξίζει να λύσει κανείς μυριάδες τετραγωνικά προβλήματα για μπορέσει να λύσει το γραμμικό συζητείται στο κεφάλαιο των πειραματικών αποτελεσμάτων (κεφ. 4) όπου και γίνεται εκτενής σύγκριση και ανάλυση όλων των μεθόδων.

4. Η λύση που προτείνεται για το τετραγωνικό υποπρόβλημα είναι κι' αυτή πρωτότυπη.

2.3.2 Απόδειξη Σύγκλισης

Σ' αυτό το κεφάλαιο θα αποδείξουμε τις ακόλουθες προτάσεις:

- Ο προτεινόμενος αλγόριθμος τερματίζει πάντα μετά από έναν πεπερασμένο αριθμό επαναλήψεων.
- Με τον τερματισμό, ο προτεινόμενος αλγόριθμος ταξινομεί σωστά όλα τα διανύσματα εισόδου στα γραμμικά διαχωρίσιμα προβλήματα.

Το πλήθος των βημάτων είναι πεπερασμένο

Το λήμμα που παρουσιάζουμε παρακάτω εισάγει μια ταξινόμηση στα σημεία που επισκέπτεται ο αλγόριθμος μέσα σ' ένα πολύτοπο. Η ταξινόμηση σχετίζεται με τις γωνίες που σχηματίζονται ανάμεσα στις διαδοχικές γωνίες $\boldsymbol{P}$, και την τοπική παράγωγο, σταθερή γι όλο το εξεταζόμενο πολύτοπο $\boldsymbol{\Delta W}$. Αυτή η ταξινόμηση θα μας βοηθήσει αργότερα να αποδείξουμε το κεντρικό θεώρημα του κεφαλαίου.

Λήμμα 1 *Έστω σημεία I και II που ανήκουν σ' ένα συγκεκριμένο πολύτοπο, τα οποία απέχουν μεταξύ τους μια επανάληψη του προτεινόμενου αλγόριθμου. Αν ϕ_1 και ϕ_2 είναι οι γωνίες που σχηματίζονται από τις βέλτιστες επιτρεπτές διευθύνσεις $\boldsymbol{P}_1$ και $\boldsymbol{P}_2$ με το $\boldsymbol{\Delta W}$ τότε $\phi_1 < \phi_2$ (σχήμα 2.7).*

Απόδειξη Ας υποθέσουμε ότι $\phi_2 \leq \phi_1$, τότε όμως θα ισχύει ότι $\cos(\phi_2) \geq \cos(\phi_1) \Rightarrow \cos(\phi_2) = \cos(\phi_1) + e$ όπου $0 \leq e < 1$.

Έστω τώρα διάνυσμα $\boldsymbol{P}$ τέτοιο, ώστε $\boldsymbol{P} = \boldsymbol{P}_1 + \boldsymbol{P}_2$, όπως φαίνεται στο σχήμα 2.7.

$$\cos(\phi) = \frac{\boldsymbol{P}\cdot\boldsymbol{\Delta W}}{|\boldsymbol{P}||\boldsymbol{\Delta W}|} \Rightarrow \tag{2.12}$$

$$\cos(\phi) = \frac{\boldsymbol{P}_1\cdot\boldsymbol{\Delta W}+\boldsymbol{P}_2\cdot\boldsymbol{\Delta W}}{|\boldsymbol{P}_1+\boldsymbol{P}_2||\boldsymbol{\Delta W}|} \Rightarrow \tag{2.13}$$

$$\cos(\phi) = \frac{|\boldsymbol{P}_1|\cos(\phi_1)+|\boldsymbol{P}_2|\cos(\phi_2)}{|\boldsymbol{P}_1+\boldsymbol{P}_2|} \Rightarrow \tag{2.14}$$

$$\cos(\phi) = \frac{\cos(\phi_1)(|\boldsymbol{P}_1|+|\boldsymbol{P}_2|)+e|\boldsymbol{P}_2|}{|\boldsymbol{P}_1+\boldsymbol{P}_2|} \Rightarrow \tag{2.15}$$

$$\cos(\phi) \geq \frac{\cos(\phi_1)(|\boldsymbol{P}_1|+|\boldsymbol{P}_2|)}{|\boldsymbol{P}_1+\boldsymbol{P}_2|} \Rightarrow \tag{2.16}$$

Χρησιμοποιώντας την τριγωνική ανισότητα $|\boldsymbol{P}_1|+|\boldsymbol{P}_2| > |\boldsymbol{P}_1+\boldsymbol{P}_2|$ μπορούμε να ξαναγράψουμε την εξίσωση 2.16 ως εξής:

$$\cos(\phi) > \cos(\phi_1) \Rightarrow \phi < \phi_1 \tag{2.17}$$

το οποίο είναι σαφώς άτοπο. Αυτό συμβαίνει διότι το πολύτοπο είναι κυρτό (convex) και το $\boldsymbol{P}$ ανήκει μέσα σε αυτό (αφού $\boldsymbol{P} = \boldsymbol{P}_1 + \boldsymbol{P}_2$), σε συνδυασμό με το γεγονός ότι υπάρχει αποδεκτή διεύθυνση ($\boldsymbol{P}$) που οδηγεί σε μεγαλύτερες μειώσεις αφού $\phi < \phi_1$. Το άτοπο προκύπτει από την αρχική μας απαίτηση ότι η γωνία ϕ_1 ήταν η βέλτιστη δυνατή επιλογή. Έτσι πρέπει να αποδεχτούμε το γεγονός ότι σε διαδοχικές κινήσεις μέσα στο πολύτοπο έχουμε $\phi_2 > \phi_1$.

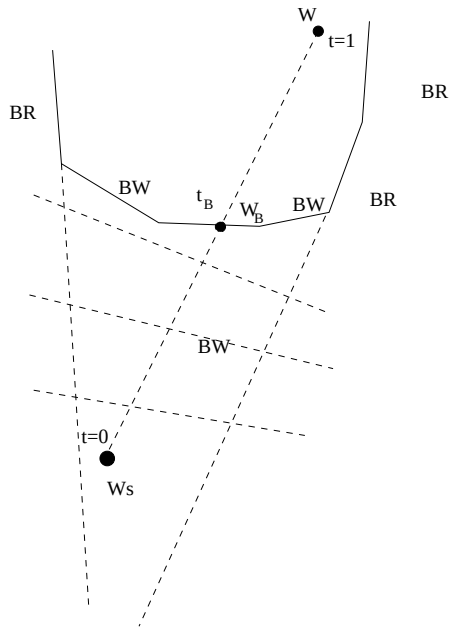
Θεώρημα 1 *Ο αλγόριθμος πάντοτε τερματίζει σε πεπερασμένα βήματα.*

Απόδειξη Ο αριθμός των διαφορετικών πολυτόπων στα οποία υπεισέρχεται ο αλγόριθμος είναι το πολύ ίσος με τον αρχικό αριθμό των BW ο οποίος προφανώς δεν μπορεί να ξεπερνά τον συνολικό αριθμό P των δεσμών που συνιστούν το πρόβλημα. Έπεται λοιπόν, ότι για να αποδείξουμε το θεώρημα αρκεί να δείξουμε ότι χρειάζεται μόνο ένας πεπερασμένος αριθμός από επαναλήψεις ανά πολύτοπο. Η γωνία ϕ που σχηματίζεται από τη βέλτιστη επιτρεπτή διεύθυνση σ' ένα σημείο του πολυτόπου και το $\boldsymbol{\Delta W}$, καθορίζεται από τα διανύσματα που αποτελούν την λίστα των ενεργών δεσμών στο δοθέν σημείο. Αφού υπάρχουν το πολύ P υπερεπίπεδα σχηματίζοντας τα σύνορα του κάθε πολυτόπου, υπάρχει ένας πεπερασμένος αριθμός από συνδυασμούς ενεργών δεσμών και συνεπώς ένας πεπερασμένος αριθμός από πιθανές τιμές του ϕ. Όμως σύμφωνα με το λήμμα 1, η γωνία ϕ αυξάνεται μονότονα. Έτσι ο αριθμός των επαναλήψεων που πρόκειται να ξοδέψει ο αλγόριθμος σ' ένα πολύτοπο είναι φραγμένος από τον αριθμό των επιτρεπτών τιμών του ϕ.

Γραμμικά διαχωρίσιμα προβλήματα

Για να αποδείξουμε ότι ο αλγόριθμος θα βρίσκει πάντοτε μια αποδεκτή λύση σε γραμμικά διαχωρίσιμα προβλήματα, πρέπει πρώτα να βεβαιωθούμε ότι θα μπορεί να διαφύγει από το τυχόν τρέχον πολύτοπο. Συγκεκριμένα, πρέπει να δείξουμε ότι δεν μπορεί να τερματίσει μέσα στο μη μηδενικό πολύτοπο όταν το πρόβλημα είναι γραμμικά διαχωρίσιμο. Αυτό δείχνεται με το λήμμα 3, το οποίο αποδεικνύει ότι στο πιο ακραίο σημείο του πολυτόπου κατά την διεύθυνση $\boldsymbol{\Delta W}$ εφάπτεται ένα τουλάχιστον BW υπερεπίπεδο και συνεπώς μπορούμε να διαφύγουμε απ' αυτό. Για να γίνει όμως αυτό πρώτα θα δείξουμε ότι για κάθε σημείο του πολυτόπου υπάρχει ένα άλλο με λιγότερη ενέργεια (πιο ακραίο ως προς $\boldsymbol{\Delta W}$), το οποίο εφάπτεται στο σύνορο του πολυτόπου και ανήκει σε ένα BW υπερεπίπεδο. Έτσι το λήμμα 2 έρχεται φυσιολογικά πριν το λήμμα 3 για να δείξει πως μια τέτοια κατασκευή είναι δυνατή.

Λήμμα 2 *Έστω ότι το αρχικό πρόβλημα 2.3 είναι γραμμικά διαχωρίσιμο, έτσι ώστε το σύστημα ανισώσεων να έχει τουλάχιστον μία λύση $\boldsymbol{W}_s$. Ας θεωρήσουμε ότι το διάνυσμα $\boldsymbol{W}$ ανήκει στο εσωτερικό ενός πολυτόπου $\mathcal{R}$ το οποίο χαρακτηρίζεται βέβαια από τον αριθμό (θετικό) των BW. Τότε υπάρχει υπερεπίπεδο $\boldsymbol{d}_B$ το οποίο*



Σχήμα 2.8: Το ευθύγραμμο τμήμα που ενώνει ένα οποιοδήποτε σημείο $\boldsymbol{W}$ που ανήκει στο πολύτοπο $\mathcal{R}$, μ' ένα οποιοδήποτε σημείο $\boldsymbol{W}_s$ που αποτελεί λύση του αρχικού προβλήματος, δεν τέμνει BR.

ανήκει στο σύνορο του $\mathcal{R}$ και είναι ταξινομημένο σαν BW από το $\boldsymbol{W}$, και ένα διάνυσμα $\boldsymbol{W}_B$ για το οποίο ισχύει:

$$\boldsymbol{W}_B^\tau \boldsymbol{d}_B = \epsilon_B \quad \textit{και} \quad \boldsymbol{W}_B = \boldsymbol{W}_s + t_B(\boldsymbol{W} - \boldsymbol{W}_s) \quad \textit{με} \quad 0 < t_B < 1 \tag{2.18}$$

Απόδειξη Η ευθεία που ενώνει τα $\boldsymbol{W}$, $\boldsymbol{W}_s$ έχει παραμετρική εξίσωση:

$$\boldsymbol{W}_t = \boldsymbol{W}_s + t(\boldsymbol{W} - \boldsymbol{W}_s). \tag{2.19}$$

Η εικόνα 2.8 δείχνει την γεωμετρία του προβλήματος στην απλή δισδιάστατη περίπτωση. Το σημαντικό σημείο της απόδειξης είναι να δειχθεί ότι το ευθύγραμμο τμήμα που σχηματίσαμε δεν μπορεί να τέμνει BR. Επιτρέπεται μόνο να τέμνει υπερεπίπεδα που έχουν ταξινομηθεί σαν BW. Το υπερεπίπεδο $\boldsymbol{d}_B$ που αναφέρεται στο λήμμα 2 είναι το πρώτο BW που τέμνει το ευθύγραμμο τμήμα ξεκινώντας από το $\boldsymbol{W}$ προς το $\boldsymbol{W}_s$.

Πράγματι, δοθέντος ενός υπερεπιπέδου $\boldsymbol{d}_p$ ταξινομημένου σαν BW από το $\boldsymbol{W}$, το σημείο τομής του με την ευθεία (t) είναι ένα σημείο $\boldsymbol{W}_p$ τέτοιο ώστε:

$$\boldsymbol{W}_p^\tau \boldsymbol{d}_p = \epsilon_p = \boldsymbol{W}_s^\tau \boldsymbol{d}_p + t_p(\boldsymbol{W}^\tau - \boldsymbol{W}_s^\tau)\boldsymbol{d}_p \tag{2.20}$$

δηλαδή,

$$t_p(\boldsymbol{W}_s^\tau - \boldsymbol{W}^\tau)\boldsymbol{d}_p = \boldsymbol{W}_s^\tau \boldsymbol{d}_p - \epsilon_p \tag{2.21}$$

Αφού το $\boldsymbol{W}_s$ είναι λύση του προβλήματος 2.3, τότε όλα τα υπερεπίπεδα $\boldsymbol{d}_p$ ταξινομούνται σωστά (BR) από το $\boldsymbol{W}_s$, και έτσι $\boldsymbol{W}_s^\tau \boldsymbol{d}_p - \epsilon_p > 0$.

Αν το $\boldsymbol{d}_p$ ταξινομείται λάθος (BW) από το $\boldsymbol{W}$, τότε $\boldsymbol{W}^\tau \boldsymbol{d}_p - \epsilon_p < 0$, έτσι ώστε

$$(\boldsymbol{W}_s^\tau - \boldsymbol{W}^\tau)\boldsymbol{d}_p > \boldsymbol{W}_s^\tau \boldsymbol{d}_p - \epsilon_p > 0 \tag{2.22}$$

Έτσι μπορούμε να εκφράσουμε το t_p σαν

$$t_p = \frac{\boldsymbol{W}_s^\tau \boldsymbol{d}_p - \epsilon_p}{(\boldsymbol{W}_s^\tau - \boldsymbol{W}^\tau)\boldsymbol{d}_p} \tag{2.23}$$

και προφανώς έπεται από την σχέση 2.22 ότι $0 < t_p < 1$.

Από την άλλη πλευρά, αν το $\boldsymbol{d}_p$ είναι BR από το $\boldsymbol{W}$, τότε $\boldsymbol{W}^\tau \boldsymbol{d}_p - \epsilon_p > 0$ και έτσι από την εξίσωση 2.23 έχουμε

$$t_p \begin{cases} < 0, & \text{αν} \quad (\boldsymbol{W}_s^\tau - \boldsymbol{W}^\tau)\boldsymbol{d}_p < 0 \\ > 1, & \text{αν} \quad (\boldsymbol{W}_s^\tau - \boldsymbol{W}^\tau)\boldsymbol{d}_p > 0 \end{cases} \tag{2.24}$$

Έτσι αποδεικνύεται ότι το ευθύγραμμο τμήμα δεν τέμνει BR, παρά μόνο BW.

Ας θεωρήσουμε τώρα το υπερεπίπεδο $\boldsymbol{d}_B$ το οποίο είναι αυτό με την μεγαλύτερη τιμή για το t_p, πάντα όμως στο διάστημα $0 < t_p < 1$. Σύμφωνα με την ανάλυση μας είναι ταξινομημένο σαν BW από το $\boldsymbol{W}$ και το σημείο τομής του με την ευθεία (t) είναι το $\boldsymbol{W}_B$ το οποίο μπορεί να υπολογιστεί σύμφωνα με τις σχέσεις 2.18.

Μένει τώρα να αποδειχθεί ότι το $\boldsymbol{W}_B$ βρίσκεται στο σύνορο του πολυτόπου $\mathcal{R}$. Αφού $\boldsymbol{W}_B^\tau \boldsymbol{d}_B = \epsilon_p$, αρκεί να δείξουμε ότι όλα τα υπόλοιπα υπερεπίπεδα ταξινομούνται από το $\boldsymbol{W}_B$ με τον ίδιο τρόπο που ταξινομούνται και από $\boldsymbol{W}$.

Έστω υπερεπίπεδο $\boldsymbol{d}_q \neq \boldsymbol{d}_B$, τότε έχουμε:

$$\boldsymbol{W}_B^\tau \boldsymbol{d}_q = \boldsymbol{W}_s^\tau \boldsymbol{d}_q + t_B(\boldsymbol{W}^\tau - \boldsymbol{W}_s^\tau)\boldsymbol{d}_q \tag{2.25}$$

Αν το $\boldsymbol{d}_q$ είναι BW, τότε $0 < t_q < t_B$ και $(\boldsymbol{W}^\tau - \boldsymbol{W}_s^\tau)\boldsymbol{d}_q < 0$, έτσι ώστε:

$$\boldsymbol{W}_B^\tau \boldsymbol{d}_q < \boldsymbol{W}_s^\tau \boldsymbol{d}_q + t_q(\boldsymbol{W}^\tau - \boldsymbol{W}_s^\tau)\boldsymbol{d}_q \tag{2.26}$$

Αντικαθιστώντας το t_q από την εξίσωση 2.23 βρίσκουμε ότι $\boldsymbol{W}_B^\tau \boldsymbol{d}_q < \epsilon_q$. Έτσι το $\boldsymbol{d}_q$ ταξινομείται σαν BW από το $\boldsymbol{W}_B$, όπως ακριβώς και από το $\boldsymbol{W}$.

Όμοια, αν το $\boldsymbol{d}_q$ είναι BR σε σχέση με το $\boldsymbol{W}$, πρέπει να εξετασθούν δύο περιπτώσεις, σύμφωνα πάντα με τις σχέσεις 2.24.

Αν $(\boldsymbol{W}_s^\tau - \boldsymbol{W}^\tau)\boldsymbol{d}_q < 0$, τότε $t_q < 0 < t_B$. Αν όμως, $(\boldsymbol{W}_s^\tau - \boldsymbol{W}^\tau)\boldsymbol{d}_q > 0$, τότε $0 < t_B < 1 < t_q$. Και στις δύο περιπτώσεις μπορούμε να χρησιμοποιήσουμε την σχέση 2.25 για να γράψουμε:

$$\boldsymbol{W}_B^\tau \boldsymbol{d}_q > \boldsymbol{W}_s^\tau \boldsymbol{d}_q + t_q(\boldsymbol{W}^\tau - \boldsymbol{W}_s^\tau)\boldsymbol{d}_q \tag{2.27}$$

Αντικαθιστώντας από την εξίσωση 2.23 βρίσκουμε ότι $\boldsymbol{W}_B^\tau \boldsymbol{d}_q > \epsilon_q$. Έτσι και στις δύο περιπτώσεις το $\boldsymbol{d}_q$ ταξινομείται σαν BR από το $\boldsymbol{W}_B$ και η απόδειξη ολοκληρώνεται.

Λήμμα 3 *Στην περίπτωση των γραμμικά διαχωρίσιμων προβλημάτων, το πιο ακραίο σημείο (ελάχιστη ενέργεια E) κατά την διεύθυνση του $\boldsymbol{\Delta W}$, μέσα σ' ένα τυχαίο πολύτοπο $\mathcal{R}$ βρίσκεται σε σημείο το οποίο ανήκει σε τουλάχιστον ένα BW υπερεπίπεδο.*

Απόδειξη Θεωρούμε το διάνυσμα $\boldsymbol{W}_B$ από το προηγούμενο λήμμα. Σύμφωνα με την εξίσωση 2.18, παίρνοντας το εσωτερικό γινόμενο με οποιοδήποτε $\boldsymbol{d}_p$, παίρνουμε:

$$-\boldsymbol{W}_B^\tau \boldsymbol{d}_p = (t_B - 1)(\boldsymbol{W}_s^\tau \boldsymbol{d}_p) + t_B(-\boldsymbol{W}^\tau \boldsymbol{d}_p) \tag{2.28}$$

Αφού το $\boldsymbol{W}_s$ είναι λύση του προβλήματος, όλα τα υπερεπίπεδα πρέπει να ταξινομούνται σωστά (BR) από το $\boldsymbol{W}_s$ άρα $\boldsymbol{W}_s^\tau \boldsymbol{d}_p > \epsilon_p$. Επίσης, $t_B - 1 < 0$ σύμφωνα με το λήμμα 2:

$$-\boldsymbol{W}_B^\tau \boldsymbol{d}_p < t_B(-\boldsymbol{W}^\tau \boldsymbol{d}_p) \tag{2.29}$$

Αθροίζοντας σ' όλα τα $\boldsymbol{d}_p$ που είναι BW σε σχέση με το $\boldsymbol{W}$ και χρησιμοποιώντας την εξίσωση 2.4 καταλήγουμε [3]. Έτσι:

$$E(\boldsymbol{W}_B) = t_B E(\boldsymbol{W}) < E(\boldsymbol{W}) \tag{2.30}$$

αφού $0 < t_B < 1$.

Από την τελευταία εξίσωση και από το προηγούμενο λήμμα, έπεται ότι για κάθε διάνυσμα $\boldsymbol{W}$ στο εσωτερικό ενός πολυτόπου $\mathcal{R}$ υπάρχει ένα διάνυσμα με χαμηλότερη ενέργεια, το οποίο ικανοποιεί την $\boldsymbol{W}_B^\tau \boldsymbol{d}_p = 0$ για τουλάχιστον ένα υπερεπίπεδο $\boldsymbol{d}_p$ το οποίο ταξινομείται σαν BW από οποιοδήποτε $\boldsymbol{W}$ μέσα στο $\mathcal{R}$.

Θεώρημα 2 (ή το θεώρημα της πόρτας) *Με τον τερματισμό του προτεινόμενου αλγόριθμου το αρχικό πρόβλημα έχει λυθεί και οι δύο κλάσεις έχουν διαχωριστεί σωστά αν και εφ' όσον το πρόβλημα είναι γραμμικά διαχωρίσιμο.*

Απόδειξη

Ας θεωρήσουμε κάποιο πολύτοπο το οποίο αντιστοιχεί σε κάποιο μη μηδενικό αριθμό BW υπερεπιπέδων. Σε σημεία που δεν αντιστοιχούν στην ελάχιστη τιμή της E ο αλγόριθμος θα έχει πάντα επιτρεπτή διεύθυνση κίνησης ακόμα και αν δεν μπορεί να κάνει 'επιταχυμένη κίνηση' βγαίνοντας από το πολύτοπο. Σ' αυτήν την περίπτωση θα συνεχίσει να εξαντλεί το πολύτοπο κατεβαίνοντας σε χαμηλότερες ενέργειες. Αν δεν συναντήσει προηγουμένως κάποιο BW[4], υποχρεούται να το συναντήσει (σε γραμμικά διαχωρίσιμα προβλήματα) όταν φτάσει στο απόλυτο ενεργειακό ελάχιστο του πολυτόπου. Τότε η 'επιταχυμένη κίνηση' είναι αναπόφευκτη, και συνεπώς ο αλγόριθμος θα συνεχίσει σ' ένα άλλο πολύτοπο, μέχρι να εξαντλήσει τα BW υπερεπίπεδα, και να λύσει τελικά το πρόβλημα.

[3] Η ενέργεια μέσα στο πολύτοπο $\mathcal{R}$ δίνεται από τον τύπο $E = \boldsymbol{W}^\tau \boldsymbol{\Delta W}$ και ελαχιστοποιείται στο πιο ακραίο σημείο του $\mathcal{R}$ κατά την διεύθυνση του $\boldsymbol{\Delta W}$

[4] Ο όρος 'πόρτα' προήλθε από την αναγκαιότητα του αλγόριθμου να εντοπίσει ένα τουλάχιστον BW στο τρέχον πολύτοπο από το οποίο θα μπορεί να διαφύγει προς χαμηλότερα ενεργειακά πολύτοπα. Η κατά σύμβαση χρησιμοποιούμενη παρομοίωση ήταν ότι ο αλγόριθμος αντιπροσωπεύεται από έναν τυφλό ο οποίος ψάχνει την έξοδο (BW) από ένα δωμάτιο (πολύτοπο) ψηλαφίζοντας τους τοίχους (BR), ακολουθώντας το ρεύμα αέρα ($\boldsymbol{\Delta W}$).

Κεφάλαιο 3

Η Βέλτιστη Διεύθυνση

Η λύση ενός προβλήματος βελτιστοποίησης που περιέχει ισοτικούς δεσμούς είναι πρόβλημα αρκετά πολύπλοκο από μόνο του. Υπάρχει όμως μια διαδικασία, όπως οι πολλαπλασιαστές Lagrange που θεωρητικά μπορούν να επιλύσουν το πρόβλημα, εφ' όσον αυτό πληρεί τις απαραίτητες προϋποθέσεις συνέχειας και διαφορισιμότητας.

Ένα πρόβλημα βελτιστοποίησης όμως, στο οποίο οι δεσμοί είναι ανισοτικοί είναι σαφώς δυσκολότερο να λυθεί αναλυτικά. Αυτό γίνεται αμέσως αντιληπτό ακόμα και σε χώρο μικρών διαστάσεων όπου είναι απαραίτητο για τον λύτη να έχει εποπτεία του προβλήματος για να μπορέσει να το λύσει. Η εποπτεία είναι απαραίτητη ώστε να μπορέσει ο λύτης να ξεχωρίσει ποιοι από τους δεσμούς είναι μέρος της λύσης και ποιοι όχι. Μ' αυτή την γνώση το πρόβλημα μετατρέπεται σε πρόβλημα βελτιστοποίησης με ισοτικούς δεσμούς που για το οποίο, όπως είδαμε, υπάρχει μια πληθώρα αναλυτικών τεχνικών και εργαλείων για την επίλυσή του.

Σε χώρους λίγων διαστάσεων η εποπτεία, αν όχι εύκολη, είναι τουλάχιστον δυνατή. Σε χώρους όμως πολλών διαστάσεων κάτι τέτοιο είναι αδύνατο. Αν υποθέσουμε ότι έχουμε M[1] δεσμούς και ξέρουμε ότι στην τελική λύση συμμετέχουν k τότε οι πιθανοί συνδυασμοί που πρέπει να εξετάσουμε είναι $\begin{pmatrix} M \\ k \end{pmatrix}$. Επειδή συνήθως δεν ξέρουμε όμως ούτε τον αριθμό k των ενεργών δεσμών ο αριθμός των πιθανών συνδυασμών είναι:

$$\sum_{k=0}^{M} \begin{pmatrix} M \\ k \end{pmatrix} = 2^M$$

Γι αυτό το λόγο προβλήματα με ανισοτικούς δεσμούς είναι πολύ πιο δύσκολα απ' ότι τα αντίστοιχα ισοτικά. Είναι κρυμμένα συνδυαστικά προβλήματα που έχουν από την φύση τους εκθετική πολυπλοκότητα.

Τα προβλήματα τετραγωνικού προγραμματισμού απαρτίζουν έναν πολύ ενδιαφέροντα και σημαντικό, με μακρά ιστορία, τομέα του μαθηματικού προγραμματισμού γενικότερα. Ο τετραγωνικός προγραμματισμός έχει εφαρμογές σε πάρα πολλές διαφορετικές περιοχές, που ποικίλλουν από μηχανολογία και μαθηματικά μέχρι φυσική και κοινωνικές επιστήμες. Πρόσφατα, το ενδιαφέρον για τον τετραγωνικό προγραμματισμό αναζωπυρώθηκε εξαιτίας του ρόλου κλειδιού που κατέχει σε πολλούς επιτυχημένους αλγόριθμους βελτιστοποίησης μη γραμμικών συναρτήσεων κάτω από δεσμούς.

Όπως δείχνουμε και με την παρούσα εργασία η αποσύνθεση μεγάλων προβλημάτων, ακόμα και γραμμικού προγραμματισμού, σε πολλά μικρότερα τετραγωνικού προγραμματισμού για την εύρεση των βέλτιστων διευθύνσεων βελτιώνει την συνολική απόδοση. Έτσι, η δυνατότητα λύσης προβλημάτων τετραγωνικού προγραμματισμού αποτελεί το σημείο κλειδί σ' ένα ζωτικό υποσύστημα που επηρεάζει την απόδοση των αλγορίθμων που επιλύουν γραμμικά ή μη προγράμματα.

[1] Ο αναγνώστης θα πρέπει να μας συγχωρήσει τη χρήση του M αλλά έχουμε δεσμεύσει το N, από το προηγούμενο κεφάλαιο σαν τον αριθμό των ελεύθερων μεταβλητών.

Εξ αιτίας της μακράς του ιστορίας, αλλά και του ευρύτατου πεδίου εφαρμογών, υπάρχει μεγάλος αριθμός προσεγγίσεων και μεθόδων για την επίλυση προβλημάτων τετραγωνικού προγραμματισμού [35]. Οι τεχνικές αυτές μπορεί να ταξινομηθούν σε δύο βασικές κατηγορίες, στις πεπερασμένες και τις επαναληπτικές.

- Οι πεπερασμένες τεχνικές περιγράφονται από μια σειρά διακριτών και πεπερασμένων σε αριθμό βημάτων.
- Οι επαναληπτικές συγκλίνουν στο όριο των πολλών επαναλήψεων. Στο παρόν θα ασχοληθούμε μόνο με τις πεπερασμένες, ενώ για τις επαναληπτικές θα αρκεστούμε να πούμε ότι λόγω της ευκολίας της διατήρησης της πληροφορίας σε αραιή μορφή (sparce) είναι πολύ ελκυστικές σε προβλήματα μεγάλης κλίμακας. Μια πολύ γνωστή επαναληπτική μέθοδος αναπτύχθηκε από τον Karmarkar [36] και βασίζεται στα συστελλόμενα ελλειψοειδή (Shrinking ellipsoids).

Στο παρόν θα κάνουμε μερικές υποθέσεις οι οποίες θα απλοποιήσουν το γενικό πρόβλημα, όμως όπως θα δούμε όχι αρκετά για να χάσει το ενδιαφέρον του. Συγκεκριμένα το πρόβλημα που θα μας απασχολήσει έχει κάποια χαρακτηριστικά που το διαφοροποιούν από τα υπόλοιπα προβλήματα μη γραμμικού, ή και του γενικού τετραγωνικού προγραμματισμού. Πιο συγκεκριμένα:

- Ο αριθμός των δεσμών που πρέπει να ικανοποιούνται με την λύση του προβλήματος είναι $M \leq N+1$, ενώ στο προηγούμενο κεφάλαιο ο αριθμός των ανισοτικών δεσμών P ήταν συνήθως μεγαλύτερος του $N+1$, αν και δεν υπήρχε ουσιαστικός περιορισμός. Θα πρέπει να υπενθυμίσουμε ότι η διάσταση του προβλήματος έχει αυξηθεί κατά ένα $(N+1)$ από την επαύξηση του αναζητούμενου διανύσματος στο κεφάλαιο 2.1.1 ώστε να συμπεριλάβει και τον όρο του κατωφλιού (threshold).
- Η συνάρτηση κόστους είναι τετραγωνική[2] ενώ στο προηγούμενο κεφάλαιο μας ήταν αδιάφορη (Ο στόχος ήταν η ικανοποίηση των δεσμών και μόνο).
- Και εδώ όπως και στο προηγούμενο κεφάλαιο, οι δεσμοί είναι γραμμικοί και ανισοτικοί.

Το πρόβλημα συζητείται διεξοδικά ενώ παρουσιάζεται και η γεωμετρική ερμηνεία του και με τη βοήθεια της γραμμικής άλγεβρας. Το παρόν κεφάλαιο μπορεί να ειδωθεί ως απαραίτητη συνέχεια του προηγούμενου κεφαλαίου, αλλά έχει και αυτόνομη αξία διότι μπορεί να αποτελέσει τη βάση για την επίλυση των περισσότερων προβλημάτων βελτιστοποίησης κάτω από γραμμικούς δεσμούς.

3.1 Το τετραγωνικό πρόβλημα

3.1.1 Πρώιμη διατύπωση

Το ζητούμενο σ' αυτό το κεφάλαιο είναι να βρεθεί διάνυσμα $\boldsymbol{P}$ τέτοιο ώστε να σχηματίζει την ελάχιστη δυνατή γωνία με το επιθυμητό διάνυσμα κατεύθυνσης $\boldsymbol{\Delta W}$, και να ικανοποιεί ταυτόχρονα τους M ενεργούς δεσμούς $\boldsymbol{P}^\tau \boldsymbol{d}_p \geq 0, p = 1 \ldots M$. Θέλουμε λοιπόν:

$$\begin{cases} max\ f = cos(\phi) = \dfrac{\boldsymbol{P}^\tau \boldsymbol{\Delta W}}{\|\boldsymbol{P}\| \|\boldsymbol{\Delta W}\|} \\ \boldsymbol{d}_p^\tau \boldsymbol{P} \geq 0, p = 1 \ldots M \end{cases} \tag{3.1}$$

Όπως μπορεί να δει κανείς από την παραπάνω εξίσωση υπάρχει απειρία λύσεων, ακριβώς όπως και στο κεφάλαιο 2.1.3, αφού αν $\boldsymbol{P}$ λύση του προβλήματος τότε και $\lambda \boldsymbol{P}$ με $\lambda > 0$ αποτελεί λύση. Αν θέλουμε να περιορίσουμε αυτόν τον επιπλέον βαθμό ελευθερίας θα πρέπει να διώξουμε τον παρονομαστή από τη

[2] Η συνάρτηση κόστους είναι τετραγωνική και αντιπροσωπεύει έλλειψη και όχι υπερβολή.

συνάρτηση κόστους. Αυτό γίνεται εύκολα αν απαιτήσουμε το $\boldsymbol{P}$ να έχει σταθερό μήκος π.χ. ίσο με 1, μια και δεν παίζει ρόλο όπως είδαμε. Τότε όμως το πρόβλημα γίνεται:

$$\begin{cases} max\ f = \boldsymbol{P}^\tau \boldsymbol{\Delta W} \\ \|\boldsymbol{P}\| = 1 \\ \boldsymbol{d}_p^\tau \boldsymbol{P} \geq 0, p = 1 \ldots M \end{cases} \tag{3.2}$$

πράγμα που σημαίνει ότι μετατοπίσαμε το μη γραμμικό κομμάτι του προβλήματος από την συνάρτηση κόστους στο τμήμα των δεσμών.

Στην πραγματικότητα υπάρχουν παραπάνω από ένα τρόποι να φράξουμε το μέτρο της λύσης. Αυτή η πολλαπλότητα επιλογών οδήγησε τον Zoutendijk [37] να παράγει μια πληθώρα παρόμοιων τεχνικών, για την εύρεση αποδεκτών λύσεων. Οι τεχνικές αυτές διαφέρουν μεταξύ τους στην 'κανονικοποίηση' της βελτιστοποιούμενης συνάρτησης, για να χρησιμοποιήσουμε έναν όρο του Zoutendijk. Στις μεθόδους του Zoutendijk γίνεται πιο εκτενής αναφορά στο κεφάλαιο 3.2.2.

3.1.2 Συναρτήσεις κόστους

Οι παραπάνω μορφές μπορεί να είναι ισοδύναμες με την έννοια ότι παράγουν ισοδύναμα προγράμματα για την επίλυσή τους αλλά έχουν το μειονέκτημα ότι δεν κάνουν την γεωμετρική ερμηνεία του προβλήματος προφανή. Εφ' όσον, όπως είδαμε, το μέτρο της λύσης δεν αποτελεί κρίσιμο παράγοντα, είμαστε ελεύθεροι να επιλέξουμε έναν περιορισμό που να μας επιτρέπει να σχεδιάσουμε έναν αλγόριθμο που να λύνει αποδοτικά αυτό το μη γραμμικό πρόβλημα.

Ο περιορισμός που διαλέγουμε απαιτεί η λύση $\boldsymbol{P}$ να αποτελεί προβολή του $\boldsymbol{\Delta W}$, δηλαδή $\boldsymbol{P}^\tau \boldsymbol{\Delta W} = \boldsymbol{P}^2$. Το δικαίωμα να επιλέξουμε κάτι τέτοιο μας το δίνει η φύση του προβλήματος που έχει έναν βαθμό ελευθερίας επιπλέον. Πράγματι, για κάθε βέλτιστη διεύθυνση είναι προφανές ότι θα υπάρχει και ένα διάνυσμα με αυτή την διεύθυνση και μέτρο το μέτρο της προβολής του $\boldsymbol{\Delta W}$ προς αυτήν. Ενώ αυτή η επιλογή φαίνεται αυθαίρετη σε πρώτη φάση, θα δούμε ότι παρέχει κάποιες διευκολύνσεις τόσο στη κατανόηση της φύσης του προβλήματος όσο και στην απλοποίηση των πράξεων που απαιτούνται σε συμβολικό επίπεδο.

Πράγματι, αν θεωρήσουμε ότι ισχύει ο δεσμός $\boldsymbol{P}^\tau \boldsymbol{\Delta W} = \boldsymbol{P}^2$ και αντικαταστήσουμε στην εξίσωση 3.1 έχουμε:

Εφ' όσον το μέτρο του $\boldsymbol{\Delta W}$ είναι δεδομένο και η συνάρτηση της τετραγωνικής ρίζας είναι αύξουσα, το πρόβλημα γράφεται ισοδύναμα:

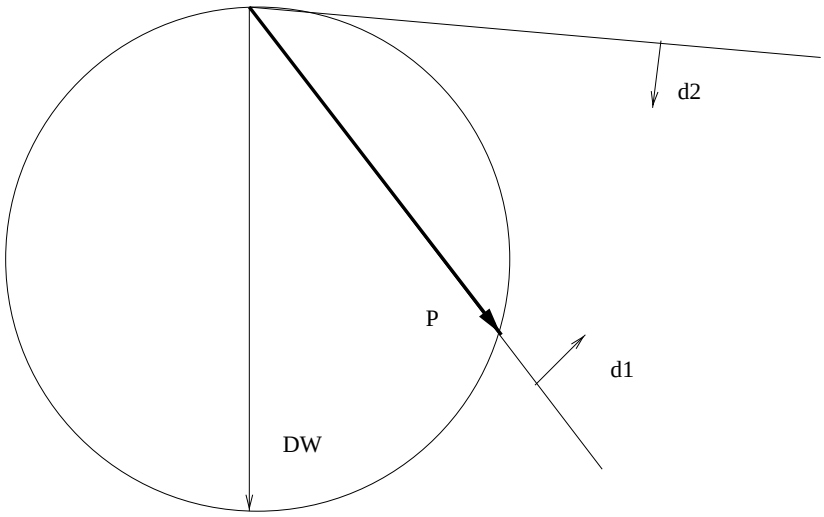
$$\begin{cases} max\ f = \boldsymbol{P}^\tau \boldsymbol{\Delta W} \\ \boldsymbol{P}^\tau \boldsymbol{\Delta W} = \boldsymbol{P}^2 \\ \boldsymbol{d}_p^\tau \boldsymbol{P} \geq 0, p = 1 \ldots M \end{cases} \tag{3.3}$$

που βέβαια δεν αποτελεί σημαντική βελτίωση από τις προηγούμενες διατυπώσεις αλλά μπορούμε να αρχίσουμε να απολαμβάνουμε τους καρπούς των επιλογών μας διότι ο περιορισμός $\boldsymbol{P}^\tau \boldsymbol{\Delta W} = \boldsymbol{P}^2$ έχει ξεκάθαρη γεωμετρική ερμηνεία όπως εικονίζεται και στο σχήμα 3.1.

Πράγματι, στις δύο διαστάσεις ο γεωμετρικός τόπος που παράγεται από όλα τα δυνατά $\boldsymbol{P}$ είναι ένας κύκλος με την αρχή των αξόνων στην αρχή του $\boldsymbol{\Delta W}$ και διάμετρο το ίδιο το $\boldsymbol{\Delta W}$. Σε περισσότερες από δύο διαστάσεις ο κύκλος γίνεται σφαίρα ή υπερσφαίρα ανώτερων διαστάσεων. Στην αρχή των αξόνων βρίσκεται επίσης και το σημείο τομής όλων των υπερεπιπέδων από το οποίο προσπαθούμε να ξεφύγουμε ακολουθώντας την βέλτιστη πορεία, αυτή που σχηματίζει την ελάχιστη γωνία ως προς το $\boldsymbol{\Delta W}$, χωρίς όμως να διασχίσουμε κάποια από τα εν λόγω υπερεπίπεδα.

Από τη στιγμή που έχουμε κάνει την βασική παραδοχή ότι η λύση είναι προβολή του $\boldsymbol{\Delta W}$ σε μια τυχούσα διεύθυνση[3] μπορούμε να το χρησιμοποιήσουμε για να απλοποιήσουμε ακόμα περισσότερο το πρόβλημα, και πιο συγκεκριμένα να απαλλαγούμε από τον μη γραμμικό δεσμό.

[3] Όπως θα δούμε, η διεύθυνση προβολής δεν είναι καθόλου τυχούσα αλλά παράγεται από έναν συγκεκριμένο μεν, άγνωστο δε, συνδυασμό δεσμών.



Σχήμα 3.1: Ένα απλό παράδειγμα στις δύο διαστάσεις με δύο δεσμούς, εκ των οποίων μόνο ο ένας συμμετέχει στη λύση.

Ξεκινάμε με την παρατήρηση ότι η Ευκλείδεια απόσταση ενός διανύσματος (σημείο) από ένα υπερεπίπεδο ελαχιστοποιείται από το διάνυσμα του υπερεπιπέδου που αποτελεί προβολή του αρχικού διανύσματος στο αυτό υπερεπίπεδο. Έτσι, μπορούμε να ξαναγράψουμε το πρόβλημα ενσωματώνοντας τον μη γραμμικό δεσμό στη συνάρτηση κόστους.

$$\begin{cases} min\ f = \frac{1}{2}(\boldsymbol{P} - \boldsymbol{\Delta W})^2 \\ \boldsymbol{d}_p^\tau \boldsymbol{P} \geq 0, p = 1 \ldots M \end{cases} \tag{3.4}$$

Αφού στο σημείο $\boldsymbol{P}$ που ελαχιστοποιείται η $\frac{1}{2}(\boldsymbol{P} - \boldsymbol{\Delta W})^2$ ισχύει ο δεσμός $\boldsymbol{P}^\tau \boldsymbol{\Delta W} = \boldsymbol{P}^2$

$$min\ f = \frac{1}{2}(\boldsymbol{P} - \boldsymbol{\Delta W})^2 = \tag{3.5}$$

$$min\ f = \frac{1}{2}(\boldsymbol{P}^2 + \boldsymbol{\Delta W}^2 - 2\boldsymbol{P}^\tau \boldsymbol{\Delta W}) = \tag{3.6}$$

$$min\ f = \frac{1}{2}(\boldsymbol{\Delta W}^2 - \boldsymbol{P}^\tau \boldsymbol{\Delta W}) = \tag{3.7}$$

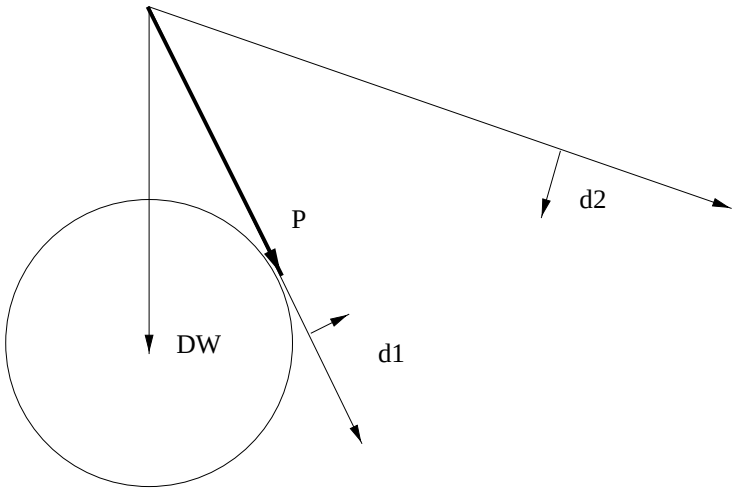
$$max\ \boldsymbol{P}^\tau \boldsymbol{\Delta W} \tag{3.8}$$

που είναι και το ζητούμενο του προβλήματος 3.3.

Το καινούριο πρόβλημα 3.4 έχει μερικά πλεονεκτήματα σε σχέση με τους ισοδύναμους προκατόχους του. Κατ΄ αρχήν δεν έχει μη γραμμικό δεσμό, γεγονός που απλοποιεί αρκετά την απαραίτητη άλγεβρα για τον υπολογισμό των πολλαπλασιαστών Lagrange, ενώ όπως θα δούμε, η εξαγωγή του δυϊκού προβλήματος γίνεται είναι προφανής. Η συνάρτηση κόστους του είναι μια απλή, κυρτή, τετραγωνικής μορφής συνάρτηση που έχει και στο παρελθόν απασχολήσει την επιστημονική κοινότητα αφού αποτελεί την πιο απλή μη γραμμική συνάρτηση κόστους.

Ακόμα περισσότερο όμως, η διατύπωση 3.4 έχει το πλεονέκτημα ότι προσδίδει στο πρόβλημα μια ακόμη γεωμετρική ερμηνεία που μαζί με την προηγούμενη (3.1) θα χρησιμοποιούνται εναλλάξιμα για την κατάδειξη συγκεκριμένων θέσεων και αντιπαραδειγμάτων ώστε να δειχθεί, στο μέτρο που αυτό είναι δυνατό, η διαδικασία από την οποία προέκυψαν οι προτεινόμενοι αλγόριθμοι.

Η καινούρια γεωμετρική ερμηνεία όπως αυτή εικονίζεται στο σχήμα 3.2 απαιτεί την ελαχιστοποίηση της ακτίνας του κύκλου (στις δύο διαστάσεις) με κέντρο το $\boldsymbol{\Delta W}$ αλλά πρέπει παρ' όλα αυτά να εφάπτεται



Σχήμα 3.2: Η γεωμετρική ερμηνεία της σχέσης 3.4

στο θετικό ημιχώρο που ορίζεται από την τομή των υπερεπιπέδων που είναι κάθετα στα διανύσματα που αποτελούν τους δεσμούς.

Το πρόβλημα 3.4 δεν αποτελεί το γενικό πρόβλημα τετραγωνικού προγραμματισμού και υπάρχουν δύο βασικοί λόγοι για αυτό.

1. Η συνάρτηση κόστους είναι συνάρτηση ελαχίστων τετραγώνων και όχι μια οποιαδήποτε κυρτή τετραγωνική συνάρτηση.

2. Το πλήθος των δεσμών ισούται με τους βαθμούς ελευθερίας του προβλήματος.

Η δεύτερη παρατήρηση πρέπει να ηχεί σαν έκπληξη διότι απαιτεί λογικό άλμα από τον αναγνώστη, σε σχέση με την ως τώρα παρουσίαση του προβλήματος, όπου δεχόμαστε ότι $M \leq N+1$ όπου $N+1$ είναι οι βαθμοί ελευθερίας του προβλήματος. Έτσι εδώ παραθέτουμε μια σύντομη απόδειξη για τον απαιτητικό αναγνώστη.

Εφ' όσον τα M διανύσματα που απαρτίζουν τους δεσμούς είναι γραμμικώς ανεξάρτητα, παράγουν και τον υπόχωρο των M διαστάσεων στον οποίο ανήκουν. Είναι λοιπόν δυνατόν να γράψουμε κάθε διάνυσμα των $N+1$ διαστάσεων σαν συνδυασμό ενός διανύσματος που ανήκει στον υπόχωρο και ενός που είναι κάθετο ως προς αυτόν[4]. Έχουμε δηλαδή:

$$\begin{cases} f = \frac{1}{2}(\boldsymbol{P} - \boldsymbol{\Delta W})^2 \\ \boldsymbol{d}_p^\tau \boldsymbol{P} \geq 0, p = 1 \ldots M \end{cases} \tag{3.9}$$

$$\begin{cases} f = \frac{1}{2}((\boldsymbol{P}_\perp + \boldsymbol{P}_\parallel) - (\boldsymbol{\Delta W}_\perp + \boldsymbol{\Delta W}_\parallel))^2 \\ \boldsymbol{d}_p^\tau (\boldsymbol{P}_\perp + \boldsymbol{P}_\parallel) \geq 0, p = 1 \ldots M \end{cases} \tag{3.10}$$

Δεδομένου ότι $\boldsymbol{d}_p^\tau \boldsymbol{P}_\perp^\tau = 0, p = 1 \ldots M$ με μια μικρή αναδιάταξη στους όρους του τετραγωνικού αναπτύγματος, έχουμε:

$$\begin{cases} f = & \frac{1}{2}((\boldsymbol{P}_\perp - \boldsymbol{\Delta W}_\perp)^2 + (\boldsymbol{P}_\parallel - \boldsymbol{\Delta W}_\parallel)^2 + \\ & 2(\boldsymbol{P}_\perp^\tau - \boldsymbol{\Delta W}_\perp^\tau)(\boldsymbol{P}_\parallel - \boldsymbol{\Delta W}_\parallel)) \\ \boldsymbol{d}_p^\tau \boldsymbol{P}_\parallel & \geq 0, p = 1 \ldots M \end{cases} \tag{3.11}$$

[4]και είναι επίσης κάθετο ως προς κάθε διάνυσμα αυτού του υποχώρου.

Όπως μπορούμε να δούμε το εσωτερικό γινόμενο μπορεί να φύγει από την εξίσωσή μας εφ' όσον πολλαπλασιάζουμε διανύσματα τα οποία είναι εξ' ορισμού κάθετα. Κοιτώντας τις συνθήκες των δεσμών παρατηρούμε ότι οι δεσμοί αναφέρονται μόνο στο $\boldsymbol{P}_{\parallel}$ κομμάτι του $\boldsymbol{P}$. Συνεπώς μπορούμε να θεωρήσουμε ότι $\boldsymbol{P}_{\perp} = \boldsymbol{\Delta W}_{\perp}$ μια και αυτή η υπόθεση ελαχιστοποιεί τον έναν τετραγωνικό όρο και δεν παραβιάζει και κανένα δεσμό. Έχουμε λοιπόν:

$$\begin{cases} f = \frac{1}{2}(\boldsymbol{P}_{\parallel} - \boldsymbol{\Delta W}_{\parallel})^2 \\ \boldsymbol{d}_p^{\tau}\boldsymbol{P}_{\parallel} \geq 0, p = 1 \ldots M \end{cases} \tag{3.12}$$

Η διατύπωση στην οποία καταλήξαμε είναι ταυτόσημη με αυτή του προβλήματος 3.4 αλλά αναφέρεται στον υπόχωρο των M διαστάσεων και όχι στον πλήρη χώρο. Ταυτόχρονα όμως είναι και ισοδύναμη γιατί όπως είδαμε λύνοντας το ένα πρόβλημα λύνεται και το άλλο. Το ενδιαφέρον στοιχείο που θα πρέπει να συγκρατήσουμε είναι ότι η διάσταση του προβλήματος καθορίζεται από το πλήθος των δεσμών και όχι από το πλήθος των αρχικών βαθμών ελευθερίας του προβλήματος. Ένα πρόβλημα $N+1$ διαστάσεων μ' ένα δεσμό είναι στην πραγματικότητα ένα μονοδιάστατο πρόβλημα όπως θα δείξουμε στην άσκηση του κεφαλαίου 3.1.3.

Λαμβάνοντας υπόψη αυτά τα δύο στοιχεία που διαφοροποιούν το πρόβλημά μας από το γενικότερο του τετραγωνικού προγραμματισμού, είμαστε σε θέση να σχεδιάσουμε αλγόριθμους με σαφώς ανώτερη απόδοση αφού θα χρησιμοποιούν και την επιπλέον διαθέσιμη πληροφορία.

3.1.3 Το μονοδιάστατο πρόβλημα

Μια αναλυτική επίθεση στο πρόβλημα καταφέρνει, χωρίς να έχει πρακτική χρησιμότητα, να μας εξοικειώνει με τις δυσκολίες, αλλά και τις συμμετρίες που παρουσιάζει το συγκεκριμένο πρόβλημα. Στο παρόν θα επιλύσουμε μέχρι τέλους την περίπτωση όπου το πλήθος των δεσμών είναι ένα (1).

Πριν προχωρήσουμε όμως στην συνέχεια της άσκησης είναι σκόπιμο να επισημάνουμε ότι το σύμβολο $\boldsymbol{Q}$ θα χρησιμοποιείται στη θέση του $\boldsymbol{P}_{\parallel}$ για λόγους αναγνωσιμότητας. Το $\boldsymbol{Q}$ δηλαδή αποτελεί το κομμάτι του $\boldsymbol{P}$ που ανήκει στον υπόχωρο που παράγεται από τους δεσμούς.

$$\begin{cases} f = \frac{1}{2}(\boldsymbol{Q} - \boldsymbol{\Delta W})^2 \\ \boldsymbol{d}^{\tau}\boldsymbol{Q} \geq 0 \end{cases} \tag{3.13}$$

Δεδομένου ότι η μέθοδος των πολλαπλασιαστών Lagrange δεν μπορεί να χρησιμοποιηθεί σε ανισοτικούς δεσμούς είμαστε αναγκασμένοι να καταφύγουμε στο ίδιο τρυκ που χρησιμοποιεί και η μέθοδος simplex, στην εισαγωγή καινούριων (slack) μεταβλητών. Έτσι, απαιτούμε $\boldsymbol{d}^{\tau}\boldsymbol{Q} = q$ όπου $q \geq 0$. Αν λ είναι ο πολλαπλασιαστής Lagrange που σχετίζεται με τον δεσμό, τότε η εκτεταμένη συνάρτηση γράφεται:

$$F = \frac{1}{2}(\boldsymbol{Q} - \boldsymbol{\Delta W})^2 + \lambda(\boldsymbol{d}^{\tau}\boldsymbol{Q} - q) \tag{3.14}$$

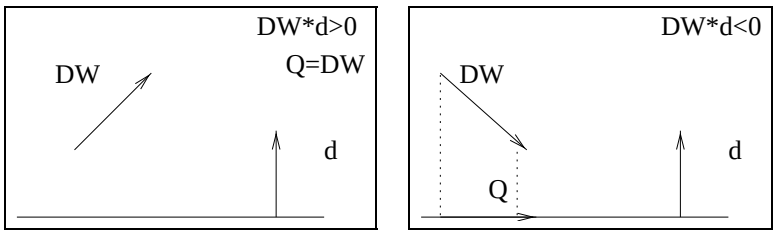
$$\boldsymbol{\nabla F} = (\boldsymbol{Q}^{\tau} - \boldsymbol{\Delta W}^{\tau}) + \lambda \boldsymbol{d}^{\tau} = 0 \tag{3.15}$$

Παίρνοντας το εσωτερικό γινόμενο της εξίσωσης 3.15 με $\boldsymbol{d}$, παίρνοντας υπόψη τον ισοτικό δεσμό και λύνοντας ως προς λ έχουμε:

$$\lambda = \frac{\boldsymbol{\Delta W}^{\tau}\boldsymbol{d} - q}{\boldsymbol{d}^2} \tag{3.16}$$

$$\boldsymbol{Q} = \boldsymbol{\Delta W} + \frac{q - \boldsymbol{\Delta W}^{\tau}\boldsymbol{d}}{\boldsymbol{d}^2}\boldsymbol{d} \tag{3.17}$$

Το πρόβλημα έχει τώρα λυθεί διότι έχουμε απαλείψει το λ και μπορούμε ελεύθερα πλέον να βελτιστοποιήσουμε την f προσέχοντας μόνο να μην παραβιάσουμε τον δεσμό $q \geq 0$.



Σχήμα 3.3: Οι δύο πιθανές λύσεις ανάλογα με το πρόσημο του γινομένου $\boldsymbol{\Delta W d}$.

$$f = \frac{1}{2}(\boldsymbol{Q} - \boldsymbol{\Delta W})^2 \tag{3.18}$$

$$f = \frac{1}{2}\left(\frac{(q - \boldsymbol{\Delta W}^\tau \boldsymbol{d})}{\boldsymbol{d}^2}\boldsymbol{d}\right)^2 \tag{3.19}$$

$$f = \frac{1}{2}\left(\frac{q^2 + (\boldsymbol{\Delta W}^\tau \boldsymbol{d})^2 - 2q(\boldsymbol{\Delta W}^\tau \boldsymbol{d})}{\boldsymbol{d}^2}\right) \tag{3.20}$$

Παρατηρούμε ότι η f εξαρτάται μόνο από το q επιβεβαιώνοντας τα συμπεράσματά μας από το κεφάλαιο 3.1.2, ότι δηλαδή πρόκειται ουσιαστικά για ένα μονοδιάστατο πρόβλημα. Η f αντιπροσωπεύει μια παραβολή, με τα κοίλα προς τα πάνω, της οποίας το ελάχιστο βρίσκεται στο σημείο $\boldsymbol{\Delta W}^\tau \boldsymbol{d}$. Αν λοιπόν αυτό το σημείο βρίσκεται στο θετικό ημιάξονα, τότε $q = \boldsymbol{\Delta W}^\tau \boldsymbol{d}$ και $\lambda = 0$ άρα και $\boldsymbol{Q} = \boldsymbol{\Delta W}$ από τις εξισώσεις 3.16 και 3.17 αντίστοιχα. Αν όμως το γινόμενο $\boldsymbol{\Delta W}^\tau \boldsymbol{d}$ είναι αρνητικό, τότε το πλησιέστερο επιτρεπτό σημείο για το q είναι το μηδέν (0). Σ' αυτή την περίπτώση έχουμε:

$$\boldsymbol{Q} = \boldsymbol{\Delta W} - \frac{(\boldsymbol{\Delta W}^\tau \boldsymbol{d})}{\boldsymbol{d}^2}\boldsymbol{d} \tag{3.21}$$

Η παρατήρηση ότι το $\boldsymbol{Q}$ και στις δύο περιπτώσεις είναι προβολή του $\boldsymbol{\Delta W}$[5] δεν πρέπει να μας ξενίζει αφού προκύπτει από την κατασκευή του προβλήματος. Το ίδιο το αποτέλεσμα δεν έχει κανένα πρακτικό αντίκρυσμα αφού η λύση στην οποία καταλήξαμε μπορεί να θεωρηθεί προφανής όπως φαίνεται κι από το σχήμα 3.3.

3.1.4 Συμβολισμός πινάκων

Προσπαθώντας να γενικεύσουμε την παραπάνω στρατηγική σε περισσότερες διαστάσεις θα οδηγηθούμε σ' έναν μετασχηματισμό προς ορθογώνιες συντεταγμένες. Ο μετασχηματισμός αυτός είναι σημαντικός για την περαιτέρω κατανόηση του κειμένου και των προτεινόμενων αλγορίθμων, διότι μετασχηματίζει το πρόβλημα στο φυσικό του σύστημα συντεταγμένων.

Πριν προχωρήσουμε όμως θα ήταν δόκιμο να εισάγουμε μερικές γεωμετρικές έννοιες που θα μας βοηθήσουν στην κατανόηση των πινάκων που θα χρησιμοποιηθούν για χάρη ενός πιο πυκνού, αλλά και πιο συνεπή, συμβολισμού.

Κατ' αρχήν, θα αντικαταστήσουμε την πλειάδα των δεσμών $\boldsymbol{d}_p^\tau \boldsymbol{Q} \geq 0, p = 1 \ldots M$ με την πολύ πιο συμπαγή μορφή $\boldsymbol{D}^\tau \boldsymbol{Q} \geq \boldsymbol{0}$ όπου ο πίνακας $\boldsymbol{D}^\tau$ είναι $M \times (N+1)$ και έχει σαν γραμμές τα διανύσματα που αποτελούν τους δεσμούς.

[5] Τη μια φορά είναι προβολή του $\boldsymbol{\Delta W}$ σε μια κενή λίστα δεσμών και συνεπώς είναι το ίδιο το $\boldsymbol{\Delta W}$, ενώ στην άλλη είναι προβολή του $\boldsymbol{\Delta W}$ στο δεσμό $\boldsymbol{d}$

$$\boldsymbol{D}^\tau = \begin{bmatrix} \boldsymbol{d}_1^\tau \dots \\ \boldsymbol{d}_2^\tau \dots \\ \vdots \\ \boldsymbol{d}_M^\tau \dots \end{bmatrix} \tag{3.22}$$

Είναι επόμενο από τη στιγμή που βλέπουμε ένα πίνακα να αρχίσουν να τίθενται ερωτήσεις σχετικά με τον αντίστροφό του μια και έτσι γίνονται περισσότερο αντιληπτές οι διάφορες ιδιότητές του. Στη συγκεκριμένη περίπτωση δεν μπορεί να υπάρχει πλήρης αντίστροφος τη στιγμή που ο πίνακας δεν είναι τετραγωνικός. Υπάρχει όμως πίνακας $\boldsymbol{V}$ τέτοιος που πολλαπλασιαζόμενος με τον $\boldsymbol{D}^\tau$ από αριστερά μας δίνει τον $M \times M$ μοναδιαίο. Πιο συγκεκριμένα ισχύει:

$$\boldsymbol{D}^\tau \boldsymbol{V} = \boldsymbol{V}^\tau \boldsymbol{D} = \boldsymbol{I} \tag{3.23}$$

Ας προσπαθήσουμε όμως να δικαιολογήσουμε αυτό το μάλλον μαγικό αποτέλεσμα. Θα προτιμήσουμε για λόγους παρουσίασης πρώτα να κατασκευάσουμε τον πίνακα $\boldsymbol{V}$ και μετά να δώσουμε την γεωμετρική ερμηνεία του.

Gramm - Schmidt

Η διαδικασία Gramm - Schmidt [27] παράγει μια ορθοκανονική βάση από ένα σύνολο, γραμμικώς ανεξάρτητων, διανυσμάτων. Η βασική ιδέα έχει και εδώ να κάνει με προβολές αφού από κάθε διάνυσμα αφαιρούμε την προβολή του στον υπόχωρο των διανυσμάτων που έχουν ήδη προηγηθεί στην διαδικασία. Αν δηλαδή τα διανύσματα $\boldsymbol{u}_i, i = 1 \dots M$ αποτελούν την ορθογώνια βάση, τότε αυτά μπορούν να παραχθούν από την εξής διαδικασία:

$$\boldsymbol{u}_i = \boldsymbol{d}_i - \sum_{j=1}^{i-1} \frac{\boldsymbol{d}_i^\tau \boldsymbol{u}_j}{\boldsymbol{u}_j^2} \boldsymbol{u}_i \tag{3.24}$$

Μερικές χρήσιμες ιδιότητες αυτής της ορθογώνιας βάσης είναι:

- $\frac{\boldsymbol{u}_i^\tau}{\|\boldsymbol{u}_i\|} \frac{\boldsymbol{u}_j}{\|\boldsymbol{u}_j\|} = \delta_{ij}\ \forall i, j$ προφανώς λόγω ορθογωνιότητος.
- $\boldsymbol{u}_i^\tau \boldsymbol{d}_j = 0\ \forall i > j$ διότι το διάνυσμα $\boldsymbol{u}_i$ είναι ουσιαστικά το $\boldsymbol{d}_i$ από το οποίο έχουν αφαιρεθεί όλες οι j προηγούμενες προβολές του στο χώρο των $\boldsymbol{d}_j$ διανυσμάτων και, συνεπώς, είναι κάθετο προς αυτά.
- $\boldsymbol{u}_i^\tau \boldsymbol{d}_i = \boldsymbol{u}_i^2$ λόγω προβολής

Ενδιαφέρουσα είναι και η παρατήρηση ότι τα διανύσματα $\boldsymbol{u}_i$ εξαρτώνται από την εκάστοτε διάταξη, τον τρόπο δηλαδή που αριθμούμε τα $\boldsymbol{d}_i$. Είναι φανερό ότι αν αλλάξουμε τον τρόπο αρίθμησης των $\boldsymbol{d}_i$ αλλάζουν και τα $\boldsymbol{u}_i$. Παρ' όλα αυτά είμαστε αρκετά κοντά στο να εξάγουμε κάποιες ποσότητες που να παραμένουν αναλλοίωτες σε πιθανές μεταθέσεις διάταξης του συνόλου των δεσμών, δηλαδή ο υπολογισμός τους να μην επηρεάζεται από την τρέχουσα αρίθμηση του συνόλου.

Αν ξεχωρίσουμε το τελευταίο διάνυσμα $\boldsymbol{u}_i$ για $i = M$ που παράγεται από την Gramm - Schmidt παρατηρούμε ότι είναι κάθετο σ' όλα τα άλλα διανύσματα $\boldsymbol{d}_i$ εκτός του ενός από το οποίο παράχθηκε και τυχαίνει να είναι τελευταίο σύμφωνα με την τρέχουσα αρίθμηση. Εκτελώντας την διαδικασία Gramm - Schmidt M φορές έχοντας διαφορετικό τελικό διάνυσμα $\boldsymbol{d}_i$ κάθε φορά και κρατώντας το τελευταίο διάνυσμα, έστω $\boldsymbol{v}_i$[6], που παράγεται από την διαδικασία, έχουμε:

$$\frac{\boldsymbol{v}_i^\tau}{\boldsymbol{v}_i^2} \boldsymbol{d}_j = \delta_{ij}\ \forall i, j \tag{3.25}$$

[6] Το $\boldsymbol{v}_i$ ισούται με το τελικό $\boldsymbol{u}_i$ φυσικά

Έτσι κατασκευάζουμε τον πίνακα $\boldsymbol{V}^\tau$ που έχει σαν γραμμές τα διανύσματα $\frac{\boldsymbol{v}_i^\tau}{\boldsymbol{v}_i^2}$ και ο οποίος εκ κατασκευής ικανοποιεί τις εξισώσεις 3.23.

$$\boldsymbol{V}^\tau = \begin{bmatrix} \frac{\boldsymbol{v}_1^\tau}{\boldsymbol{v}_1^2} \cdots \\ \frac{\boldsymbol{v}_2^\tau}{\boldsymbol{v}_2^2} \cdots \\ \vdots \\ \frac{\boldsymbol{v}_M^\tau}{\boldsymbol{v}_M^2} \cdots \end{bmatrix} \tag{3.26}$$

Γεωμετρική ερμηνεία

Ακόμα και ένας τόσο σημαντικός πίνακας όσο ο αντίστροφος ενός άλλου δεν αξίζει τόσο μόνος του όσο μαζί με την γεωμετρική του ερμηνεία, όπου αυτή είναι δυνατή. Για να μπορέσουμε να αντιληφθούμε την γεωμετρική του ερμηνεία, θα καταφύγουμε σ' ένα απλό και προσφιλές τρυκ. Θα υπολογίσουμε τον πίνακα σε λίγες διαστάσεις. Ο σκοπός της άσκησης πέρα από τον στείρο υπολογισμό και την πολύτιμη γεωμετρική ερμηνεία είναι να μας δώσει και κάποια αυτοπεποίθηση στον χειρισμό αυτών των πολύ σπουδαίων πινάκων.

Ένας δεσμός Στην περίπτωση που έχουμε μια λίστα που αποτελείται από ένα δεσμό, έστω $\boldsymbol{d}_1$ τότε εφαρμόζοντας την τεχνική Gramm - Schmidt έχουμε:

$$\boldsymbol{u}_1 = \boldsymbol{d}_1 \tag{3.27}$$

$$\boldsymbol{v}_1 = \boldsymbol{d}_1 \tag{3.28}$$

Παρ' όλο που δεν προκύπτει τίποτα το άμεσα ενδιαφέρον κρατάμε την πληροφορία και προσθέτουμε ένα δεσμό ακόμα στην άσκηση για να την κάνουμε λίγο πιο παραγωγική.

Δύο δεσμοί Έστω η λίστα που αποτελείται από δύο δεσμούς $\{\boldsymbol{d}_1, \boldsymbol{d}_2\}$, τότε έχουμε:

$$\boldsymbol{u}_1 = \boldsymbol{d}_1 \tag{3.29}$$

$$\boldsymbol{u}_2 = \boldsymbol{d}_2 - \frac{\boldsymbol{d}_2^\tau \boldsymbol{u}_1}{\boldsymbol{u}_1^2}\boldsymbol{u}_1 \tag{3.30}$$

$$\boldsymbol{v}_2 = \boldsymbol{d}_2 - \frac{\boldsymbol{d}_2^\tau \boldsymbol{d}_1}{\boldsymbol{d}_1^2}\boldsymbol{d}_1 \tag{3.31}$$

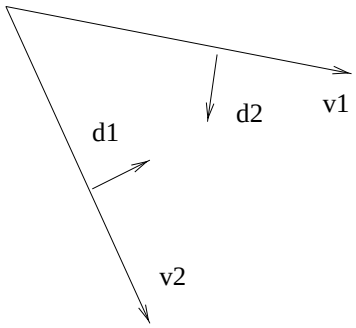
Αν θέλουμε τώρα να υπολογίσουμε και το $\boldsymbol{v}_1$ θα πρέπει να εναλλάξουμε τα στοιχεία της λίστας μας σε $\{\boldsymbol{d}_2, \boldsymbol{d}_1\}$, οπότε και έχουμε:

$$\boldsymbol{u}_1 = \boldsymbol{d}_2 \tag{3.32}$$

$$\boldsymbol{u}_2 = \boldsymbol{d}_1 - \frac{\boldsymbol{d}_1^\tau \boldsymbol{u}_1}{\boldsymbol{u}_1^2}\boldsymbol{u}_1 \tag{3.33}$$

$$\boldsymbol{v}_1 = \boldsymbol{d}_1 - \frac{\boldsymbol{d}_1^\tau \boldsymbol{d}_2}{\boldsymbol{d}_2^2}\boldsymbol{d}_2 \tag{3.34}$$

Κατ' αρχήν θα πρέπει να παρατηρήσουμε ότι τα $\boldsymbol{v}_1$ και $\boldsymbol{v}_2$ είναι πλήρως συμμετρικά, και για αυτό αναλλοίωτα σε αναδιατάξεις, ακριβώς όπως είχαμε παρατηρήσει προηγουμένως. Εκτός του ότι ικανοποιούν όλες τις συνθήκες ορθογωνιότητας που είχαμε θέσει η γεωμετρική τους ερμηνεία είναι τώρα προφανής. Τα διανύσματα $\boldsymbol{v}_i$ αποτελούν τις ακμές του πολυδιάστατου κώνου του θετικού ημιχώρου που παράγεται από την τομή των υπερεπιπέδων των δεσμών. Αυτό προκύπτει αν δούμε πως τα $\boldsymbol{v}_1$ και $\boldsymbol{v}_2$ είναι προβολές των $\boldsymbol{d}_1$ και $\boldsymbol{d}_2$ στο εναπομείναν διάνυσμα.



Σχήμα 3.4: Οι ακμές όπως απεικονίζονται στην περίπτωση των δύο διαστάσεων.

Εφαρμογή

Λήμμα 4 *Ο $M \times M$ πίνακας $\boldsymbol{V}^\tau \boldsymbol{V}$ έχει αντίστροφο τον $\boldsymbol{D}^\tau \boldsymbol{D}$, δηλαδή ισχύει:*

$$(\boldsymbol{V}^\tau \boldsymbol{V})(\boldsymbol{D}^\tau \boldsymbol{D}) = (\boldsymbol{D}^\tau \boldsymbol{D})(\boldsymbol{V}^\tau \boldsymbol{V}) = \boldsymbol{I} \tag{3.35}$$

Απόδειξη Ξεκινώντας από την εξίσωση 3.23, έχουμε:

$$\boldsymbol{D}^\tau \boldsymbol{V} = \boldsymbol{I} \Rightarrow \tag{3.36}$$
$$\boldsymbol{V}\boldsymbol{D}^\tau \boldsymbol{V} = \boldsymbol{V} \tag{3.37}$$

Παρατηρούμε ότι όταν ο τελεστής $\boldsymbol{V}\boldsymbol{D}^\tau$ δρα πάνω σε ένα οποιοδήποτε διάνυσμα που ανήκει στον υπόχωρο των M διαστάσεων το αφήνει αναλλοίωτο. Λέμε οποιοδήποτε, διότι κάθε διάνυσμα του υποχώρου μπορεί να γραφτεί σαν γραμμικός συνδυασμός των στηλών του $\boldsymbol{V}$, και αφού η βάση του υποχώρου δεν επηρεάζεται, τότε το ίδιο ισχύει για όλα τα διανύσματα του υποχώρου.

Ο πίνακας $\boldsymbol{V}\boldsymbol{D}^\tau$ (και ο $\boldsymbol{D}\boldsymbol{V}^\tau$ αντίστοιχα) παρ' όλο που είναι διάστασης $(N+1) \times (N+1)$ αποτελεί ένα είδος μοναδιαίου για τα διανύσματα του υποχώρου των M διαστάσεων. Ισχύει λοιπόν ότι:

$$\boldsymbol{V}\boldsymbol{D}^\tau \boldsymbol{D} = \boldsymbol{D} \Rightarrow \tag{3.38}$$
$$\boldsymbol{V}^\tau \boldsymbol{V}\boldsymbol{D}^\tau \boldsymbol{D} = \boldsymbol{V}^\tau \boldsymbol{D} = \boldsymbol{I} \tag{3.39}$$

Το αντίστροφο μπορεί να αποδειχτεί είτε με παρόμοιο τρόπο ξεκινώντας όμως από τη σχέση $\boldsymbol{V}^\tau \boldsymbol{D} = \boldsymbol{I}$, είτε προσπαθώντας να υπολογίσει κανείς τον ανάστροφο του $(\boldsymbol{V}^\tau \boldsymbol{V})(\boldsymbol{D}^\tau \boldsymbol{D})$.

3.1.5 Πολλαπλασιαστές Lagrange σε περισσότερες διαστάσεις

Ας προσπαθήσουμε να αναπαράγουμε πιο συστηματικά την τακτική που ακολουθήσαμε στη μονοδιάστατη περίπτωση σε περισσότερες διαστάσεις. Το πρόβλημα γράφεται:

$$\begin{cases} f = \frac{1}{2}(\boldsymbol{Q} - \boldsymbol{\Delta W})^2 \\ \boldsymbol{D}^\tau \boldsymbol{Q} \geq \boldsymbol{0} \end{cases} \tag{3.40}$$

Η μετατροπή των ανισοτικών δεσμών σε ισοτικούς γίνεται με την απαίτηση $\boldsymbol{q} = \boldsymbol{D}^\tau \boldsymbol{Q}, \boldsymbol{q} > \boldsymbol{0}$ και οι πολλαπλασιαστές Lagrange αποθηκεύονται πλέον στο συναλλοίωτο[7] διάνυσμα $\boldsymbol{\lambda}$. Έτσι, η εκτεταμένη

[7] το διάνυσμα $\boldsymbol{\lambda}$ είναι διάνυσμα γραμμής και όχι στήλης όπως το $\boldsymbol{q}$. Η διάκριση γίνεται για λόγους συνέπειας στο συμβολισμό, αλλά έχει και την πρακτική συνέπεια ότι υποδηλώνει ότι το $\boldsymbol{\lambda}$ ανήκει στο δυϊκό χώρο.

συνάρτηση γίνεται:

$$F = \frac{1}{2}(\boldsymbol{Q} - \boldsymbol{\Delta W})^2 + \boldsymbol{\lambda}(\boldsymbol{D}^\tau \boldsymbol{Q} - \boldsymbol{q}) \tag{3.41}$$

$$\boldsymbol{\nabla F} = (\boldsymbol{Q}^\tau - \boldsymbol{\Delta W}^\tau) + \boldsymbol{\lambda D}^\tau = 0 \tag{3.42}$$

κατ' αναλογία με την εξίσωση 3.15.

Προσπαθώντας να εμφανίσουμε το $\boldsymbol{q}$ στην εξίσωση πολλαπλασιάζουμε με το $\boldsymbol{D}$ από δεξιά. Έτσι έχουμε:

$$(\boldsymbol{Q}^\tau - \boldsymbol{\Delta W}^\tau)\boldsymbol{D} + \boldsymbol{\lambda D}^\tau \boldsymbol{D} = 0 \tag{3.43}$$

για να λύσουμε ως προς $\boldsymbol{\lambda}$ είναι απαραίτητο να πολλαπλασιάσουμε με το γινόμενο $\boldsymbol{V}^\tau \boldsymbol{V}$ που αντιπροσωπεύει τον αντίστροφο του γινομένου $\boldsymbol{D}^\tau \boldsymbol{D}$. Αντικαθιστώντας το $\boldsymbol{Q}^\tau \boldsymbol{D}$ με το $\boldsymbol{q}^\tau$ παίρνουμε:

$$\boldsymbol{\lambda} = -(\boldsymbol{q}^\tau - \boldsymbol{\Delta W}^\tau \boldsymbol{D})\boldsymbol{V}^\tau \boldsymbol{V} \tag{3.44}$$

ή

$$\boldsymbol{\lambda}^\tau = \boldsymbol{V}^\tau \boldsymbol{V}(\boldsymbol{D}^\tau \boldsymbol{\Delta W} - \boldsymbol{q}) \tag{3.45}$$

Δεδομένου ότι:

$$\boldsymbol{Q}^\tau = \boldsymbol{\Delta W}^\tau - (\boldsymbol{\Delta W}^\tau - \boldsymbol{q}^\tau)\boldsymbol{V}^\tau \boldsymbol{V} \boldsymbol{D}^\tau \tag{3.46}$$

αντικαθιστώντας το στην f παίρνουμε:

$$\begin{cases} f = \frac{1}{2}(\boldsymbol{\Delta W}^\tau \boldsymbol{D} - \boldsymbol{q}^\tau)\boldsymbol{V}^\tau \boldsymbol{V}(\boldsymbol{D}^\tau \boldsymbol{\Delta W} - \boldsymbol{q}) \\ \boldsymbol{q} \geq \boldsymbol{0} \end{cases} \tag{3.47}$$

Κατ' αναλογία με τη μονοδιάστατη περίπτωση η συνάρτηση f αντιστοιχεί σ' ένα ελλειπτικό παραβολοειδές, μια παραβολή δηλαδή που έχει σαν βάση μια έλλειψη. Η βάση του παραβολοειδούς είναι έλλειψη και όχι υπερβολή διότι:

- Ο πίνακας $\boldsymbol{V}^\tau \boldsymbol{V}$ είναι πραγματικός συμμετρικός και θετικά ορισμένος, άρα οι ιδιοτιμές του είναι θετικές.
- Στο αρχικό πρόβλημα η βάση του παραβολοειδούς ήταν μια υπερσφαίρα, δηλαδή ένα κλειστό σχήμα. Δεν είναι δυνατόν μ' έναν γραμμικό μετασχηματισμό να το αντικαταστήσουμε μ' ένα ανοιχτό.

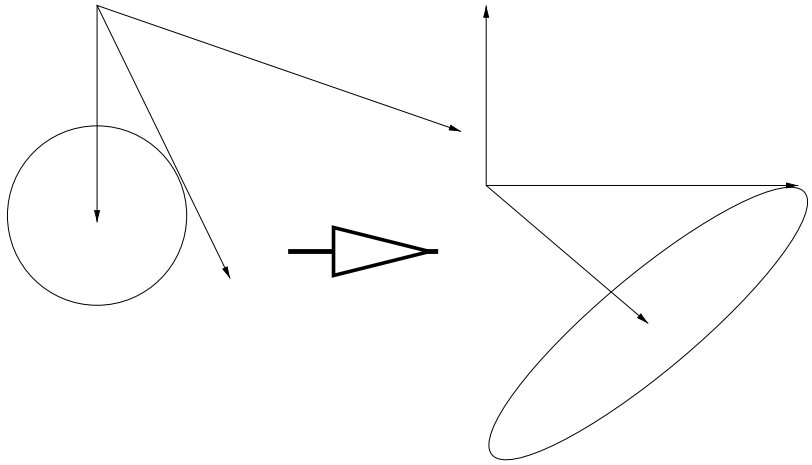
Το ζητούμενο στο οποίο καταλήξαμε είναι να βρούμε την ελάχιστη από τις όμοιες ελλείψεις[8] έτσι ώστε να έχει ένα τουλάχιστον σημείο στο θετικό τεταρτημόριο[9]. Είναι προφανές ότι ορθογωνιοποιήσαμε τους δεσμούς μας, αλλά στην προσπάθεια μετατρέψαμε τον κύκλο σε έλλειψη όπως φαίνεται και στο σχήμα 3.5. Το πρόβλημα αυτό δεν είναι ευκολότερο απ' ότι το προηγούμενο αλλά έχει το σημαντικό πλεονέκτημα ότι διατυπώνεται στις φυσικές του συντεταγμένες. Δηλαδή οι ανεξάρτητες μεταβλητές ($\boldsymbol{q}$) είναι M και όχι $N+1$, ενώ το σύστημα είναι πλέον ορθογώνιο. Αυτό δεν βελτιώνει καθόλου την κατάσταση από μόνο του γιατί ούτως ή άλλως δεν υπάρχει αναλυτικός τρόπος λύσης, αλλά μπορεί κανείς τώρα να δει ποια η σχέση του προβλήματος του τετραγωνικού προγραμματισμού με τις διάφορες τεχνικές συμπυκνούμενων ελλειψοειδών.

3.1.6 Ορθογώνιες συντεταγμένες

Εφ' όσον είναι φανερό ότι προσπαθώντας να εφαρμόσουμε τη μέθοδο των πολλαπλασιαστών Lagrange καταλήξαμε σε κάποιου είδους μετασχηματισμό σε ορθογώνιες συντεταγμένες, είναι σκόπιμο να τον μελετήσουμε λίγο πιο διεξοδικά. Παρατηρούμε λοιπόν ότι αφού απαιτούμε το $\boldsymbol{Q}$ να ανήκει στο πεδίο που ορίζεται από

[8] Ελλειψοειδή, στις περισσότερες από δύο διαστάσεις

[9] Ο όρος τεταρτημόριο χρησιμοποιείται καταχρηστικά εδώ θέλοντας να δηλώσουμε τον γεωμετρικό τόπο των σημείων που έχουν όλες τις συντεταγμένες τους θετικές αλλά έτσι που να το καταλάβει ο αναγνώστης.



Σχήμα 3.5: Οπτική αναπαράσταση του μετασχηματισμού σε ορθογώνιους δεσμούς.

την τομή των θετικών ημιχώρων, θα μπορούσαμε να το εκφράσουμε σαν θετικό γραμμικό συνδυασμό των ακμών αυτού του κώνου. Δηλαδή:

$$\boldsymbol{Q} = \sum_{i=1}^{M} q_i \frac{\boldsymbol{v}_i}{\boldsymbol{v}_i^2}, q_i > 0 \forall i \in \{1 \dots M\} \tag{3.48}$$

η σε πιο συμπαγή μορφή:

$$\boldsymbol{Q} = \boldsymbol{V}\boldsymbol{q}, \boldsymbol{q} > \boldsymbol{0} \tag{3.49}$$

Αν τώρα ξέρουμε το $\boldsymbol{Q}$ και θέλουμε να υπολογίσουμε το $\boldsymbol{q}$ αυτό γίνεται εύκολα αν πολλαπλασιάσουμε με τον από αριστερά αντίστροφο του $\boldsymbol{V}$, τον $\boldsymbol{D}^\tau$. Πράγμα που μας οδηγεί στη βασική σχέση που χρησιμοποιήσαμε για να μετατρέψουμε το ανισοτικό πρόβλημα σε ισοτικό, δηλαδή $\boldsymbol{q} = \boldsymbol{D}^\tau \boldsymbol{Q}$. Όπότε τώρα που έχουμε τον ευθύ και τον αντίστροφο μετασχηματισμό μπορούμε να μετατρέψουμε οποιοδήποτε διάνυσμα $\boldsymbol{X}$ των $N+1$ σε $\boldsymbol{x}$ του υποχώρου των M διαστάσεων.

$$\boldsymbol{X} = \boldsymbol{V}\boldsymbol{x} \Leftrightarrow \boldsymbol{x} = \boldsymbol{D}^\tau \boldsymbol{X} \tag{3.50}$$

Είναι μάλλον προφανές ότι αν αντικαταστήσουμε το $\boldsymbol{Q}$ στην αρχική f θα καταλήξουμε στο πρόβλημα 3.47.

Είναι σημαντική η παρατήρηση στο ευθύ τμήμα του μετασχηματισμού, ότι ενώ το παραγόμενο διάνυσμα $\boldsymbol{X}$ είναι $N+1$ διαστάσεων, στην πραγματικότητα άπτεται πλήρως στον υπόχωρο τον M διαστάσεων. Αυτό γίνεται αμέσως αντιληπτό αν αναλογιστούμε ότι μιλάμε για ένα θετικό γραμμικό συνδυασμό των M διανυσμάτων $\boldsymbol{v}_i$, που είναι γραμμικώς ανεξάρτητα και αποτελούν συνεπώς πλήρη βάση του υποχώρου. Έτσι λοιπόν το κομμάτι του $\boldsymbol{Q}$ που είναι κάθετο στον υπόχωρο δεν είναι δυνατόν να διατηρηθεί μετά από το μετασχηματισμό. Ευτυχώς είναι τέτοια η δομή του προβλήματος, που σύμφωνα με τον ορισμό του, το $\boldsymbol{Q}$ δεν έχει κάθετη συνιστώσα στον υπόχωρο (βλ. 3.2), και συνεπώς μπορούμε να εκτελούμε το μετασχηματισμό κατά βούληση χωρίς να μας απασχολεί τυχόν απώλεια πληροφορίας.

3.1.7 Συνθήκες τερματισμού

Στην πραγματικότητα υπάρχει ακόμα ένας μετασχηματισμός που μετασχηματίζει το πρόβλημα σε ορθογώνιες συντεταγμένες. Ο χώρος στον οποίο οδηγεί είναι ο δυϊκός αυτού που καταλήξαμε στο προηγούμενο κεφάλαιο. Για να μπορέσουμε να εστιάσουμε στην ουσία της δυϊκότητας του χώρου, αλλά και για να εκμεταλλευτούμε αυτή την πληροφορία κατάλληλα, θα πρέπει να κατανοήσουμε πρώτα τα κριτήρια τερματισμού της λύσης του τετραγωνικού προγράμματος.

Η εξίσωση 3.45 ουσιαστικά μας λέει ότι:

$$\boldsymbol{\lambda}^{\tau} = -\nabla f \tag{3.51}$$

Είναι φανερό ότι αν το πρόβλημα δεν είχε περιορισμούς, τότε αν εκκινούσαμε το διάνυσμα $\boldsymbol{q}$ από επιτρεπτή θέση ($\boldsymbol{q} \geq \mathbf{0}$), θα μπορούσαμε να μετακινήσουμε το διάνυσμα θέσης $\boldsymbol{q}$ κατά τη διεύθυνση του $\boldsymbol{\lambda}^{\tau}$[10]. έτσι ώστε να πετύχουμε μείωση της συνάρτησης κόστους. Από τη στιγμή που υπάρχουν περιορισμοί είναι σαφές ότι αυτή η κίνηση θα είναι επιτρεπτή αν και εφ' όσον οι δεσμοί δεν παραβιάζονται.

Η αλγοριθμική διατύπωση της παραπάνω πρότασης άλλωστε αποτελεί και τη βάση όλων των γνωστών αλγορίθμων ενεργών δεσμών. Αυτό σημαίνει ότι θα επιτρέπεται η κίνηση, κατά μήκος της τυχούσας συνιστώσας i, κατά λ_i αν συμβαίνει ένα από τα δυο τινά:

- $q^i > 0$ άρα ο δεσμός δεν είναι ενεργός και συνεπώς μπορούμε να κινηθούμε προς ή από αυτόν.
- $q^i = 0$ και $\lambda_i > 0$ ο δεσμός είναι μεν ενεργός αλλά ακολουθώντας το λ_i δεν τον παραβιάζουμε αφού απομακρυνόμαστε απ' αυτόν προς το εσωτερικό της επιτρεπτής περιοχής ($\boldsymbol{q} \geq \mathbf{0}$).

Είναι τώρα μάλλον προφανές το πότε θα σταματήσει αυτή η διαδικασία. Θα πρέπει για κάθε ορθογώνια συνταγμένη i να ισχύει ένα από τα δύο:

- $q^i = 0$ και $\lambda_i < 0$ που σημαίνει ότι ο δεσμός i είναι ενεργός συμμετέχει στη τελική λύση και το σύστημα δεν έχει την τάση να απομακρυνθεί απ' αυτόν.
- $q^i > 0$ και $\lambda_i = 0$ για να σταματήσει το σύστημα να κινείται ως προς αυτή τη διεύθυνση.

Στο επιθυμητό ακρότατο θα έχουμε δηλαδή $\forall i$:

$$q^i \lambda_i = 0 \tag{3.52}$$
$$q^i \geq 0 \tag{3.53}$$
$$\lambda_i \leq 0 \tag{3.54}$$

Στο ίδιο συμπέρασμα καταλήγουμε αν ξεκινήσουμε από την εξίσωση 3.45 και πολλαπλασιάσουμε από αριστερά με $\boldsymbol{q}^{\tau}$. Δηλαδή έχουμε:

$$\boldsymbol{q}^{\tau}\boldsymbol{\lambda}^{\tau} = \boldsymbol{q}^{\tau}\boldsymbol{V}^{\tau}\boldsymbol{V}(\boldsymbol{D}^{\tau}\boldsymbol{\Delta W} - \boldsymbol{q}) \tag{3.55}$$
$$\boldsymbol{q}^{\tau}\boldsymbol{\lambda}^{\tau} = \boldsymbol{Q}^{\tau}(\boldsymbol{\Delta W} - \boldsymbol{Q}) \tag{3.56}$$
$$\boldsymbol{q}^{\tau}\boldsymbol{\lambda}^{\tau} = 0 \tag{3.57}$$

με την διαφορά ότι δεν εξάγεται η πληροφορία για τα πρόσημα των δυϊκών μεταβλητών $\boldsymbol{q}$ και $\boldsymbol{\lambda}$. Ενώ, για το μεν $\boldsymbol{q}$, η απαίτηση $\boldsymbol{q} \geq \mathbf{0}$ είναι ζητούμενο του προβλήματος, η πληροφορία ότι στο σημείο που αποτελεί λύση του τετραγωνικού προγράμματος ισχύει $\boldsymbol{\lambda}^{\tau} \leq \mathbf{0}$ προέρχεται από μια διαισθητική ανάλυση. Δεν θα ζητούσαμε από τον αναγνώστη να εμπιστευτεί τη διαίσθησή μας αν αυτή δεν συνέπιπτε με τα κριτήρια βελτιστοποίησης των Kuhn - Tucker οι οποίοι ευτυχώς έχουν δώσει μια πολύ πιο αυστηρή απόδειξη στο γενικότερο πρόβλημα για τις συνθήκες ελαχιστοποίησης συναρτήσεων κάτω από δεσμούς.

Στην πραγματικότητα οι συνθήκες Kuhn - Tucker αποτελούν εγγύηση ότι η λύση ανήκει σ' ένα πολύ πιο περιορισμένο πεδίο απ' αυτό της επιτρεπτής περιοχής του υποχώρου των M διαστάσεων. Οι συνθήκες των Kuhn - Tucker μας επισημαίνουν, αν και κάπως έμμεσα, ότι η λύση είναι προβολή του $\boldsymbol{\Delta W}$ πάνω σε μια άγνωστη μεν, συγκεκριμένη δε, λίστα δεσμών. Το πως συνάγεται αυτό το συμπέρασμα δεν είναι άμεσα αντιληπτό, αν και ξέρουμε ότι το $\boldsymbol{Q}$ είναι προβολή του $\boldsymbol{\Delta W}$.

[10] Η παράλειψη της αναφοράς στον συντελεστή βήματος είναι σκόπιμη, μια και η αναλυτική διατύπωση ενός αλγορίθμου δεν είναι ο σκοπός αυτού του κεφαλαίου.

Είναι σημαντικό να συνειδητοποιήσουμε ότι τα μοναδικά αντικείμενα που υπάρχουν στον εξεταζόμενο υπόχωρο είναι τα υπερεπίπεδα των δεσμών. Είναι λογικό λοιπόν να θεωρήσουμε ότι η λύση θα ανήκει σε κάποιο συνδυασμό αυτών. Ακόμα και αν δεν άπτεται σε κανένα από τα υπερεπίπεδα ($\boldsymbol{d}_i^\tau \boldsymbol{\Delta W} > \boldsymbol{0}\, \forall i \in \mathcal{L} \Rightarrow \boldsymbol{Q} = \boldsymbol{\Delta W}$) πάλι μπορούμε να θεωρήσουμε ότι πρόκειται για το συνδυασμό που περιέχει μηδενικό αριθμό δεσμών.

Αν απαιτήσουμε από το $\boldsymbol{Q}$ να ισούται με τη προβολή του $\boldsymbol{\Delta W}$ πάνω σε μια αυθαίρετη λίστα δεσμών $\mathcal{L}'$, σύμφωνα με την τεχνική που περιγράψαμε στο κεφάλαιο 3.1.4 έχουμε:

$$\boldsymbol{Q} = \boldsymbol{\Delta W} - \sum_{i \in \mathcal{L}'} \frac{\boldsymbol{\Delta W}^\tau \boldsymbol{u}_i}{\boldsymbol{u}_i^2} \boldsymbol{u}_i \tag{3.58}$$

όπου τα $\boldsymbol{u}_i$ υπολογίζονται αναδρομικά από τη σχέση

$$\boldsymbol{u}_i = \boldsymbol{d}_i - \sum_{j=1}^{i-1} \frac{\boldsymbol{d}_i^\tau \boldsymbol{u}_j}{\boldsymbol{u}_j^2} \boldsymbol{u}_j \tag{3.59}$$

Αυτό το $\boldsymbol{Q}$ που αποτελεί περικοπή του $\boldsymbol{\Delta W}$ κατά Gramm - Schmidt ικανοποιεί τις εξής σχέσεις από κατασκευής:

$$\boldsymbol{d}_i^\tau \boldsymbol{Q} = q_i = 0\, \forall i \in \mathcal{L}' \tag{3.60}$$

$$\frac{\boldsymbol{v}_i^\tau}{\boldsymbol{v}_i^2} (\boldsymbol{\Delta W} - \boldsymbol{Q}) = \lambda_i = 0\, \forall i \ni \mathcal{L}' \tag{3.61}$$

Αυτό σημαίνει ότι δεδομένης της τελικής ενεργής λίστας δεσμών $\mathcal{L}'$ το σημείο που ελαχιστοποιεί την συνάρτηση κόστους είναι το κομμένο κατά Gramm - Schmidt $\boldsymbol{\Delta W}$, και επειδή το $\boldsymbol{Q}$ πρέπει να άπτεται πάνω σε κάποιους δεσμούς (ή και κανένα $\mathcal{L}' = \emptyset$) η λύση $\boldsymbol{Q}$ μπορεί πάντα να γραφεί σαν προβολή του $\boldsymbol{\Delta W}$ πάνω στη λίστα δεσμών $\mathcal{L}'$.

Η τελική λίστα $\mathcal{L}'$ μπορεί να είναι άγνωστη αλλά δεν παύει να είναι κάποιο υποσύνολο της αρχικής λίστας εισόδου $\mathcal{L}$. Αξίζει να σημειωθεί ότι η διάταξη των στοιχείων της λίστας δεν παίζει κανένα ρόλο στην εξαγωγή της λύσης $\boldsymbol{Q}$ που είναι μια διαδικασία που στοιχίζει $(N+1)M'^2$ πράξεις, όπου M' είναι το πλήθος των στοιχείων της $\mathcal{L}'$.

Αν θέλουμε να περιγράψουμε το πρόβλημα σαν αλγοριθμικό μπλοκ με όρους εισόδου - εξόδου τότε θα κατατάσσαμε σαν είσοδο τη λίστα των δεσμών $\mathcal{L}$ των M δεσμών, και σαν έξοδο τη λίστα $\mathcal{L}'$ με τα M' στοιχεία, όπου βέβαια $\mathcal{L}' \subseteq \mathcal{L}$. Θα μπορούσαμε να χρησιμοποιήσουμε την τεχνική της εξαντλητικής αναζήτησης, αλλά δυστυχώς το πλήθος των περιπτώσεων, που ανέρχεται σε 2^M συνδυασμούς, είναι απαγορευτικό. Η γνώση όμως τελικά ότι η λύση είναι μια προβολή του $\boldsymbol{\Delta W}$ σε έναν από τους 2^M δυνατούς συνδυασμούς θα αποδεχτεί το σημείο κλειδί στο σχεδιασμό ενός πρωτότυπου και αποδοτικού αλγορίθμου.

Ο δυϊκός χώρος

Θα μπορούσαμε να είχαμε επιλέξει έναν άλλο μετασχηματισμό με το διάνυσμα $\boldsymbol{\lambda}^\tau$ σαν πρωτεύουσα μεταβλητή αντί του $\boldsymbol{q}$. Αν ορίσουμε δηλαδή τον μετασχηματισμό:

$$\boldsymbol{\lambda}^\tau = \boldsymbol{V}^\tau (\boldsymbol{\Delta W} - \boldsymbol{Q}) \Leftrightarrow \boldsymbol{Q} = \boldsymbol{\Delta W} - \boldsymbol{D}\boldsymbol{\lambda}^\tau \tag{3.62}$$

μπορούμε να γράψουμε την f σαν

$$\begin{cases} f = \frac{1}{2} \boldsymbol{\lambda} \boldsymbol{D}^\tau \boldsymbol{D} \boldsymbol{\lambda}^\tau \\ \boldsymbol{\lambda}^\tau \leq \boldsymbol{V}^\tau \boldsymbol{\Delta W} \end{cases} \tag{3.63}$$

Το πρόγραμμα που καλούμαστε να επιλύσουμε, είναι αντίστοιχο με το 3.47, αφού αντιπροσωπεύει στην πραγματικότητα το δυϊκό του. Ενδιαφέρουσα είναι και η παρατήρηση ότι ενώ προηγουμένως το κέντρο της έλλειψης ήταν μετατοπισμένο, τώρα έχουμε μετατοπισμένους τους άξονες του προβλήματος.

3.2 Προηγούμενες εργασίες

Τα προβλήματα βελτιστοποίησης απασχόλησαν κυρίως μεταπολεμικά τη διεθνή βιβλιογραφία. Οι περισσότερες δημοσιεύσεις αφορούσαν στη Simplex, αλλά όπως ήταν αναμενόμενο ένα σημαντικό μέρος ερευνητών προσπαθούσαν να λύσουν γενικότερα προβλήματα μη γραμμικού προγραμματισμού.

Στο παρόν έχουμε επιλέξει τις μεθόδους των Rosen και Zoutendijk για να μας βοηθήσουν να εντοπίσουμε τις ομοιότητες αλλά και τις διαφορές της κλασσικής θεώρησης με την παρούσα. Επειδή και οι δύο μέθοδοι επιδιώκουν να λύσουν προβλήματα βελτιστοποίησης κάτω από γραμμικούς δεσμούς, είναι πιθανό να μπορούν να εφαρμοστούν τόσο στο τετραγωνικό υποπρόβλημα που είναι το θέμα αυτού του κεφαλαίου, όσο και στο αρχικό γραμμικό πρόβλημα ανισοτήτων όπως αυτό ορίζεται στο προηγούμενο κεφάλαιο. Έτσι θα περιγράψουμε όλα τα πιθανά σενάρια, αναλύοντας τα πιθανά πλεονεκτήματα και μειονεκτήματα της κάθε περίπτωσης.

3.2.1 Rosen

Ο Rosen πρότεινε τη μέθοδο της προβολής του gradient (gradient projection method) το 1960 [38] και ένα χρόνο αργότερα τη γενίκευσε ώστε να συμπεριλαμβάνει και μη γραμμικούς δεσμούς [39]. Η μέθοδος εισάγει τον πίνακα προβολής. Ο πίνακας προβολής είναι ένας πίνακας που προβάλλει ένα οποιοδήποτε διάνυσμα από τον πλήρη χώρο των N διαστάσεων στον μηδενόχωρο των διανυσμάτων των ενεργών δεσμών. Η κεντρική ιδέα και εδώ είναι να χρησιμοποιεί αυτόν τον πίνακα για να προβάλει το $\boldsymbol{\Delta W}$ σε μια επιτρεπτή διεύθυνση.

Αν αγνοήσουμε τις διαφορές στον συμβολισμό ανάμεσα στην παρούσα διατριβή και στη διεθνή βιβλιογραφία, εύκολα συνάγεται ότι ο πίνακας προβολής του Rosen δίνεται από:

$$\boldsymbol{\mathcal{P}} = \boldsymbol{I} - \boldsymbol{D}\boldsymbol{V}^{\tau} = \boldsymbol{I} - \boldsymbol{V}\boldsymbol{D}^{\tau} \tag{3.64}$$

Ο αλγόριθμος του Rosen αρκείται στον υπολογισμό μίας επιτρεπτής διεύθυνσης και όχι της καλύτερης δυνατής επιτρεπτής διεύθυνσης. Στην περίπτωση που ο πίνακας $\boldsymbol{D}^{\tau}$ είναι τετραγωνικός, έχουμε δηλαδή $N+1$ ενεργούς δεσμούς, αυτό σημαίνει ότι όλες οι προβολές καταλήγουν στο μηδενικό διάνυσμα. Ο αλγόριθμος του Rosen προνοεί για μια τέτοια περίπτωση και διαφεύγει από την βέλτιστη ακμή, ή αν αυτό δεν είναι δυνατόν τερματίζει, αφού αποδεικνύεται εύκολα ότι το συγκεκριμένο σημείο, είναι σημείο που ικανοποιεί τις συνθήκες Kuhn - Tucker.

Το πρόβλημα της γραμμικής διαχωρισιμότητας

Στην περίπτωση που κάποιος προσπαθεί να λύσει το πρόβλημα της γραμμικής διαχωρισιμότητας με τη μέθοδο του Rosen, τότε θα αναπαράγει την ακολουθία σημείων που παράγει και η μέθοδος του Bobrowski. Είναι γεγονός ότι και οι δύο αλγόριθμοι χειρίζονται τη λίστα των ενεργών δεσμών με τον ίδιο τρόπο. Δεν έχουν την πολυτέλεια να αποφανθούν για τον βέλτιστο συνδυασμό δεσμών με αποτέλεσμα να αλλάζουν την τρέχουσα λίστα τους μόνο κατά ένα δεσμό. Ο δεσμός που ο αλγόριθμος εγκαταλείπει βρίσκεται απέναντι από την ακμή που σχηματίζει την μικρότερη γωνία με το $\boldsymbol{\Delta W}$.

Δεδομένου ότι ο Rosen προηγείται του Bobrowski κατά 24 χρόνια η σωστή διατύπωση θα ήταν μάλλον ότι ο αλγόριθμος του Bobrowski συμπεριφέρεται ακριβώς σαν τον αλγόριθμο του Rosen. Όπως έχουμε ξαναπεί η επιστημονική συνεισφορά του Bobrowski έγκειται στο γεγονός ότι μπόρεσε να συνδέσει τον κανόνα ανανέωσης του off-line perceptron με ένα πρόβλημα βελτιστοποίησης. Από τη στιγμή που μπόρεσε και απέδειξε σύγκλιση σε πεπερασμένο αριθμό βημάτων, καταργώντας ουσιαστικά το μοναδικό πλεονέκτημα της on-line μάθησης, έδωσε το έναυσμα για εμπλοκή σχεδόν οποιουδήποτε αλγόριθμου βελτιστοποίησης στα νευρωνικά δίκτυα. Θα ήταν πραγματικά παράξενο αν ο Bobrowski κατέληγε σ' έναν αλγόριθμο που κανείς δεν είχε σκεφτεί, σε έναν τόσο διαδεδομένο, και με τόσο μεγάλη πρακτική αξία κλάδο των μαθηματικών όπως η βελτιστοποίηση.

Το τετραγωνικό υποπρόβλημα

Ας υποθέσουμε τώρα ότι προσπαθούμε να χρησιμοποιήσουμε τον αλγόριθμο του Rosen για την επίλυση του τετραγωνικού υποπροβλήματος. Ο κανόνας του Rosen όπως τον έχουμε διατυπώσει αλλά και όπως τον έχουμε δει διατυπωμένο [40, 41, 42, 43, 44] δεν επιτρέπει τη λύση αυτού του προβλήματος, πολύ απλά επειδή δεν επιτρέπει τον αποχαρακτηρισμό περισσότερου του ενός ενεργού δεσμού, και αυτό μόνο με την προϋπόθεση ότι το πλήθος των ενεργών δεσμών έχει φτάσει τη διάσταση του προβλήματος.

Στην πραγματικότητα ο αλγόριθμος του Rosen αποχαρακτηρίζει έναν δεσμό όταν η προβολή του gradient στον μηδενόχωρο των ενεργών δεσμών γίνει 0. Αυτό πρακτικά συμβαίνει όταν:

- όταν ο αλγόριθμος φτάσει στο ελάχιστο της συνάρτησης στον υπόχωρο των ενεργών δεσμών.
- όταν το πλήθος των ενεργών δεσμών έχει φτάσει το πλήθος των διαστάσεων του αρχικού χώρου.

Στη γενική περίπτωση, με μια τυχούσα αντικειμενική συνάρτηση, ο αλγόριθμος του Rosen δεν μπορεί ποτέ να ικανοποιήσει την πρώτη συνθήκη[11] λόγω εκτεταμένου zigzgging. Επειδή όμως ξέρουμε ότι το ελάχιστο της συγκεκριμένης συνάρτησης στον υπό εξέταση υπόχωρο βρίσκεται στην προβολή του $\boldsymbol{\Delta W}$ στον συγκεκριμένο μηδενόχωρο, ο αλγόριθμος του Rosen χρειάζεται μόνο μια επανάληψη για να συγκλίνει στο ελάχιστο του υπόχωρου που ορίζεται από τους ενεργούς δεσμούς. Πράγματι στο κεφάλαιο 3.3.5 προτείνουμε μια ανάλογη τεχνική για την επίλυση του προβλήματος σε πεπερασμένο αριθμό βημάτων.

3.2.2 Zoutendijk

Ο Zoutendijk ήταν από τους πρώτους (1960), [37, 45] που διέκρινε τα πιθανά οφέλη από τον υπολογισμό της βέλτιστης επιτρεπτής διεύθυνσης. Συγκεκριμένα ο Zoutendijk επέμεινε στον υπολογισμό της βέλτιστης επιτρεπτής διεύθυνσης μέσω ενός πλήθους διαφορετικών κανονικοποιήσεων. Αυτό έχει σαν αποτέλεσμα την παραγωγή διαφορετικών διαδικασιών ανάλογα με την επιλεχθείσα κανονικοποίηση. Ο Zoutendijk παρουσίασε 5 διαφορετικές κανονικοποιήσεις και 3 διαφορετικές anti - zigzagging τεχνικές. Στο παρόν παρουσιάζουμε ένα υποσύνολο αυτών, σε μια πιο απλοποιημένη έκδοση από του Zoutendijk, αφού δεν μας απασχολούν μη γραμμικοί δεσμοί, ή περιορισμοί κουτιού (box constraints).

Ας υποθέσουμε ότι θέλουμε να ελαχιστοποιήσουμε μια συνάρτηση F, και ότι με $\boldsymbol{d}_p$ δίνονται τα διανύσματα των ενεργών δεσμών κατά την τρέχουσα επανάληψη. Ο στόχος του Zoutendijk είναι να βρει ένα διάνυσμα διεύθυνσης $\boldsymbol{P}$ που να ελαχιστοποιεί το γινόμενο $\nabla F^\tau \boldsymbol{P}$, ώστε να το χρησιμοποιήσει σαν διεύθυνση αναζήτησης. Είναι φανερό ότι η λύση του προβλήματος εκτείνεται στο άπειρο, και άρα κάποιου είδους περιορισμός είναι απαραίτητος. Ο Zoutendijk πρότεινε 5, εκ των οποίων οι δύο πιο σημαντικοί, κατά τη γνώμη μας, είναι:

$$\begin{cases} min \nabla F^\tau \boldsymbol{P} \\ \boldsymbol{d}_p^\tau \boldsymbol{P} \geq 0 \\ \boldsymbol{P}^\tau \boldsymbol{P} \leq 1 \end{cases} \tag{3.65}$$

και

$$\begin{cases} min \nabla F^\tau \boldsymbol{P} \\ \boldsymbol{d}_p^\tau \boldsymbol{P} \geq 0 \\ -1 \leq \boldsymbol{P}_j \leq 1 \end{cases} \tag{3.66}$$

Αμέσως μπορεί να δει κανείς ότι το πρόβλημα της 3.65 είναι ισοδύναμο με το 3.2, ενώ το πρόβλημα 3.66 δεν έχει ισοδύναμο στην μέχρι τώρα συζήτησή μας. Στην πραγματικότητα το πρόγραμμα της 3.66, αλλά και όλες οι υπόλοιπες προτεινόμενες κανονικοποιήσεις, δεν λύνουν το πρόβλημα που θέλουμε να λύσουμε, με την έννοια ότι δεν βρίσκουν την βέλτιστη επιτρεπτή διεύθυνση, αλλά μια απλή επιτρεπτή διεύθυνση. Δεδομένης της εμμονής του Zoutendijk στη μέθοδο Simplex είναι σχεδόν σίγουρο ότι επινοήθηκαν διότι επέτρεπαν την χρησιμοποίηση της Simplex ή κάποιας παραλλαγής της, για τη λύση τους. Αντίθετα όμως

[11] Αυτό ισχύει και για την περίπτωση του γραμμικού προβλήματος για άλους λόγους

το πρόβλημα 3.65 έχει έναν μη γραμμικό δεσμό με αποτέλεσμα να μην μπορεί να εφαρμοστεί η Simplex, όχι άμεσα τουλάχιστον.

Ο ίδιος ο Zoutendijk δηλώνει μέσα από τις σελίδες του βιβλίου του, στο σημείο που κάνει μια σύγκριση μεταξύ των κανονικοποιήσεων:

> Είναι πολύ δύσκολο να κάνει κανείς μια σύγκριση των πιθανών κανονικοποιήσεων, χωρίς υπολογιστική εμπειρία (1960).

και λίγο παρακάτω όταν μιλάει για την κανονικοποίηση που οδηγεί στο πρόβλημα 3.65:

> Θα πρέπει να οδηγεί σε μικρότερο αριθμό επαναλήψεων, σε σχέση με τις άλλες κανονικοποιήσεις, διότι δίνει την επιτρεπτή διεύθυνση που ελαχιστοποιεί τη γωνία με το gradient.

Είναι περιττό να θυμίσουμε πόσο σύμφωνους μας βρίσκει αυτή η δήλωση του Zoutendijk.

Το πρόβλημα της γραμμικής διαχωρισιμότητας

Ο Zoutendijk αφιερώνει ένα ολόκληρο κεφάλαιο στο βιβλίο του αναφέροντας πώς θα μπορούσε να επιτευχθεί ο υπολογισμός ενός επιτρεπτού σημείου. Δυστυχώς όμως όλες οι μέθοδοι που προτείνει βασίζονται στη Simplex, με τα γνωστά προβλήματα που αυτό συνεπάγεται.

Αν για χάρη της συζήτησης υποθέσουμε ότι χρησιμοποιούμε την μερικώς γραμμική (piecewise linear) συνάρτηση του perceptron, σαν συνάρτηση ελαχιστοποίησης, ακριβώς όπως κάνει ο Bobrowski, και υπήρχε μια μέθοδος που να λύνει αξιόπιστα το τετραγωνικό υποπρόβλημα, τότε η μέθοδος του Zoutendijk όπως ορίζεται από το πρόβλημα 3.65 θα ήταν ισοδύναμη με τον **FLF**, με την έννοια ότι και οι δύο κινούνται κατά τη βέλτιστη επιτρεπτή διεύθυνση.

Το τετραγωνικό υποπρόβλημα

Ας υποθέσουμε τώρα ότι επιχειρούμε να λύσουμε το τετραγωνικό υποπρόβλημα με τη μέθοδο του Zoutendijk. Μπορεί κανείς να δει ότι με τον τρόπο που υποτίθεται ότι ο αλγόριθμος θα βγάζει δεσμούς από τη λίστα των ενεργών δεσμών πολύ σύντομα θα βρεθεί αντιμέτωπος με το ίδιο πρόβλημα που προσπαθεί να λύσει, αλλά σε μικρότερη διάσταση. Είναι ένα πρόβλημα που αναπαράγει τον εαυτό του, και συνεπώς θα πρέπει να ακολουθήσει κανείς μια αναδρομική διαδικασία μέχρι να φτάσει σε μονοδιάστατα προβλήματα, όπου η λύση είναι προφανής όπως δείξαμε στο κεφάλαιο 3.1.3.

Μια τέτοια ιδέα φτάνει πολύ μακρύτερα από την αρχική σύλληψη του Zoutendijk ο οποίος φυσικά προτείνει και για το τετραγωνικό υποπρόβλημα τη μέθοδο Simplex. Το κόστος μιας τέτοιας υλοποίησης είναι μεγάλο διότι απαιτεί $O(N^3)$ μνήμη, ενώ το υπολογιστικό κόστος παραμένει μη υπολογίσιμο, αφού το συνολικό πλήθος των επαναλήψεων δεν είναι φραγμένο από κάποιο γνωστό αριθμό, αν και είναι πεπερασμένο. Το πρόβλημα της αναδρομής μπορεί να αντιμετωπισθεί με τη ορθογωνιοποίηση των δεσμών, σύμφωνα με τη μέθοδο που προτείνουμε στο κεφάλαιο 3.3.4, ο οποίος εισάγει επιπλέον και μια anti - zigzagging τεχνική (διπλή αναζήτηση) για επιτάχυνση της διαδικασίας σύγκλισης.

3.3 Μια καινούρια προσέγγιση

3.3.1 Ιστορική αναδρομή

Η μέχρι τώρα παρουσίαση του θέματος είναι αναλυτική και συμπαγής, με την έννοια ότι δεν αφήνει περιθώρια αμφιβολίας, αφού τεκμηριώνει τα επιμέρους θέματα τόσο με γεωμετρική ερμηνεία, όσο και με μαθηματικές κατασκευές. Κατά τη διάρκεια της ανάπτυξης αυτής της εργασίας όμως το θεωρητικό υπόβαθρο δεν ήταν πάντα τόσο σταθερό.

Το αρχικό πρόβλημα από το οποίο ξεκίνησε η παρούσα εργασία, ήταν να κατασκευαστεί ένα perceptron που όμως θα εκτελούσε σφαιρικό διαχωρισμό και όχι γραμμικό. Μια τέτοια κατασκευή έχει πολλές πρακτικές

εφαρμογές στον εντοπισμό και στη μοντελοποίηση συσσωματωμάτων (clusters). Το πρόβλημα αυτό λύθηκε με την μεταφορά του μοντέλου του Rosenblatt σε μια σφαιρική συνάρτηση βάσης σε συνάρτηση με έναν ακριβή τρόπο υπολογισμού του βέλτιστου βήματος. Σ' αυτή τη φάση οι συνάψεις του δικτύου έπαιζαν το ρόλο των κέντρων της σφαίρας, και το βάρος κατωφλιού (threshold) αντιπροσώπευε την ακτίνα του κύκλου (ή το τετράγωνο της ακτίνας).

Μετά από τα πρώτα καλά πειραματικά αποτελέσματα σε τεχνητά προβλήματα, έπρεπε ν' αντιμετωπιστεί η προφανής γενίκευση της μετατροπής της επιφάνειας διαχωρισμού από σφαιρική σε ελλειψοειδή. Η συνάρτηση βάσης δεν ήταν πρόβλημα, ούτε και ο κανόνας ανανέωσης των βαρών. Όμως το βέλτιστο βήμα ήταν διαφορετικό για κάθε διαφορετικού τύπου σύναψη (κέντρο της έλλειψης, ή ημιάξονας). Αυτό σήμαινε ουσιαστικά ότι η μέγιστη πτώση δεν βρισκόταν κατά την διεύθυνση του gradient, αλλά κατά μήκος μιας παραλλαγής του. Το γεγονός αυτό από μόνο του ήταν αρκετά ενοχλητικό, και παρ' όλο που το ελλειπτικό perceptron δεν είχε τόσο καλή απόδοση όσο το κυκλικό χρειαζόταν ένα άλλο έναυσμα για να σηματοδοτήσει το ότι η συγκεκριμένη κατεύθυνση οδηγούσε σε αδιέξοδο.

Το έναυσμα ήρθε με την μορφή του ακόλουθου πολύ απλού προβλήματος. Αν σ' ένα πεπερασμένο χώρο ορίσουμε ένα γραμμικά διαχωρίσιμο πρόβλημα, το κυκλικό perceptron θα έπρεπε να μπορεί να το λύσει μεταφέροντας του κέντρο του κύκλου μακριά από την διαχωριστική γραμμή, και μεγαλώνοντας την ακτίνα έτσι ώστε να συμπεριλάβει μόνο τα σημεία της μιας κατηγορίας. Αυτή ήταν και η συμπεριφορά του για προβλήματα με μικρό αριθμό διαστάσεων. Όσο μεγάλωνε η διάσταση του χώρου τόσο πιο δύσκολο γινόταν για το κυκλικό perceptron να λύσει το πρόβλημα.

Από την άλλη πλευρά τώρα, το κλασσικό perceptron πρέπει να είναι σε θέση να λύσει τα προβλήματα του τετραγωνικού αν το τροφοδοτήσουμε με το γραμμικοποιημένο πρόβλημα. Κι εδώ όμως παρατηρείται η ίδια συμπεριφορά. Το κλασσικό perceptron έλυνε πολύ γρήγορα τα γραμμικά διαχωρίσιμα προβλήματα, αλλά καθόλου καλά τα τετραγωνικά. Επιπλέον όσο μεγάλωνε η διάσταση του προβλήματος, τόσο μεγάλωνε και η δυσκολία του. Αυτή η παρατήρηση δεν ήταν νέα, την είχε κάνει πρώτος ο Volper [46].

Αυτό που ήταν καινούριο όμως και χρειαζόταν περισσότερη ανάλυση ήταν η παρατήρηση ότι οι ειδικού τύπου συναρτήσεις βάσης ήταν βέλτιστες για τα προβλήματα του ιδίου τύπου. Δηλαδή το κυκλικό perceptron έλυνε προβλήματα που ήταν σφαιρικά διαχωρίσιμα, και το γραμμικό έλυνε προβλήματα που ήταν γραμμικά διαχωρίσιμα. Οποιαδήποτε προσπάθεια γενίκευσης της παραπάνω πρότασης σε προβλήματα του άλλου τύπου οδηγούσε σε απογοητευτικές συμπεριφορές από άποψη σύγκλισης και ταλαντώσεων.

Το αδιέξοδο αυτής της στρατηγικής είναι φανερό. Το να απαιτείται ο εκ νέου σχεδιασμός μιας ξεχωριστής συνάρτησης βάσεως για κάθε διαφορετικό πρόβλημα έρχεται σε απευθείας αντίθεση με το κυριότερο πλεονέκτημα των νευρωνικών δικτύων που είναι ότι *δεν εμπεριέχουν άλλη γνώση για το πρόβλημα, πέρα από αυτή που τους παρουσιάζεται μέσω των παραδειγμάτων*. Ήταν λοιπόν σαφές ότι προείχε η κατανόηση της στενής σύνδεσης της απόδοσης του μοντέλου με τη συνάρτηση βάσης του, και, το κυριότερο, που οφειλόταν η κακή απόδοση του στα ξένα προς τη φύση του προβλήματα.

Αν τέθηκε θέμα για το σε ποια συνάρτηση βάσης (γραμμική ή τετραγωνική) θα έπρεπε να δοθεί βάρος, ξεπεράστηκε πολύ γρήγορα, όταν διαπιστώθηκε ο όγκος των εξισώσεων που θα έπρεπε να λυθούν για τις πρώτες πειραματικές εκδόσεις του **FLF**. Οι περισσότερες από αυτές δεν είχαν καν αναλυτική λύση. Επιπλέον οι κύκλοι και οι σφαίρες είναι ειδικές περιπτώσεις των ελλειψοειδών, και υπήρχε ήδη η εμπειρία από τις κακές ιδιότητες των ελλειπτικών συναρτήσεων βάσης. Από την άλλη πλευρά το γραμμικό perceptron μπορούσε εν δυνάμει να λύσει όλα τα προβλήματα (με γραμμικοποίηση), και η διεθνής επιστημονική βιβλιογραφία είχε αρκετές αναφορές ώστε να υπάρχει μια ικανή βάση για σύγκριση.

Η επιλογή ήταν σαφής λοιπόν, μια και αν τα αποτελέσματα ήταν ενθαρρυντικά, θα ήταν δυνατόν να μεταφερθεί η επιπλέον γνώση και εμπειρία από τις γραμμικές, πίσω στις τετραγωνικές συναρτήσεις. Αυτό το τελευταίο βήμα δεν έγινε ποτέ απαραίτητο. Με τις προτεινόμενες μεταβολές το perceptron μπορεί να λύσει οποιοδήποτε επιλύσιμο (γραμμικά διαχωρίσιμο) πρόβλημα, είτε αυτό έχει προέλθει από ομογενή κατανομή, είτε από υψηλόβαθμες γραμμικοποιήσεις σε πολυδιάστατους χώρους.

3.3.2 Πρώτες προσπάθειες

Η βάση της κατασκευής κάθε αλγορίθμου είναι η ενδελεχής ανάλυση του προς επίλυση προβλήματος. Το πρόβλημα, προς ανάλυση σε πρώτη φάση και προς επίλυση σε δεύτερη, ήταν οι πολύ αργοί χρόνοι σύγκλισης που παρουσίαζε το perceptron σε προβλήματα που οι είσοδοι του ήταν κατάλληλα γραμμικοποιημένες, έτσι ώστε το πρόβλημα να είναι γραμμικά διαχωρίσιμο.

Κατά τη διάρκεια των πρώτων πειραμάτων επιβεβαιώθηκαν, διαισθητικά τουλάχιστον, οι θέσεις των Casasent [24] και Volper [46] αφού η ποσοτική επιβεβαίωση ήταν ούτως ή άλλως εκτός στόχων. Το perceptron κατά τη διάρκεια της εκπαίδευσης κατάφερνε να φτάσει αρκετά χαμηλά τη συνάρτηση κόστους, συνήθως στα ένα (1) ή δύο (2) BW. Σ' εκείνη την περιοχή όμως άρχιζαν να εμφανίζονται τα φαινόμενα ταλαντώσεως, που είναι κοινός τόπος για τα νευρωνικά δίκτυα.

Ο κανόνας ανανέωσης των βαρών του perceptron άλλαζε το διάνυσμα θέσης $\boldsymbol{W}$, αν το βήμα ήταν κατάλληλο, έτσι ώστε να παραμένει στη περιοχή των λίγων BW αλλά με διαφορετική ταξινόμηση. Με πιο απλά λόγια, ο κανόνας του perceptron προσπαθώντας να λύσει πλήρως το πρόβλημα, χαλούσε μέρος της λύσης του, άλλαζε δηλαδή, τα ήδη σωστά ταξινομημένα, διανύσματα.

Αν αυτό συμβαίνει για λίγες επαναλήψεις, τότε σε γενικές γραμμές δεν υπάρχει ιδιαίτερο πρόβλημα. Δυστυχώς όμως αυτό δεν είναι αλήθεια. Τα πειράματα μας επιβεβαίωσαν πλήρως τον Volper, δηλαδή ότι το πλήθος των επαναλήψεων που απαιτούνται είναι ανάλογο του N^8, όπου N είναι η διάσταση του προβλήματος. Ακόμα χειρότερα, η ταχύτητα σύγκλισης εξαρτιόταν από την κατανομή όπως και από την απόσταση που χωρίζει μεταξύ τους τις κλάσεις.

Ο Volper μένοντας πιστός σε μια αυστηρή θεωρητική ανάλυση, διέταξε το σύνολο των παραδειγμάτων του με ομογενή τρόπο πάνω σ' ένα πλέγμα. Αυτό του επέτρεψε την πολυτέλεια του ακριβή υπολογισμού του ρυθμού σύγκλισης (N^8), αλλά επισκίασε την σημαντικότητα των υπόλοιπων παραγόντων. Παραδείγματος χάρη, αν η απόσταση που χωρίζει τις κλάσεις είναι σχετικά 'μικρή', τότε μικρό θα είναι και το πολύτοπο στόχος στο χώρο των βαρών, κάνοντας πολύ πιο δύσκολη την επιτυχία για μια στοχαστική διαδικασία όπως το perceptron.

Από την άλλη πλευρά, αν η κατανομή των παραδειγμάτων είναι ομογενής, τότε τα πολύτοπα έχουν παρόμοια μεγέθη και σχήματα, χωρίς κάποιο να ξεχωρίζει ιδιαίτερα, γεγονός που εκμεταλλεύεται η στατιστική φύση του perceptron. Ο όρος N^8 που υπέδειξε ο Volper οφείλεται εξ' ολοκλήρου στην απομογενοποίηση του προβλήματος που προέρχεται από τη γραμμικοποίηση, δηλαδή την αύξηση των διαστάσεων του χώρου λύσης του προβλήματος με την εισαγωγή όρων ανώτερης τάξης.

Το πρόβλημα λοιπόν που σε πρώτη φάση αντιμετωπίζαμε ήταν η ταλάντωση που παρουσίαζε η σύγκλιση του perceptron στα δύσκολα προβλήματα. Ενώ αυτό το πρόβλημα και το συμπληρωματικό του, η μακροχρόνια κίνηση σε πλατώ (σχεδόν επίπεδες περιοχές με μικρή κλίση) της συνάρτησης κόστους, είναι σχεδόν άλυτο στην υπόλοιπη περιοχή των νευρωνικών δικτύων, είναι σαφώς πιο εύκολο για στην περίπτωση των μονοστρωματικών δικτύων.

Στα μονοστρωματικά δίκτυα τα όρια μεταξύ των περιοχών με διαφορετική ταξινόμηση, είναι σαφώς καθορισμένα και πολύ εύκολα στον υπολογισμό τους εφ' όσον, πρόκειται για υπερεπίπεδα. Το αποτέλεσμα αυτού είναι η διαμέριση του χώρου σε πολλαπλά πολύτοπα όπως δείξαμε στο κεφάλαιο 2.1.1. Το συμπέρασμα αυτό όμως δεν βγήκε ούτε τόσο γρήγορα, ούτε τόσο καθαρά, όπως στο εν λόγω κεφάλαιο.

Επειδή οι περισσότερες ιδέες ήταν καινούριες, η οπτικοποίηση του χώρου αλλά και των προτεινόμενων στρατηγικών επίλυσης ήταν εντελώς απαραίτητες. Η οπτικοποίηση όμως προβλημάτων πολλών διαστάσεων δεν είναι καθόλου προφανής. Γι' αυτό το λόγο στα αρχικά στάδια ανάπτυξης του αλγορίθμου οι δοκιμές γίνονταν σε προβλήματα δύο (2) ή τριών (3) διαστάσεων. Οι παρατηρήσεις σ' αυτά τα προβλήματα έδειχναν ότι όταν το perceptron έμπαινε στη τελική φάση των ταλαντώσεων, που φαινόταν να διαρκεί για πάντα, η λύση ήταν να ακολουθήσει την διεύθυνση του BR αντί να το περάσει.

Αυτή η διεύθυνση φαινόταν να οδηγεί κατ' ευθείαν στη λύση του προβλήματος, ακόμα και αν ο αριθμός των BW ήταν σχετικά ψηλός (και όχι ένα ή δύο). Κάπως αυθαίρετα, αλλά έτσι ήταν τότε η τρέχουσα αντίληψη για το πρόβλημα, οι οδηγοί γραμμές[12] ονομάστηκαν φαράγγια, γιατί φαινόντουσαν να οδηγούν

[12] Στην πραγματικότητα ήταν υπερεπίπεδα

στην κοιλάδα της συναρτήσεις κόστους. Το πρόβλημα λοιπόν σε πρώτη φάση ήταν να μπορέσουμε να εντοπίσουμε τα φαράγγια, για να μπορέσουμε να τα ακολουθήσουμε.

Χρειάστηκαν πέντε ή έξι εκδόσεις του αλγόριθμου, για να μπορέσει να γίνει η σύνδεση των BR με τα φαράγγια, αλλά στο μεταξύ ήταν ήδη διαθέσιμη μια περιορισμένη έκδοση του 'Fast Moving' η οποία επέτρεπε απότομες πτώσεις στον αριθμό των BW. Τα αποτελέσματα ήταν πολύ εντυπωσιακά αφού ο προτεινόμενος αλγόριθμος έλυνε σε τρεις ή τέσσερις επαναλήψεις προβλήματα που το perceptron χρειαζόταν αρκετές χιλιάδες. Το μηχάνημα που αναπτυσσόταν ο αλγόριθμος ήταν 486/100. Τα προβλήματα ήταν τυπικά επίπεδα, παραβολές, ελλείψεις και κύκλοι, από 500 μέχρι 2000 σημεία στις δύο διαστάσεις.

Σειρά είχαν προβλήματα μεγαλύτερων διαστάσεων. Η αναθεώρηση της οπτικής των φαραγγιών ήταν αναγκαστική. Ο αλγόριθμος είχε αναπτυχθεί με τον έμμεσο περιορισμό ότι δεν μπορεί να ακολουθεί περισσότερα από ένα φαράγγι κάθε φορά. Σε περισσότερες διαστάσεις γινόταν φανερό ότι τα φαράγγια δεν ήταν γραμμές που έπρεπε να ακολουθηθούν, αλλά τοίχοι που δεν έπρεπε να διασχιστούν (BR). Σ' αυτό το στάδιο αναπτύχθηκε όλη η τεχνολογία που επέτρεπε την ακριβή κίνηση κατά μήκος των δεσμών. Η αποθήκευση των δεσμών σε μια λίστα, γινόταν για την προφύλαξή τους από αριθμητικά σφάλματα στρογγύλευσης (moving near). Η προβολή του $\boldsymbol{\Delta W}$ πάνω στους ενεργούς δεσμούς γινόταν σύμφωνα με την διαδικασία Gramm - Schmidt.

Εκείνη την περίοδο άρχισε να γίνεται φανερό ότι το πρόβλημα προς επίλυση είχε μετατραπεί σε μια σειρά προβλημάτων επιτρεπτών κατευθύνσεων (feasible direction). Χρειάστηκαν αρκετά πειράματα μ' όλο και πιο δύσκολα σύνολα δεδομένων για να αποδειχτεί ότι οι δεσμοί που έπρεπε να ακολουθήσει η λύση κατά την έξοδο της από την μια κορυφή του πολυτόπου δεν περιλαμβάνει απαραίτητα όλους τους δεσμούς εισόδου. Είναι εύκολο να αποδειχτεί ότι ο νεοεισερχόμενος δεσμός είναι πάντα μέρος της λύσης λόγω της κυρτότητας των πολυτόπων, αλλά αυτό δεν ισχύει απαραίτητα για όλους τους προηγούμενους δεσμούς.

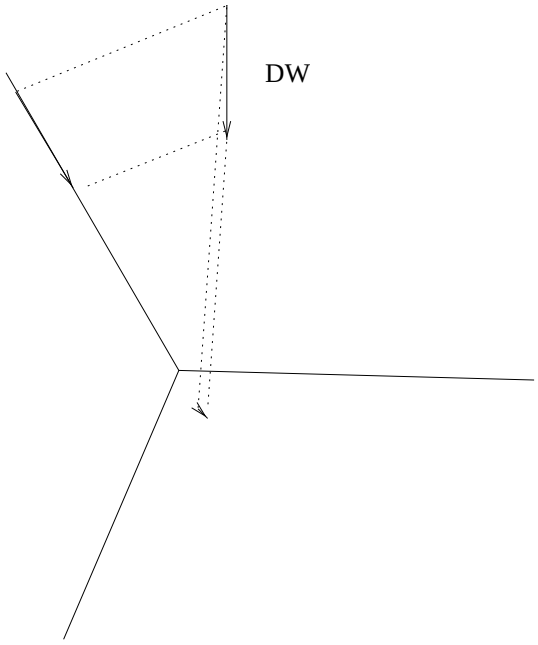
Χρειάστηκε το πρόβλημα του σχήματος 3.6 για να δειχτεί ότι απαιτείται ένας πιο συστηματικός τρόπος για την αναγνώριση και τον εντοπισμό των ενεργών δεσμών. Στο σχήμα φαίνεται μια τέτοια περίπτωση όπου ο αλγόριθμος κινείται πάνω σε μια ακμή του πολυτόπου (στερεού σε τρεις διαστάσεις), κατευθυνόμενος προς το κομβικό σημείο της κορυφής, όπου συναντά και έναν τρίτο δεσμό. Αν τώρα ο αλγόριθμος δεν είναι επαρκής ώστε να βγάλει από τη λίστα τους δύο παλιούς δεσμούς και να συνεχίσει με τον καινούριο είναι καταδικασμένος ή σε αδιέξοδο ή σε μη βέλτιστη κίνηση.

Τότε άρχισε να γίνεται σαφής ο διαχωρισμός του προβλήματος γραμμικού προγραμματισμού και του υποπροβλήματος του τετραγωνικού, δηλαδή της εύρεσης της βέλτιστης διεύθυνσης. Παρ' όλο που ο συμβολισμός δεν ήταν ακόμα τόσο σαφής όσο παρουσιάστηκε, ήταν προφανές μια φυσική μέθοδος αποσύνθεσης του μεγάλου προβλήματος γραμμικού προγραμματισμού ($P \times (N+1)$) σε πολλά μικρότερα ($K \times (N+1)$) τετραγωνικού προγραμματισμού, έχει πλεονεκτήματα σε όρους απόδοσης και ταχύτητας. Επίσης τότε συνδέθηκε και το πρόβλημα με το φυσικό του ανάλογο στις τρεις διαστάσεις, την κίνηση υλικού σημείου πάνω σε επίπεδους δεσμούς υπό την επήρεια της βαρύτητας ($\boldsymbol{\Delta W}$).

Άρχισαν λοιπόν να προτείνονται διάφοροι επιμέρους αλγόριθμοι που έλυναν το συγκεκριμένο υποπρόβλημα. Έτσι διαχωρίστηκαν πλήρως τα δύο μέρη του αλγόριθμου και σε επίπεδο προγραμματισμού και κώδικα ηλεκτρονικού υπολογιστή. Κάτω από αυτό το φως ο αλγόριθμος των Bobrowski και Niemiro [34, 47] μπορεί να θεωρηθεί ένα μέλος της οικογένειας του **FLF** με μια ειδική, όχι βέλτιστη, τεχνική για να λύνει το υποπρόβλημα του τετραγωνικού προγραμματισμού.

Για να είναι δυνατή η επαλήθευση της απόδοση των επιμέρους αλγορίθμων, αλλά και η εγκυρότητα τους, αναπτύχθηκε μια ακόμα παραλλαγή του **FLF**, που εκτελούσε εξαντλητικό ψάξιμο στο χώρο των λύσεων. Επειδή οι απαιτούμενοι χρόνοι ήταν εκθετικοί, στην πράξη ήταν αδύνατο να λυθούν προβλήματα πέρα των επτά (7) διαστάσεων. Αργότερα βέβαια ο εκθετικός κώδικας αντικαταστάθηκε με τις συνθήκες των Kuhn - Tucker αλλά η συμβολή του ήταν πολύτιμη στην κατανόηση του χώρου, και η αντικατάσταση αυτή δεν θα μπορούσε να γίνει πριν αποκτηθεί πρακτική και θεωρητική άποψη του προβλήματος. Συνολικά χρειάστηκαν πέντε με έξι εκδόσεις αλγορίθμων ακόμα πριν ωριμάσουν αρκετά ώστε καταλήξουν σ' αυτούς που θα παρουσιαστούν.

Σε χρόνο παράλληλο, εισήχθηκε και η ιδέα των επιπλέον μεταβλητών ώστε να λυθεί το πρόβλημα του διαχωρισμού του χώρου σε δύο ξεχωριστές περιοχές. Πράγματι μέχρι τότε, ο αλγόριθμος δεν ήταν σε θέση



Σχήμα 3.6: Ο αλγόριθμος πρέπει να εγκαταλείψει τους δύο δεσμούς και να συνεχίσει μόνο με τον τρίτο.

να λύσει προβλήματα αν είχε το λάθος πρόσημο στο βάρος κατωφλίου (w_0). Η μέθοδος για την άρση αυτού του περιορισμού, περιγράφεται στο κεφάλαιο 2.1.4, είναι πρωτότυπη, και οδηγεί απ' ευθείας στην απόδειξη του βασικού θεωρήματος σύγκλισης του αλγορίθμου.

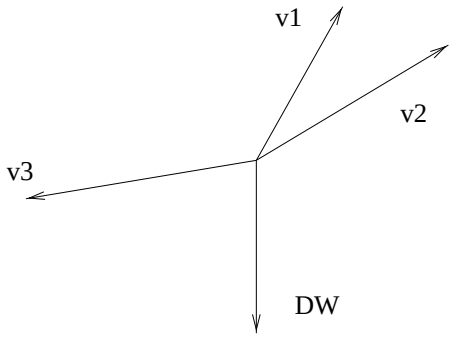
3.3.3 HooverFe

Μετά τις πρώτες απόπειρες λύσης του τετραγωνικού υποπροβλήματος, σχεδιάστηκε ο HooverFe[13]. Ο HooverFe αποτέλεσε μια πολύ επιτυχημένη προσπάθεια, και χρειάστηκαν εξαντλητικά πειράματα σε προβλήματα επτά διαστάσεων ώστε να βρεθούν αντιπαραδείγματα και να αποδειχθεί η αστοχία του σε ορισμένες περιπτώσεις.

Εξαιρετικά στρεβλές ελλείψεις των 20 (με τη γραμμικοποίηση 40) διαστάσεων, με πάνω από 20000 σημεία δεν αποτελούσαν πρόβλημα για τον HooverFe. Ακόμα με τον HooverFe λύθηκε για πρώτη φορά, με γραμμικά διαχωρίσιμο, τρόπο το πρόβλημα των Gorman και Sejnowski [29, 30]. Το περίφημο sonar.dat είναι ένα πρόβλημα που υποτίθεται είναι από τα πρώτα προβλήματα του πραγματικού κόσμου που επειδή δεν είναι γραμμικά διαχωρίσιμο σηματοδότησε την ανάγκη για νευρωνικά δίκτυα με κρυμμένους κόμβους, τα οποία μπορούν να λύσουν μη γραμμικά προβλήματα. Στην πραγματικότητα όμως οι συγγραφείς του άρθρου επηρεάστηκαν από την αργή πορεία σύγκλισης του perceptron και έβγαλαν το συμπέρασμα ότι το πρόβλημα είναι μη γραμμικά διαχωρίσιμο. Σε επανάληψη των πειραμάτων του Sejnowski με το perceptron σ' ένα μηχάνημα κλάσης Pentium/133 και χρειάστηκαν 146000 επαναλήψεις για να συγκλίνει. Στο μεταξύ στο 486/100 ο HooverFe έλυνε χαρούμενα το πρόβλημα σε 20 λεπτά και σε λιγότερο από 200 επαναλήψεις.

Οι δυνατότητες ελέγχου της προτεινόμενης λύσης έφταναν μέχρι προβλήματα επτά διαστάσεων. Ο δοκιμές γινόντουσαν σε μια συλλογή 30 ή 40 προβλημάτων που κάλυπταν δειγματοληπτικά τον χώρο των πιθανών προβλημάτων. Κατά την διάρκεια επίλυσης ενός προβλήματος με την χρήση του αλγόριθμου **FLF**, ο HooverFe έτρεχε αρκετές δεκάδες φορές, όσες φορές ήταν απαραίτητο για την σύγκλιση του αλγόριθμου.

[13]Η ρίζα του ονόματός του ήταν το αστείο της εποχής εκείνης, διότι φαινόταν να 'καθαρίζει' τα προβλήματα όπως τόσο γενναιόδωρα υποσχόταν η σχετική διαφήμιση. Τα δύο τελευταία γράμματα αποτελούν πληροφορία έκδοσης.



Σχήμα 3.7: Ο HooverFe επιλέγει και τους τρεις δεσμούς ενώ η βέλτιστη λύση είναι πάνω στην ακμή $\boldsymbol{v}_3$ (δεσμοί 1 και 2).

Γίνεται φανερό λοιπόν ότι ο HooverFe είχε ήδη τρέξει αποτελεσματικά σε χιλιάδες προβλήματα, προτού βρεθεί η ασυμφωνία σύγκλισης με τη μέθοδο του εξαντλητικού ψαξίματος σ' ένα από τα προβλήματα.

Η αναπαραγωγή του προβλήματος σε χαμηλότερες διαστάσεις στάθηκε αδύνατη. Παρ' όλα αυτά βρέθηκαν και άλλες περιπτώσεις όπου ο αλγόριθμος παρουσίαζε μη βέλτιστη συμπεριφορά. Μετά από λίγες εβδομάδες ενασχόλησης βρέθηκε και ένα αντιπαράδειγμα στις 3 διαστάσεις που επεδείκνυε ξεκάθαρα την αδυναμία του αλγόριθμου να βρει την βέλτιστη λύση. Το αντιπαράδειγμα ονομάστηκε το πρόβλημα των πιστωτικών, διότι αναπαράγεται εύκολα με 3 πιστωτικές κάρτες.

Ο ίδιος αλγόριθμος, είναι πολύ απλός στην περιγραφή και στην υλοποίησή του. Η βασική του ιδέα είναι ότι με αλλεπάλληλες πτώσεις κατά τη διεύθυνση του $\boldsymbol{\Delta W}$ είναι δυνατό να εντοπιστούν οι ενεργοί δεσμοί. Προκειμένου να έχουμε αλλοίωση του αποτελέσματος επειδή η αρχική τοποθέτηση από την οποία αρχίζει η πτώση ευνοεί κάποιον δεσμό θεωρούμε το σημείο $\boldsymbol{\rho}$ το οποίο ισαπέχει από όλους τους δεσμούς. Δηλαδή:

$$\frac{\boldsymbol{d}_i^\tau}{\|\boldsymbol{d}_i\|}\boldsymbol{\rho} = \varrho \, \forall i \in \mathcal{L} \tag{3.67}$$

Τότε αφήνουμε το διάνυσμα θέσης να κινηθεί κατά το $\boldsymbol{\Delta W}$ μέχρι να συναντήσει κάποιο δεσμό.

$$\frac{\boldsymbol{d}_i^\tau}{\|\boldsymbol{d}_i\|}(\boldsymbol{\rho} + \eta_i \boldsymbol{\Delta W}) = 0 \Rightarrow \eta_i = \frac{-\varrho\|\boldsymbol{d}_i\|}{\boldsymbol{d}_i^\tau \boldsymbol{\Delta W}} \tag{3.68}$$

Διαλέγουμε τον δεσμό με το μικρότερο θετικό η_i, και εισάγουμε το δεσμό i στη λίστα των ενεργών δεσμών. Στη συνέχεια αντικαθιστούμε το $\boldsymbol{\Delta W}$ με την προβολή του πάνω στο σύνολο των μέχρι στιγμής ενεργών δεσμών και η διαδικασία συνεχίζεται μέχρις ότου η προβολή του $\boldsymbol{\Delta W}$στη λίστα των ενεργών δεσμών αποκτήσει θετικό εσωτερικό γινόμενο με τους υπόλοιπους δεσμούς.

Αποδεικνύεται ότι ο HooverFe συγκλίνει πάντα στο σωστό αποτέλεσμα στις 2 διαστάσεις. Γι' αυτό το λόγο το αντιπαράδειγμα έπρεπε να βρεθεί στις τρεις ή και περισσότερες διαστάσεις. Στο πρόβλημα των πιστωτικών καρτών, που περιγράφεται από το σχήμα 3.7, εικονίζονται μόνο οι ακμές, και όχι τα ίδια τα υπερεπίπεδα για λόγους αναγνωσιμότητας. Τα δύο υπερεπίπεδα που αποτελούν και τη λύση του προβλήματος σχηματίζουν μια πολύ οξεία γωνία, ακριβώς όπως δύο σελίδες ενός βιβλίου που απέχουν πολύ λίγο μεταξύ τους. Η ακμή που σχηματίζεται από την τομή τους ($\boldsymbol{v}_3$) είναι σχεδόν κάθετη στο $\boldsymbol{\Delta W}$ αλλά έχει θετικό εσωτερικό γινόμενο μ' αυτό. Γίνεται αντιληπτό ότι όσο στενεύουμε την γωνία μεταξύ των δύο σελίδων, και όσο οριζοντιώνουμε τον τρίτο δεσμό, η μεν λύση δεν αλλάζει, αλλά το αρχικό διάνυσμα $\boldsymbol{\rho}$ μετατοπίζεται πάνω από τον τρίτο δεσμό. Σύμφωνα με τον HooverFe ο τρίτος δεσμός είναι ο πρώτος που εισάγεται στη λίστα, και καθώς δεν υπάρχει πρόβλεψη από τον αλγόριθμο για απομάκρυνση του δεσμού από τη λίστα είναι προφανές ότι η διαδικασία οδηγεί σε λάθος αποτέλεσμα, αφού με οπτική επισκόπηση είναι φανερό ότι στη τελική λύση συμμετέχουν μόνο οι δύο δεσμοί (αυτοί που αποτελούν τις σελίδες του βιβλίου).

3.3.4 Ο αλγόριθμος διπλής αναζήτησης

Ο βασικός λόγος αποτυχίας του HooverFe είναι η αδυναμία του να αφαιρεί δυναμικά δεσμούς από τη λίστα. Η ιδέα της τεχνικής της διπλής αναζήτησης (Dual Search) είναι αρκετά απλή και προσφέρει λύση σ' αυτή την αδυναμία. Ξεκινώντας από ένα σημείο που ανήκει στην επιτρεπτή περιοχή, ο αλγόριθμος εξετάζει ποια από τις δύο παρακάτω δυνατές διευθύνσεις του προσφέρει τη μέγιστη δυνατή μείωση στη συνάρτηση κόστους.

1. Η διεύθυνση κατά τη φυσική κατεύθυνση της παραγώγου της τετραγωνικής συνάρτησης (gradient).

2. Η διεύθυνση κατά την κατεύθυνση της προβολής του $\boldsymbol{\Delta W}$ πάνω στη τρέχουσα λίστα των δεσμών.

Η τεχνική αυτή φαίνεται να χρησιμοποιεί το γεγονός ότι η λύση που ψάχνουμε είναι προβολή του $\boldsymbol{\Delta W}$ πάνω σε μια λίστα δεσμών, επιταχύνοντας σημαντικά την αναζήτηση, ελαττώνοντας το φαινόμενο της ταλάντωσης γύρω από το σημείο ισορροπίας που προκαλείται ακολουθώντας την παράγωγο[14]. Παρ' όλα αυτά, η προτεινόμενη ιδέα προαπαιτεί τη λύση δύο προβλημάτων προτού μπορέσει να εφαρμοσθεί.

1. Ακολουθώντας είτε το gradient είτε την διεύθυνση της τοπικής προβολής του $\boldsymbol{\Delta W}$ είναι δυνατό να διασχίσουμε τα όρια της επιτρεπτής περιοχής. Κάτι τέτοιο όμως δεν είναι επιθυμητό διότι θα κάνει τον προτεινόμενο αλγόριθμο να συμπεριφέρεται μη αναμενόμενα. Το πρόβλημα όμως της εύρεσης της βέλτιστης διεύθυνσης κάτω από δεσμούς είναι δυστυχώς αυτό που προσπαθούμε να λύσουμε. Το τρυκ που χρησιμοποιήσαμε ήταν να μεταφέρουμε όλο το πρόβλημα σ' έναν χώρο που να είναι εύκολα επιλύσιμο. Ο χώρος αυτός είναι ο χώρος των ορθογώνιων δεσμών. Πράγματι, με ορθογώνιους δεσμούς, η εύρεση της λίστας των ενεργών δεσμών είναι προφανής. Στη συνέχεια λοιπόν θα επικεντρωθούμε στην λύση του προγράμματος 3.47.

2. Εφ' όσον θέλουμε να έχουμε μέγιστη μείωση της συνάρτησης κόστους, πρέπει να λύνουμε το πρόβλημα της γραμμικής ελαχιστοποίησης της συνάρτησης κόστους σε κάθε επανάληψη. Ευτυχώς, αντίθετα από άλλα προβλήματα η λύση είναι αναλυτική και μπορεί να αποφευχθεί η χρονοβόρα διαδικασία μιας αριθμητικής ελαχιστοποίησης σε κάθε επανάληψη του αλγόριθμου.

   Αν υποθέσουμε ότι θέλουμε να ελαχιστοποιήσουμε τη συνάρτηση κόστους 3.47 ως προς το βήμα η κατά τη διεύθυνση $\boldsymbol{\delta q}$ τότε έχουμε:

$$f' = \frac{1}{2}(\boldsymbol{\delta w}^\tau - \boldsymbol{q}^\tau - \eta\boldsymbol{\delta q}^\tau)\boldsymbol{V}^\tau\boldsymbol{V}(\boldsymbol{\delta w} - \boldsymbol{q} - \eta\boldsymbol{\delta q}) \tag{3.69}$$

$$f' = f + \eta\nabla f^\tau\boldsymbol{\delta q} + \frac{1}{2}\eta^2\boldsymbol{\delta q}^\tau\boldsymbol{V}^\tau\boldsymbol{V}\boldsymbol{\delta q} \tag{3.70}$$

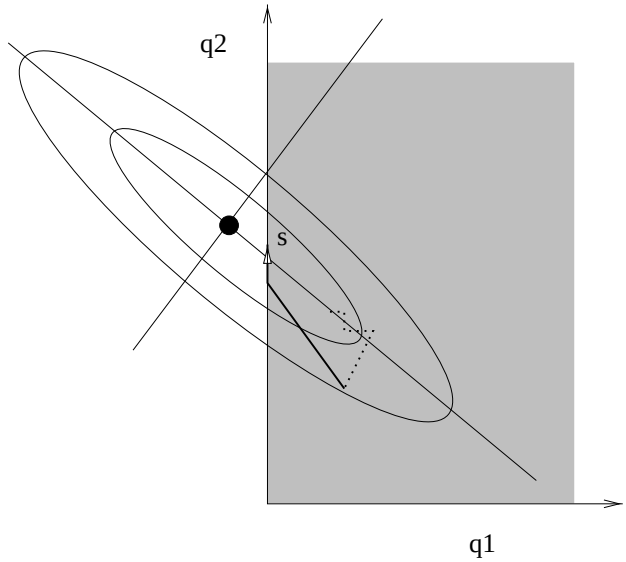
$$\frac{df'}{d\eta} = \nabla f^\tau\boldsymbol{\delta q} + \eta\boldsymbol{\delta q}^\tau\boldsymbol{V}^\tau\boldsymbol{V}\boldsymbol{\delta q} = 0 \Rightarrow \tag{3.71}$$

$$\eta = \frac{-\nabla f^\tau\boldsymbol{\delta q}}{\|\boldsymbol{\Delta Q}\|^2} \tag{3.72}$$

   όπου για ευκολία έχουμε θέσει ότι $\boldsymbol{\Delta Q} = \boldsymbol{V}\boldsymbol{\delta q}$, $\boldsymbol{\delta w} = \boldsymbol{D}^\tau\boldsymbol{\Delta W}$ και φυσικά ισχύει ότι $\nabla f = -\boldsymbol{V}^\tau\boldsymbol{V}(\boldsymbol{\delta w} - \boldsymbol{q})$.

Αφού λοιπόν έχουμε τον τρόπο να ελαχιστοποιούμε την συνάρτηση κόστους σε κάθε πιθανή διεύθυνση $\boldsymbol{\delta q}$, το μοναδικό σημείο που πρέπει να είμαστε προσεκτικοί είναι ο χειρισμός των δεσμών. Υπάρχουν δύο σημεία που πρέπει να προσέξουμε:

[14] Ενώ η τεχνική των συζυγών διευθύνσεων (conjugate gradient) φαίνεται καλή ιδέα, δεν είναι διότι η τεχνική αυτή έχει διατυπωθεί για προβλήματα χωρίς δεσμούς.



Σχήμα 3.8: Η διακεκομμένη γραμμή δείχνει την πιθανή πορεία του αλγόριθμου αν ακολουθούσε μόνο τη διεύθυνση του gradient.

1. Η διεύθυνση $\boldsymbol{\delta q}$ μας οδηγεί προς δεσμό, πριν σημειωθεί η ελαχιστοποίηση της συνάρτησης κόστους. Σ' αυτή τη περίπτωση διαλέγουμε το βήμα έτσι που να κινηθούμε ακριβώς ($q_i + \eta_i \delta q_i = 0$) μέχρι το κοντινότερο δεσμό τον οποίο θα εισάγουμε και στην εσωτερική μας λίστα για να τον προστατέψουμε από αριθμητικά σφάλματα. Δηλαδή:

$$\eta = \min\{\frac{-\nabla f^\tau \boldsymbol{\delta q}}{\|\boldsymbol{\Delta Q}\|^2},\ -\frac{q_i}{\delta q_i} : \eta > 0\} \tag{3.73}$$

2. Η διεύθυνση $\boldsymbol{\delta q}$ μας οδηγεί έξω από δεσμούς που είναι ήδη στη λίστα μας, και άρα έχουν $q_i = 0$. Τότε είμαστε υποχρεωμένοι να περιορίσουμε το αρχικό $\boldsymbol{\delta q}$ με τον εξής προφανή τρόπο:

$$\delta q_i = \begin{cases} 0, \text{ αν } q_i = 0 \text{ και } \delta q_i < 0 \\ \delta q_i, \text{ σε άλλη περίπτωση} \end{cases} \tag{3.74}$$

Τώρα ο προτεινόμενος αλγόριθμος διπλής αναζήτησης αρχίζει να διαγράφεται κάπως καθαρότερα. Ξεκινάμε από το σημείο $\boldsymbol{q} = \boldsymbol{0}$ ($\boldsymbol{Q} = \boldsymbol{0}$), όπου όλοι οι δεσμοί ανήκουν στη λίστα. Μετά υπολογίζονται διαδοχικά οι δύο πιθανές διευθύνσεις:

1.
$$\boldsymbol{\delta q} = -\nabla f = \boldsymbol{V}^\tau \boldsymbol{V}(\boldsymbol{\delta w} - \boldsymbol{q}) = \boldsymbol{V}^\tau(\boldsymbol{\Delta W} - \boldsymbol{Q}) \tag{3.75}$$

2.
$$\boldsymbol{\delta q} = \boldsymbol{D}^\tau(\boldsymbol{\Delta W}_{GS} - \boldsymbol{Q}) \tag{3.76}$$

   όπου $\boldsymbol{\Delta W}_{GS}$ είναι η προβολή του $\boldsymbol{\Delta W}$ πάνω στη λίστα των ενεργών δεσμών, ή με όρους γραμμικής άλγεβρας, στο μηδενόχωρο των ενεργών δεσμών.

Στη συνέχεια, σύμφωνα με το κανόνα της εξίσωσης 3.74, κρατάμε μόνο τις συνιστώσες εκείνες που δεν αντιβαίνουν στους υπάρχοντες δεσμούς. Χρησιμοποιώντας την σχέση 3.73 είναι δυνατόν να υπολογιστούν

τα βέλτιστα βήματα, εκείνα δηλαδή που ελαχιστοποιούν τη συνάρτηση κόστους χωρίς να βγαίνουν όμως έξω από την επιτρεπτή περιοχή, και για τις δύο πιθανές αρχικές διευθύνσεις. Ο αλγόριθμος επιλέγει τη διεύθυνση εκείνη που ρίχνει τη συνάρτηση κόστους χαμηλότερα.

Ο αλγόριθμος τερματίζει όταν το διάνυσμα $\boldsymbol{\delta q}$ του gradient είναι 0 και συνεπώς ο αλγόριθμος δεν μπορεί να κινήσει επιπλέον το διάνυσμα θέσης $\boldsymbol{q}$. Αυτό μπορεί να συμβεί για δύο λόγους. Πρώτον, ο αλγόριθμος βρήκε το ελάχιστο μέσα στην επιτρεπτή περιοχή ως προς τη διεύθυνση i και συνεπώς $\delta q_i = 0$. Δεύτερον ο αλγόριθμος βρίσκεται πάνω στο δεσμό i και το αρχικό $\delta q_i < 0$. Η ένωση των δύο περιπτώσεων αποτελούν τις συνθήκες Kuhn - Tucker και συνεπώς ο αλγόριθμος τερματίζει φυσιολογικά.

Ο χειρισμός των δεσμών τώρα, ίσως φαίνεται ιδιαίτερα απλός, αλλά είναι απολύτως ορθός. Ο τρόπος χειρισμού που περιγράφεται από την σχέση 3.74 είναι ισοδύναμος με την προβολή του $\boldsymbol{\delta q}$ πάνω στη λίστα των ενεργών δεσμών. Επειδή οι δεσμοί είναι ορθογώνιοι και συνεπώς τα μεταξύ τους εσωτερικά γινόμενα είναι μηδενικά, η διαδικασία Gramm - Schmidt περικόπτεται σ' αυτή που περιγράφουμε, με μια ακόμα απλοποίηση. Το όλο πρόβλημα, στο οποίο ψάχνουμε να δώσουμε λύση, δεν είναι το πως γίνεται η προβολή, αλλά σε ποιους δεσμούς γίνεται κάθε φορά η προβολή. Στη περίπτωση των ορθογωνίων δεσμών η απάντηση είναι προφανής, και γι' αυτό είμαστε αναγκασμένοι να εμπλακούμε σ' αυτή τη παλινδρόμηση μεταξύ του αρχικού και του χώρου των ορθογώνιων δεσμών.

Για να εξασφαλίσουμε τη σύγκλιση σε σχετικά μικρό αριθμό βημάτων είναι απαραίτητο να χρησιμοποιήσουμε όλη την υπάρχουσα πληροφορία είτε αυτή προέρχεται από το gradient, είτε από τη διεύθυνση προβολής του $\boldsymbol{\Delta W}$. Ο αλγόριθμος πράγματι αποφεύγει τα ζικ-ζακ, και κατά την εμπειρία μας εντοπίζει την ακριβή λύση σε ελάχιστες επαναλήψεις. Στο σχήμα 3.8 φαίνεται μια τέτοια περίπτωση που η συνδυασμένη γνώση βοηθά τον αλγόριθμο να συγκλίνει πολύ γρηγορότερα.

3.3.5 Ένας αλγόριθμος πεπερασμένων βημάτων

Ο αλγόριθμος της διπλής αναζήτησης δεν έχει αποτύχει μέχρι στιγμής ούτε σ' ένα από τα εξαντλητικά τεστ που τον έχουμε υποβάλλει. Παρ' όλα αυτά έχει ένα σημαντικό μειονέκτημα. Εξ' αιτίας της δυϊκής υπόστασής του στάθηκε αδύνατο να επιτευχθεί μια αναλυτική απόδειξη όσον αφορά στη σύγκλισή του σε πεπερασμένο αριθμό βημάτων. Επιπλέον ο τρόπος που αποκόπτει τα διανύσματα διεύθυνσης δυσκολεύει ακόμα περισσότερο οποιαδήποτε τέτοια προσπάθεια.

Στο παρόν θα παρουσιάσουμε μια παραλλαγή του αλγόριθμου της διπλής αναζήτησης που έχει όμως την εξαιρετική ιδιότητα να είναι πεπερασμένων βημάτων. Η κεντρική ιδέα είναι ότι ακολουθώντας μόνο τη διεύθυνση της προβολής του $\boldsymbol{\Delta W}$ πάνω στη τρέχουσα λίστα δεσμών, ρίχνουμε πάντα τη τιμή της συνάρτησης κόστους, χωρίς όμως να βγαίνουμε από τα όρια της επιτρεπτής περιοχής. Άρα είναι δυνατόν να συμβούν δύο τινά όταν ο αλγόριθμος δεν μπορεί να κινηθεί άλλο. Είτε έχει φτάσει στη λύση, είτε έχει φτάσει σε μια προβολή του $\boldsymbol{\Delta W}$, που ανήκει όμως μέσα στην επιτρεπτή περιοχή. Σε κάθε περίπτωση ελέγχουμε τον αλγόριθμο με τις συνθήκες Kuhn - Tucker. Αν πρόκειται για τη λύση έχει καλώς. Αν όχι τότε μπορούμε να ξεκολλήσουμε τον αλγόριθμο από εκεί ακολουθώντας τη διεύθυνση του gradient, κίνηση η οποία θα ρίξει επιπλέον τη συνάρτηση κόστους.

Συνεχίζοντας τη διαδικασία, προσπαθώντας να προσεγγίσουμε, τη τρέχουσα προβολή του $\boldsymbol{\Delta W}$, ακολουθώντας την εξίσωση 3.76, είναι αδύνατο να περάσουμε από το ίδιο σημείο, εφ' όσον αυτό βρίσκεται πιο ψηλά από την τρέχουσα θέση από ενεργειακή άποψη. Δεδομένου ότι όλες οι πιθανές προβολές του $\boldsymbol{\Delta W}$ είναι πεπερασμένες, και δεν ανήκουν όλες στην επιτρεπτή περιοχή, είναι σαφές ότι αυτός ο αλγόριθμος θα φτάσει στο σωστό συνδυασμό δεσμών σε πεπερασμένο αριθμό βημάτων.

Το μοναδικό λεπτό σημείο που απομένει, είναι να δώσουμε επαρκείς εγγυήσεις, για το ότι ακολουθώντας τη διεύθυνση της προβολής του $\boldsymbol{\Delta W}$ η συνάρτηση κόστους μειώνεται. Για να μπορέσουμε να δείξουμε κάτι τέτοιο αρκεί να δείξουμε ότι το εσωτερικό γινόμενο $-\nabla f^\tau \boldsymbol{D}^\tau(\boldsymbol{\Delta W}_{GS} - \boldsymbol{Q})$ είναι θετικό. Πράγματι:

$$-\nabla f^\tau \boldsymbol{D}^\tau(\boldsymbol{\Delta W}_{GS} - \boldsymbol{Q}) = \tag{3.77}$$

$$(\boldsymbol{\Delta W} - \boldsymbol{Q})^\tau \boldsymbol{V} \boldsymbol{D}^\tau(\boldsymbol{\Delta W}_{GS} - \boldsymbol{Q}) = \tag{3.78}$$

$$(\boldsymbol{\Delta W} - \boldsymbol{Q})^\tau(\boldsymbol{\Delta W}_{GS} - \boldsymbol{Q}) = \tag{3.79}$$

$$(\boldsymbol{\Delta W}_{GS} - \boldsymbol{Q})^2 > 0 \tag{3.80}$$

διότι το $\boldsymbol{Q}$ που είναι το διάνυσμα θέσης μας, επαφίεται ήδη πάνω στους ίδιους δεσμούς που ικανοποιεί και το $\boldsymbol{\Delta W}_{GS}$ αφού ισχύει $\boldsymbol{\Delta W Q} = \boldsymbol{\Delta W}_{GS}\boldsymbol{Q}$.

Ένας άλλος τρόπος, πιο διαισθητικός, να καταλήξουμε στο ίδιο συμπέρασμα προκύπτει αν αναλογιστούμε ότι πάνω σ' ένα συγκεκριμένο σύνολο δεσμών, η συνάρτηση κόστους ελαχιστοποιείται στο σημείο που αποτελεί προβολή του $\boldsymbol{\Delta W}$ στον συγκεκριμένο υπόχωρο ($\boldsymbol{\Delta W}_{GS}$). Άρα, η προς τα εκεί κίνηση, είναι αδύνατο να μην μειώνει τη συνάρτηση κόστους.

Μια παραλλαγή της παραπάνω μεθόδου είναι ή μέθοδος του Rosen, αν την εφαρμόσουμε ειδικά για το τετραγωνικό υποπρόβλημα. Η μοναδική διαφορά είναι ότι αντί να ακολουθήσουμε την διεύθυνση του gradient, προκειμένου να ξεφύγουμε από ένα σημείο που είναι προβολή του $\boldsymbol{\Delta W}$ στην τρέχουσα λίστα δεσμών, ακολουθούμε την βέλτιστη ακμή. Ο αλγόριθμος εξακολουθεί να έχει το πλεονέκτημα ότι παραμένει πεπερασμένος σε πλήθος επαναλήψεων, αλλά είναι και πολύ πιο κομψός διότι δεν απαιτεί υπολογισμούς και στους δύο χώρους. Πράγματι η λίστα των ενεργών δεσμών αρκεί για να μπορέσουμε να προκαθορίσουμε αν έχουμε φτάσει στη λύση και την κίνηση της επόμενης επανάληψης αν όχι. Αν τώρα διαλέξουμε το γινόμενο $\frac{\boldsymbol{v}_i^T}{\boldsymbol{v}_i^2}\boldsymbol{\Delta W}$ σαν το κριτήριο της καλύτερης ακμής αντί για $\frac{\boldsymbol{v}_i^T}{\|\boldsymbol{v}_i\|}\boldsymbol{\Delta W}$ τότε ο αλγόριθμος έχει κόστος επανάληψης $O(NK^2)$ όπως και ο αλγόριθμος διπλής αναζήτησης, αλλά χωρίς το επιπλέον κόστος $O(NM^3)$ που απαιτείται για τον υπολογισμό των ακμών.

Κεφάλαιο 4

Πειραματικά αποτελέσματα

4.1 Περιγραφή προβλημάτων

Τα πειραματικά αποτελέσματα των εξαντλητικών εξομοιώσεων που θα παρουσιάσουμε έχουν διττό στόχο. Συγκεκριμένα θέλουμε να δείξουμε ότι:

1. Η αποσύνθεση του γραμμικού προγράμματος σε μικρότερα τετραγωνικού προγραμματισμού είναι ωφέλιμη και γόνιμη τακτική που μπορεί να προσφέρει μεγάλη βελτίωση σε απόδοση, ιδιαίτερα σε προβλήματα μεγάλης κλίμακας.

2. Η μέθοδος που προτείνουμε για την επίλυση του τετραγωνικού προγράμματος είναι πρωτότυπη, και εφ' όσον δεν έχουμε απόδειξη πολυωνυμικής σύγκλισης, κρίνεται απαραίτητο να μελετήσουμε τη συμπεριφορά της σε εξομοιώσεις για να μπορέσουμε να εκτιμήσουμε την απόδοσή της.

Τα προβλήματα ταξινόμησης τα οποία αποτέλεσαν τους άξονες των πολλαπλών τεστ σύγκρισης χωρίζονται στις εξής κατηγορίες:

Προβλήματα ομογενούς κατανομής Τα προβλήματα εδώ σχηματίζονται από σημεία τα οποία είναι τυχαίως κατανεμημένα στο εσωτερικό ενός υπερκύβου N διαστάσεων. Ένα τυχαία εκλεγμένο υπερεπίπεδο αποτελεί το σύνορο των δύο κλάσεων. Στο παρόν μελετούνται τρία προβλήματα αυτού του τύπου με διαφορετικό αριθμό σημείων και διαστάσεων.

Ελλειπτικά διαχωρίσιμα προβλήματα Σ' αυτού του τύπου τα προβλήματα απαιτείται από το δίκτυο να διαχωρίσει πλήρως τα σημεία που βρίσκονται στο εσωτερικό ενός ελλειψοειδούς στις M διαστάσεις, από αυτά που βρίσκονται στο εξωτερικό του. Το ελλειψοειδές τυπικά έχει εξίσωση $\sum_{i=1}^{M}(x_i - c_i)^2/a_i^2 = 1$. Προφανώς το πρόβλημα μπορεί εύκολα να μετατραπεί σε γραμμικά διαχωρίσιμο, αν χρησιμοποιήσουμε και τους όρους της δεύτερης τάξης των x_i. Έτσι η διάσταση των προτύπων εισόδου είναι $N = 2M$. Με αυτή τη διατύπωση το πρόβλημα σχετίζεται άμεσα με δίκτυα τύπου Casasent [24]. Η ειδική περίπτωση της υπερσφαίρας που μελετήθηκε από τον Volper [46] έδειξε ότι ο κανόνας του perceptron απαιτεί περίπου $O(P^3N^8)$ επαναλήψεις. Και σ' αυτή τη περίπτωση μελετούνται τρία προβλήματα αυτού του τύπου.

Αναγνώριση στόχων sonar Αυτό είναι ένα πολύ γνωστό πρόβλημα διαχωρισμού ανακλώμενων σημάτων sonar σε δύο κλάσεις, σε μεταλλικούς κυλίνδρους[1] και βράχους. Χρησιμοποιούμε το αρχικό σύνολο δεδομένων που μελετήθηκε από τους Gorman και Sejnowski [29, 30], το οποίο αποτελείται από 208 διανύσματα εισόδου, το καθένα με 60 συνιστώσες. Σ' αυτό το πρόβλημα οι Gorman και

[1] νάρκες

Sejnowski ανέφεραν μόνο 85% επιτυχία για το μονοστρωματικό perceptron, το οποίο μπορεί να γίνει 100% μόνο με την εισαγωγή 12 κρυμμένων κόμβων, σ' ένα ενδιάμεσο επίπεδο, στην αρχιτεκτονική του νευρωνικού δικτύου.

Ιονοσφαιρικά δεδομένα Σ' αυτό το πρόβλημα τα δεδομένα προέρχονται από εκπομπή ραντάρ στην ιονόσφαιρα. Το πρόβλημα προτάθηκε και μελετήθηκε από τον Sigillito [48]. Αποτελείται από 351 διανύσματα εισόδου, το καθένα με 33 συνιστώσες.

Αναγνώριση Χαρακτήρων Οι χαρακτήρες έχουν ψηφιοποιηθεί σε εικόνες 25×25. Ο καθένας από τους χαρακτήρες μετασχηματίζεται χρησιμοποιώντας 3 μετασχηματισμούς κλίμακας και 5 διαφορετικές στροφές, σύμφωνα με τη μέθοδο των Perantonis και Lisboa [49]. Έτσι παράγονται 32 αναλλοίωτα σε στροφές χαρακτηριστικά.

OCR1 Αυτό είναι ένα σχετικά μικρό πρόβλημα που περιέχει 32 χαρακτήρες και 160 εικόνες χαρακτήρων συνολικά. Το πρόβλημα είναι γραμμικά διαχωρίσιμο.

OCR2 Το αρχικό σύνολο δεδομένων αποτελείτο από 240 πρότυπα εισόδου (20 παραλλαγές, 12 γραμμάτων του ελληνικού αλφάβητου) ψηφιοποιημένα σε μια εικόνα 25×25. Ο καθένας από τους χαρακτήρες μετασχηματίζεται χρησιμοποιώντας 3 μετασχηματισμούς κλίμακας και 5 διαφορετικές στροφές, παράγοντας συνολικά $240 \times 15 = 3240$ εικόνες χαρακτήρων. Από αυτές τις εικόνες, 32 χαρακτηριστικά, σχεδόν αναλλοίωτα σε μετασχηματισμούς κλίμακας και στροφής, εξάγονται, χρησιμοποιώντας συσχετισμούς τρίτης τάξης, σύμφωνα με τη μέθοδο των Perantonis και Lisboa [49]. Ένα δίκτυο 32 εισόδων και 12 εξόδων χρησιμοποιείται για το διαχωρισμό των χαρακτήρων σε κλάσεις. Σε αντίστοιχα πειράματα που έγιναν [49], αναφέρεται ότι τα μονοστρωματικά δίκτυα παρουσιάζουν δυσκολίες στην εκπαίδευσή τους σε τέτοιου είδους προβλήματα, και συνεπώς ένα επίπεδων κρυμμένων κόμβων ήταν απαραίτητο για την αύξηση της ταξινομητικής ικανότητας. Στο παρόν εξετάζουμε αυτό το πρόβλημα ξανά κάτω από το φως της ανάλυσής μας για το γραμμικό προγραμματισμό.

4.2 Ταχύτητα εκμάθησης

Για να μπορέσουμε να εκτιμήσουμε την ταχύτητα σύγκλισης του προτεινόμενου αλγόριθμου τον συγκρίνουμε με άλλους, ήδη γνωστούς αλγόριθμους, στην εκπαίδευση μονοστρωματικών δικτύων. Η σύγκριση περιλαμβάνει τέσσερις αλγόριθμους οι οποίοι δοκιμάστηκαν σε όλα τα προβλήματα:

1. Ο προτεινόμενος αλγόριθμος (**FLF**, βλ. 2.3.1).
2. Ο αλγόριθμος που προτάθηκε το 1984 από τους Bobrowski και Niemiro (BN [34, 47].
3. Ο αλγόριθμος του αυθεντικού perceptron με τον κανόνα του Rosenblatt.
4. Ο αλγόριθμος που αποτελεί προσαρμογή του conjugate gradient στα νευρωνικά δίκτυα [50].

Ο conjugate gradient είναι ένας βελτιωμένος αλγόριθμος, σε σχέση με το perceptron του Rosenblatt αλλά απαιτεί συνεχή συνάρτηση εξόδου (σιγμοειδή), και όχι τη συνάρτηση βήματος. Για να έχουμε μια πιο πλήρη εικόνα, δοκιμάζουμε και τον conjugate gradient στα προβλήματα που παρουσιάσαμε.

Τα αποτελέσματα που αφορούν την ταχύτητα παρουσιάζονται στον πίνακα 4.1 σε δευτερόλεπτα, ενώ ο αριθμός των επαναλήψεων που απαιτήθηκαν από τον κάθε αλγόριθμο βρίσκεται στον πίνακα 4.2. Το ποσοστό των σωστά ταξινομημένων διανυσμάτων εισόδου που επιτεύχθηκαν κατά τον τερματισμό του αλγορίθμου παρουσιάζεται στον πίνακα 4.3.

Όλα τα αποτελέσματα είναι μέσοι όροι δέκα προσπαθειών, των οποίων η εκκίνηση έγινε από διαφορετικά, τυχαία αρχικά βάρη τα οποία ανήκαν στο διάστημα από -0.5 έως 0.5. Ο προτεινόμενος αλγόριθμος και αυτός

	FLF	BN	Perceptron	CG
Ομογενής				
P=100 N=2	0.006	0.015	0.435	0.326
P=1000 N=4	0.243	0.527	1.379	4.466
P=20000 N=10	33.0	108.6	3600	184.1
Έλλειψη				
P=100 N=4	0.023	0.031	0.059	0.227
P=10000 N=20	74.3	259.4	2661.4	328.2
P=20000 N=40	647.6	4554	3600	582.9
Sonar	117.8	405.9	1174.9	3600
OCR1	4.75	5.65	42.38	3600
Ionosphere	22.1	31.8	3600	3600
OCR2	308.3	1337.7	3600	3600

Πίνακας 4.1: Ο χρόνος σε δευτερόλεπτα που χρειάστηκε ο κάθε αλγόριθμος για να τερματίσει.

	FLF	BN	Perceptron	CG
Ομογενής				
P=100 N=2	4.6	6.5	581	12.6
P=1000 N=4	16.4	32.0	157	18.5
P=20000 N=10	76.9	205	14439	208
Έλλειψη				
P=100 N=4	13.4	19.5	63.1	9.0
P=10000 N=20	214.0	567	14749	104.5
P=20000 N=40	488.7	2162	6082	77.8
Sonar	223.8	190.5	146047	123549
OCR1	16.0	17.9	94.4	2855
Ionosphere	155.5	142.8	430735	101658
OCR2	294.4	533.9	4554	589.3

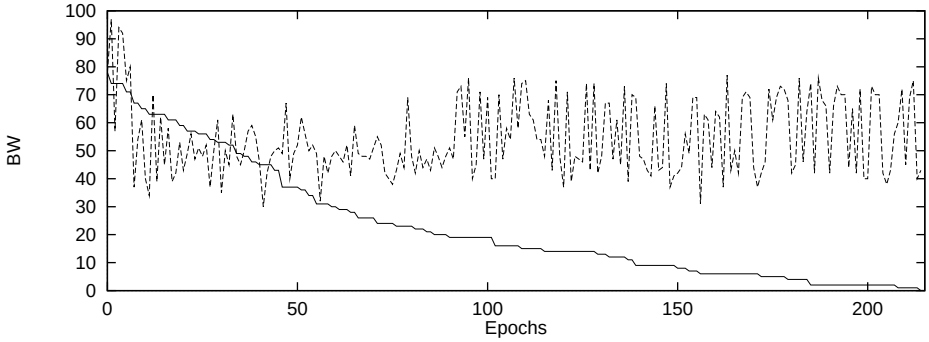
Πίνακας 4.2: Ο αριθμός των επαναλήψεων που χρειάστηκε ο κάθε αλγόριθμος για να τερματίσει.

	FLF	BN	Perceptron	CG
Ομεγενής				
P=100 N=2	100	100	100	100
P=1000 N=4	100	100	100	100
P=20000 N=10	100	100	99.91	100
Έλλειψη				
P=100 N=4	100	100	100	100
P=1000 N=20	100	100	99.90	100
P=20000 N=40	100	100	99.72	100
Sonar	100	100	100	94.13
OCR1	100	100	100	97.51
Ionosphere	96.35	94.87	92.22	97.07
OCR2	99.72	99.64	97.47	94.36

Πίνακας 4.3: Η απόδοση του κάθε αλγορίθμου σε ποσοστά σωστά ταξινομημένων προτύπων εισόδου.

	Simplex	**FLF** (Dsearch)	Bobrowski	**FLF** (Rosen)
Ομογενής				
P=100 N=2	0.03 (107)	0 (3)	0 (3)	0 (3)
P=1000 N=4	3.83 (1098)	0.12 (12)	0.37 (35)	0.12 (12)
P=20000 N=10	3117 (23510)	15.95 (64)	46.22 (135)	16.89 (64)
Έλλειψη				
P=100 N=4	0.04 (117)	0.01 (6)	0.01 (6)	0.01 (6)
P=10000 N=20	1278.8 (13199)	45.97 (219)	163.33 (582)	45.38 (219)
P=20000 N=40	9840 (30278)	336.73 (461)	2108.9 (1866)	321.09 (461)
Sonar	2.37 (664)	57.36 (256)	141.01 (177)	5.01 (256)
OCR1	347.93 (6952)	2.35 (16)	2.46 (16)	2.34 (16)

Πίνακας 4.4: Σύγκριση με την Simplex. Τα αποτελέσματα είναι σε δευτερόλεπτα, σε παρένθεση είναι ο αριθμός των επαναλήψεων που απαιτούνται.



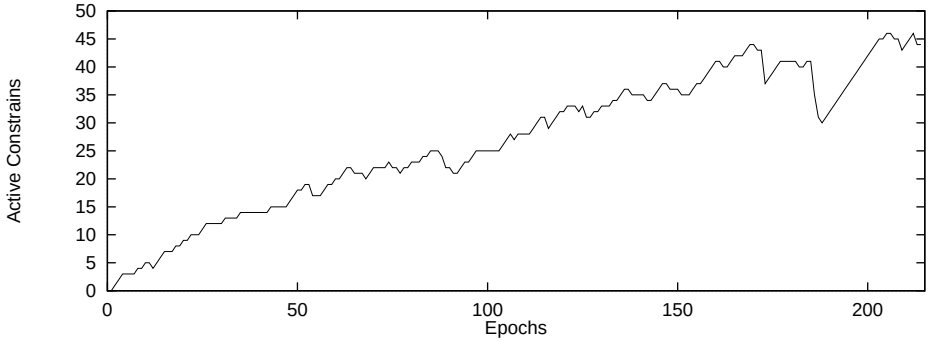
Σχήμα 4.1: Ο αριθμός των BW σε σχέση με το πλήθος των επαναλήψεων, τόσο για τον **FLF** όσο και για το perceptron.

των Bobrowski και Niemiro (BN) αφέθηκαν να συνεχίσουν μέχρι να λύσουν το πρόβλημα ή να ικανοποιήσουν το κριτήριο τερματισμού τους. Το perceptron και ο conjugate gradient τερματίζονται αυτόματα σε μια ώρα αν δεν έχουν λύσει το πρόβλημα. Αυτό γίνεται αναγκαστικά διότι και οι δύο αυτοί αλγόριθμοι δεν έχουν κριτήριο τερματισμού. Και ενώ για τα γραμμικά διαχωρίσιμα προβλήματα, το perceptron έχει απόδειξη σύγκλισης, και συνεπώς τερματισμού σε πεπερασμένο αριθμό βημάτων, κάτι τέτοιο δεν ισχύει για τα μη γραμμικώς διαχωρίσιμα προβλήματα. Όλες οι εξομοιώσεις ήρθαν σε πέρας χρησιμοποιώντας έναν υπολογιστή Pentium/133, και ένα πακέτο που εξομοιώνει μια πληθώρα νευρωνικών τεχνικών εκμάθησης (billnet), και έχει αναπτυχθεί τοπικά [2].

Συγκρινόμενη με το perceptron η προτεινόμενη μέθοδος παρουσιάζει σαφή υπεροχή, σε όρους ταχύτητας εκμάθησης. Η παρατηρούμενη υπεροχή είναι ήδη υπαρκτή στα μικρής κλίμακας προβλήματα, αλλά γίνεται πολύ έντονη στα μεσαίας, και μεγάλης κλίμακας προβλήματα. Το σχήμα 4.1 δείχνει μια τυπική πορεία εκμάθησης τόσο του perceptron όσο και του **FLF**, για το sonar. Στο σχήμα 4.1 εικονίζεται το πλήθος των λάθος ταξινομημένων προτύπων, για κάθε επανάληψη του αλγορίθμου. Όπως βλέπουμε, ο **FLF** διατηρεί φθίνουσα πορεία, ενώ το perceptron παρουσιάζει τη γνωστή ταλάντωση.

Στο σχήμα 4.2 παρουσιάζεται το M, το πλήθος των ενεργών δεσμών, σε σχέση με τον αριθμό των επαναλήψεων, την χρονική δηλαδή εξέλιξη του αλγορίθμου. Όπως μπορεί να δει κανείς, ο αλγόριθμος ξεκινά με λίγους ενεργούς δεσμούς στην αρχή του προβλήματος, και όσο το πρόβλημα δυσκολεύει, επειδή πλησιάζει προς τη λύση του, ο αριθμός των δεσμών αυξάνει. Στην πραγματικότητα, όσο η διαδικασία της εκμάθησης προχωρά, ο αλγόριθμος ανακαλύπτει τις πιο αποδοτικές διευθύνσεις κίνησης, χτίζοντας μια

[2] http://www.iit.demokritos.gr/ vasvir/billnet



Σχήμα 4.2: Ο αριθμός των ενεργών δεσμών σε κάθε επανάληψη.

εσωτερική βάση δεδομένων από τους δεσμούς που τις συγκροτούν, ενώ το perceptron είναι μακριά από τη λύση και η καμπύλη εκμάθησής του είναι σχεδόν επίπεδη.

Είναι σημαντικό να παρατηρήσουμε ότι ο προτεινόμενος αλγόριθμος έχει ταξινομήσει σωστά όλα τα διανύσματα εισόδου στα γραμμικά διαχωρίσιμα προβλήματα, ενώ το perceptron απέτυχε να τα καταφέρει στο δοθέν χρονικό διάστημα της μίας ώρας για τα μεγάλης κλίμακας προβλήματα.

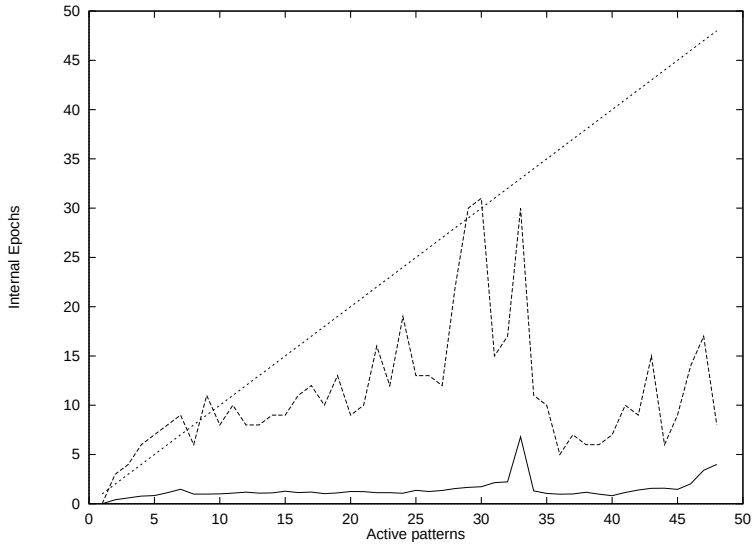
Για το πρόβλημα των δεδομένων από sonar ο αλγόριθμός μας επέτυχε τον πλήρη διαχωρισμό των δύο κλάσεων και συνεπώς το πρόβλημα είναι γραμμικά διαχωρίσιμο. Το ίδιο βέβαια κατάφερε και το perceptron έπειτα από 146000 επαναλήψεις. Δεν είναι λοιπόν παράξενο που οι Gorman και Sejnowski [29, 30] (1988) δεν κατάφεραν να διακρίνουν ότι το πρόβλημα ήταν γραμμικά διαχωρίσιμο. Κατά τη διάρκεια της συγγραφής του παρόντος, υπέπεσε στην αντίληψη μας, και μια άλλη δημοσίευση που φτάνει στο ίδιο συμπέρασμα, ακολουθώντας μιαν άλλη τεχνική (Moreno και Gordon (1998), [51]). Το θέμα όμως είναι ότι με τον προτεινόμενο αλγόριθμο, είναι δυνατόν να αποφανθούμε με βεβαιότητα το κατά πόσο ένα πρόβλημα είναι γραμμικά διαχωρίσιμο.

Είναι επίσης σαφές από τους πίνακες 4.1 και 4.2 ότι ο προτεινόμενος αλγόριθμος παρουσιάζει μια σαφή βελτίωση και σε σχέση με τον αλγόριθμο των Bobrowski και Niemiro, κυρίως για δύο λόγους:

1. Ο αλγόριθμος εκ φύσεως επιλέγει την βέλτιστη διεύθυνση, λύνοντας το δύσκολο τετραγωνικό πρόβλημα του κεφαλαίου 3, σε σχέση με τον BN που διαλέγει την βέλτιστη ακμή. Αυτό τον βοηθά να πετύχει μεγαλύτερη ευθυγράμμιση με τη φυσική παράγωγο του perceptron, έχοντας έτσι καλύτερη απόδοση, και φτάνοντας στη λύση του προβλήματος σε μικρότερο αριθμό βημάτων.

2. Το γεγονός ότι η μέθοδος BN ακολουθεί μόνο τις ακμές των πολυτόπων, σημαίνει ότι από τη στιγμή που ο αλγόριθμος συναντήσει $N+1$ BR πρότυπα, το πλήθος των ενεργών δεσμών αποκόπτεται στους $N+1$. Στη δική μας περίπτωση όμως, όπως φαίνεται και από το σχήμα 4.2, ο αριθμός των ενεργών δεσμών M μεταβάλλεται συνέχεια, μένοντας για το μεγαλύτερο διάστημα της εκπαίδευσης, μικρότερο από $N+1$. Επειδή και οι δύο αλγόριθμοι χρειάζονται να κτίσουν τις ακμές του πολυτόπου, διαδικασία που είναι κόστους $O((N+1)M^3)$, γίνεται αντιληπτό ότι ο προτεινόμενος αλγόριθμος παρουσιάζει ένα σαφές πλεονέκτημα σε σχέση με τον BN που απαιτεί $O((N+1)^4)$ πράξεις για την λύση του τετραγωνικού προβλήματος για την εύρεση της βέλτιστης διεύθυνσης, σχεδόν σε κάθε επανάληψη του αλγόριθμου.

   Το πλεονέκτημα που αποκτούμε απαιτώντας την βέλτιστη λύση για την διεύθυνση αναζήτησης, και όχι απλώς μια αποδεκτή λύση (ακμή) όπως οι Bobrowski και Niemiro, θα ήταν εύκολο να το χάσουμε αν ο αλγόριθμος διπλής αναζήτησης χρειαζόταν πολύ χρόνο για να φέρει σε πέρας το πρόγραμμα που υπολογίζει την βέλτιστη διεύθυνση. Για κάθε εσωτερική επανάληψη του αλγόριθμου διπλής αναζήτησης απαιτούνται πράξεις $O((N+1)M'^2)$, όπου M' είναι το πλήθος των δεσμών σε κάθε



Σχήμα 4.3: Το μέγιστο πλήθος, και ο μέσος όρος των εσωτερικών επαναλήψεων που απαιτούνται από τον αλγόριθμο διπλής αναζήτησης για την εύρεση της βέλτιστης διεύθυνσης, σε συνάρτηση με τον αριθμό των ενεργών δεσμών.

εσωτερική επανάληψη του αλγορίθμου. Αν το πλήθος των επαναλήψεων υπερβαίνει κατά πολύ το M' τότε το πλεονέκτημα που έχουμε αποκτήσει κατά τον υπολογισμό των ακμών χάνεται.

Ευτυχώς τα πειραματικά αποτελέσματα μας δικαιώνουν αφού όπως βλέπουμε στο σχήμα 4.3, το πλήθος των εσωτερικών επαναλήψεων δεν ξεπερνάει το πλήθος των ενεργών δεσμών M. Στην πραγματικότητα στο σχήμα εικονίζονται 2 καμπύλες. Η καμπύλη που είναι ψηλότερα αποτελείται από το μέγιστο αριθμό εσωτερικών επαναλήψεων που απαιτήθηκαν καθ' όλη την διάρκεια της εκπαίδευσης για όλα τα προβλήματα που απαντήθηκαν με συγκεκριμένο πλήθος δεσμών. Αντίθετα η καμπύλη που είναι χαμηλότερα απαρτίζεται από το μέσο όρο εσωτερικών επαναλήψεων στα αντίστοιχα, σε πλήθος ενεργών δεσμών, προβλήματα.

Το ενδιαφέρον συμπέρασμα που προκύπτει από τη μελέτη του σχήματος 4.3 είναι ότι ένα σχετικά μικρό ποσοστό χρόνου σπαταλάται στον αλγόριθμο της διπλής αναζήτησης, αφού τα σχήματα και οι χρόνοι δείχνουν ότι ο υπολογισμός των ακμών καταναλώνει τον περισσότερο χρόνο.

Η παραλλαγή του αλγορίθμου του Rosen όπως περιγράφεται στο κεφάλαιο 3.3.5 υλοποιήθηκε και συγκρίθηκε με την Simplex και τις άλλες μεθόδους σε ένα υποσύνολο των προβλημάτων στον πίνακα 4.4. Η μέθοδος του Rosen είναι αποδεδειγμένα πεπερασμένη και έχει λιγότερες πράξεις από την μέθοδο της διπλής αναζήτησης. Παρ' όλα αυτά επειδή οι δοκιμές αποτελούνται από προβλήματα μεγάλης κλίμακας όπου το $P >> N$ βλέπουμε ότι ουσιαστικά οι δύο μεθόδοι δεν έχουν διαφορά. Από την άλλη πλευρά η Simplex είναι κατώτερη της μεθόδου των Bobrowski και Niemiro η οποία με τη σειρά της είναι κατώτερη από τις προτεινόμενες. Μοναδική εξαίρεση αποτελεί το sonar που, με τον εξαιρετικά χαμηλό λόγο P/N, αναδεικύεται καλύτερη του Bobrowski. Σ' αυτό το πρόβλημα φαίνεται και η ανωτερότητα της μεθόδου του Rosen έναντι στον αλγόριθμο της διπλής αναζήτησης, όπου έχει αποτελέσματα συγκρίσιμα με την Simplex.

Τέλος, η προτεινόμενη μέθοδος (**FLF**) παρουσιάζει και μια σημαντική βελτίωση στην ταχύτητα εκμάθησης και σε σχέση με τον conjugate gradient. Αντίθετα από το perceptron όμως έδωσε πολύ πιο καλή συμπεριφορά από άποψη κλιμάκωσης (scaling).Στη πραγματικότητα είναι τόσο καλά κλιμακούμενος που ξεπέρασε σε απόδοση ακόμα και τον προτεινόμενο σ' ένα πρόβλημα μεγάλης κλίμακας (P=20000 N=40), αν και θα πρέπει να υπογραμμιστεί ότι δεν κατάφερε να διαχωρίσει σχετικά απλά πραγματικά προβλήματα, που ακόμα και το perceptron τα κατάφερε (sonar, OCR1) στο δοθέν διάστημα της μίας ώρας.

	FLF	BN	Perceptron
Ομογενής			
P=100 N=2	97.13	96.93	96.27
P=1000 N=4	99.90	99.94	99.89
P=20000 N=10	99.96	99.95	99.85
Έλλειψη			
P=100 N=4	96.93	96.53	97.60
P=10000 N=20	99.82	99.85	99.80
P=20000 N=40	99.77	87.47	99.57
Sonar	75.00	74.03	76.61
OCR1	98.40	97.99	99.44
Ionosphere	86.77	86.87	85.75
OCR2	99.20	99.11	98.94

Πίνακας 4.5: Η γενικευτική ικανότητα του κάθε αλγορίθμου σε ποσοστά σωστά ταξινομημένων προτύπων εισόδου.

4.3 Γενικευτική Ικανότητα

Ένα από τα βασικά χαρακτηριστικά των νευρωνικών δικτύων είναι η γενικευτική ικανότητα, η ικανότητα δηλαδή να εξάγουν σχετικά ασφαλή συμπεράσματα σε δεδομένα που δεν έχουν ποτέ τροφοδοτηθεί στο δίκτυο κατά τη διάρκεια της εκπαίδευσης. Έτσι και εμείς, εκτελέσαμε μια σειρά πειραμάτων, για να μπορέσουμε να αξιολογήσουμε την γενικευτική ικανότητα του προτεινόμενου αλγόριθμου σε σχέση με αυτούς των Bobrowski και Niemiro, και το perceptron του Rosenblatt. Ο conjugate gradient μένει έξω απ' αυτό το γύρο δοκιμών διότι χρησιμοποιεί σιγμοειδείς συναρτήσεις αντί για συναρτήσεις βήματος, γεγονός που μετατρέπει την σχετική σύγκριση σε άκυρη.

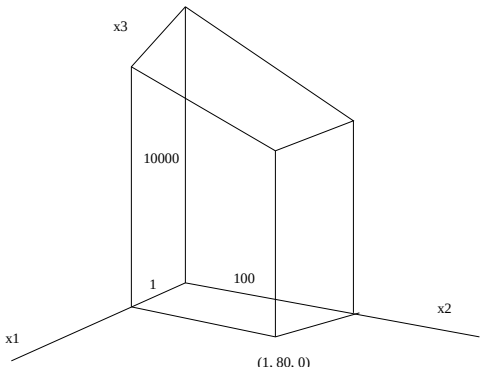
Τα πειράματα σ' αυτό το στάδιο έγιναν ως εξής: Το κάθε πρόβλημα προς επίλυση διαμερίσθηκε κατά 80%, επί του συνολικού αριθμού προτύπων, σε σύνολο εκπαίδευσης, και κατά 20% σε σύνολο αξιολόγησης. Συνολικά χρησιμοποιήθηκαν 10 διαφορετικές διαμερίσεις και 10 διαφορετικές επανεκκινήσεις με τυχαία βάρη για την κάθε διαμέριση ώστε να επιτύχουμε στατιστική αξιοπιστία. Τα βάρη ως συνήθως προκύπτουν από μια γεννήτρια τυχαίων αριθμών στο διάστημα από -0.5 έως 0.5 ακολουθώντας μια ομογενή κατανομή. Τα κριτήρια τερματισμού ήταν ακριβώς ίδια με πριν.

Τα αποτελέσματα δίνονται από τον πίνακα 4.5 σαν μέσος όρος του ποσοστού των σωστά ταξινομημένων προτύπων επί του συνολικού αριθμού προτύπων στο σύνολο δοκιμής (test set). Σε τέσσερα από τα εξεταζόμενα προβλήματα ο προτεινόμενος αλγόριθμος έχει την καλύτερη γενικευτική ικανότητα, ενώ στα υπόλοιπα έχει την δεύτερη καλύτερη. Παρατηρούμε λοιπόν ότι η γενικευτική ικανότητα δεν θυσιάζεται στο βωμό της απόδοσης που προσφέρει ο **FLF**.

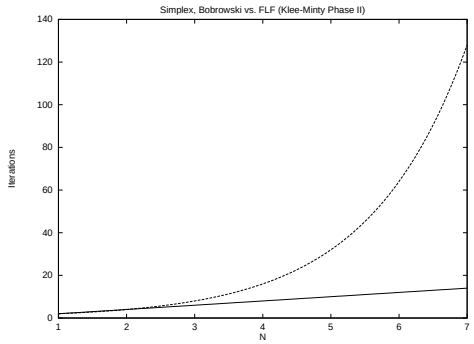
4.4 Πολυπλοκότητα και πολυωνυμική συμπεριφορά

Υπάρχουν 2 ανοιχτά θέματα όσον αφορά το συνολικό κόστος του προτεινόμενου αλγόριθμου:

Το συνολικό πλήθος των επαναλήψεων του αλγόριθμου Για να βγει ο αλγόριθμος από ένα πολύτοπο πρέπει να μειώσει τον αριθμό των BW τουλάχιστον κατά ένα. Αυτό σημαίνει ότι ο αλγόριθμος θα επισκεφτεί το πολύ P διαφορετικά πολύτοπα, γιατί είναι φυσικά αδύνατον να έχουμε περισσότερα BW, από όσα διανύσματα να ταξινομήσουμε. Έτσι για να έχουμε πολυωνυμική πολυπλοκότητα πρέπει να δειχτεί ότι ο αλγόριθμος εξερευνά το πολύτοπο σε πολυωνυμικό χρόνο. Η εξερεύνηση αυτή αντιπροσωπεύεται από το γενικό γραμμικό πρόγραμμα που βελτιστοποιεί γραμμική συνάρτηση 2.4 με ανισοτικούς δεσμούς (φάση 2 της Simplex).



Σχήμα 4.4: Ο 'κομμένος κύβος' των Klee και Minty στις 3 διαστάσεις.



Σχήμα 4.5: Ο αριθμός των επαναλήψεων για την Simplex και **FLF** για το πρόβλημα του 'κομμένου κύβου' των Klee και Minty σε σχέση με τον αριθμό των διαστάσεων.

Οι Klee και Minty [33] επινόησαν ένα πρόβλημα στο οποίο η Simplex αναγκάζεται να επισκεφτεί και τις 2^N κορυφές ενός κομμένου κύβου. Πιο συγκεκριμένα το πρόβλημα των Klee και Minty περιγράφεται ως εξής:

$$\begin{cases} max\ f = \sum_{i=1}^{N} 10^{N-i} x_i \\ x_i \geq 0\ \forall i = 1 \ldots N \\ \sum_{j=1}^{i-1} 2 \cdot 10^{i-j} x_j + x_i \leq 100^{i-1}\ \forall i = 1 \ldots N \end{cases} \tag{4.1}$$

πράγμα που σημαίνει ότι έχει $P = 2N$ δεσμούς. Η λύση του περιλαμβάνει πάντα τους πρώτους $N-1$ δεσμούς και τον τελευταίο $(2N)$. Έχουμε δηλαδή στη λύση:

$$x_i = 0\ \forall i = 1 \ldots N-1,\ x_N = 100^{N-1} \tag{4.2}$$

με τιμή της συνάρτησης $f = x_N = 100^{N-1}$. Στο σχήμα 4.4 εικονίζεται το πρόβλημα των Klee και Minty για $N = 3$.

Επαναλάβαμε το πείραμα και όπως αναμενόταν η μέθοδος του Bobrowski είχε και αυτή εκθετική συμπεριφορά[3]. Αντίθετα όμως ο **FLF** είχε γραμμική συμπεριφορά λύνοντας το πρόβλημα σε $2N$ βήματα για οποιοδήποτε αριθμό διαστάσεων δοκιμάσαμε να λύσουμε, όπως φαίνεται και από το σχήμα 4.5.

Λύνοντας βέβαια σε γραμμικό χρόνο, το αντιπαράδειγμα που έδειξε ότι η Simplex είναι εκθετικής πολυπλοκότητας, δεν αποτελεί απόδειξη ότι η προτεινόμενη μέθοδος είναι γραμμικής ή έστω πολυωνυμικής πολυπλοκότητας. Παρ' όλο που διαισθητικά είμαστε πεπεισμένοι ότι ο **FLF** είναι πολυωνυμικής πολυπλοκότητας, μια πιο αυστηρή απόδειξη φαίνεται σαφώς πιο δύσκολη. Άλλωστε δεν πρέπει να ξεχνάμε ότι και η Simplex υποτίθεται ότι είχε πολυωνυμική συμπεριφορά μέχρι να έρθουν οι Klee και Minty με το αντιπαράδειγμα τους.

Το τετραγωνικό υποπρόβλημα Ο αλγόριθμος διπλής αναζήτησης που λύνει το τετραγωνικό υποπρόβλημα αλλά και ο **FLF** μπορούν να θεωρηθούν ότι ανήκουν σε μια υποκατηγορία των αλγορίθμων επιτρεπτών κατευθύνσεων. Δεν ανήκουν απλώς στην υποκατηγορία αλγορίθμων ενεργών δεσμών, αλλά στο υποσύνολο των αλγορίθμων των βέλτιστων ενεργών δεσμών (best active constraints). Με βάση αυτό το διαχωρισμό έχει άμεση σχέση με τις μεθόδους του Zoutendijk (best feasible directions). Ο Zoutendijk βέβαια επιχειρεί να λύσει το τετραγωνικό υποπρόβλημα που του παρουσιάζεται με Simplex, πράγμα που εκτός από άκομψο είναι και μη πρακτικό.

Ο αλγόριθμος διπλής αναζήτησης, παρ' όλο που δεν έχει αυστηρή απόδειξη, συμπεριφέρεται με σχεδόν βέλτιστο τρόπο. Αν τύχει και επισκεφτεί το σωστό συνδυασμό δεσμών, βρίσκει τη λύση σε μια επανάληψη. Στο μεταξύ επειδή η κατεύθυνση του gradient που ακολουθεί είναι τετραγωνικής συνάρτησης, δεν αναμένεται πάνω από μια αλλαγή κατεύθυνσης ανά δεσμό. Ο αλγόριθμος κινείται στον υπόχωρο των ενεργών δεσμών με βέλτιστο τρόπο, μετά την ορθογωνιοποίηση που έχουμε κάνει και έτσι δεν αναμένεται συμπεριφορά τεθλασμένης γραμμής ή κυκλική τροχιά. Τα παραπάνω δεν αποτελούν βέβαια αυστηρή απόδειξη, αλλά είναι βάσιμες σκέψεις που ενισχύονται από τα πειραματικά αποτελέσματα που δείχνουν υπογραμμικό πλήθος επαναλήψεων.

4.5 Συμπεράσματα

Στο παρόν σύγγραμμα έχουμε εισάγει έναν εξαιρετικά αποδοτικό αλγόριθμο για την εκπαίδευση μονοστρωματικών δικτύων τύπου perceptron με τη συνάρτηση βήματος σαν συνάρτηση εξόδου. Ο αλγόριθμος ελαχιστοποιεί τη συνάρτηση κόστους του perceptron, ακολουθώντας τη διεύθυνση της μέγιστης πτώσης (steepest descent), έχοντας σαν απαίτηση ταυτόχρονα να μην αυξήσει των αριθμό των λάθος ταξινομημένων

[3] Δεν πρέπει να ξεχνάμε ότι η μέθοδος του Bobrowski και η Simplex παράγουν την ίδια ακολουθία σημείων για προβλήματα της δεύτερης φάσης

προτύπων. Μ' αυτό τον τρόπο το πρόβλημα της εκπαίδευσης ενός perceptron διασπάται σε μια ακολουθία από μικρά προβλήματα τετραγωνικού προγραμματισμού, των οποίων η λύση καθορίζει την αποδεκτή, σύμφωνα με τους τρέχοντες δεσμούς, διεύθυνση μέγιστης πτώσης.

Η συνεισφορά του παρόντος στον ερευνητικό τομέα μπορεί να συνοψιστεί στα εξής σημεία:

- Η απόδειξη ότι η προτεινόμενη στρατηγική τερματίζει σ' έναν πεπερασμένο αριθμό βημάτων (βλ. 2.3.1) άσχετα από τη φύση του προβλήματος (γραμμικά διαχωρίσιμο ή όχι) αποτελώντας έτσι ένα φυσικό κριτήριο για τη γραμμική διαχωρισιμότητα.

- Η απόδειξη ότι ο αλγόριθμος πάντα συγκλίνει στη λύση του προβλήματος, στην περίπτωση που το πρόβλημα είναι γραμμικά διαχωρίσιμο.

- Η πειραματική απόδειξη, σε γραμμικά και όχι διαχωρίσιμα προβλήματα, του ότι χρησιμοποιώντας έναν σχετικά αποτελεσματικό αλγόριθμο τετραγωνικού προγραμματισμού (αλγόριθμος διπλής αναζήτησης) για την εύρεση της βέλτιστης διεύθυνσης είναι δυνατόν να εκπαιδεύσει κανείς μονοστρωματικά δίκτυα τύπου perceptron πολύ αποδοτικότερα απ' ότι με άλλες τεχνικές.

- Ο προτεινόμενος αλγόριθμος για την επίλυση του τετραγωνικού προγράμματος είναι κι' αυτός πρωτότυπος, μια και κάνει χρήση της πληροφορίας ότι η λύση είναι προβολή του $\boldsymbol{\Delta W}$ πάνω σ' ένα συνδυασμό δεσμών.

- Η δυνατότητα του αλγόριθμου να μην αυξάνει τον αριθμό των λάθος ταξινομημένων προτύπων είναι πραγματικά πολύτιμη. Χρησιμοποιώντας αυτή την ιδιότητα σαν δομικό λίθο είναι δυνατόν να στηρίξει κανείς το οικοδόμημα ενός διστρωματικού δικτύου. Επίσης, με ελάχιστες μετατροπές είναι δυνατόν ο αλγόριθμος να μετασχηματιστεί έτσι ώστε να εντοπίζει τα πραγματικά τοπικά ελάχιστα της γραμμικής διαχωρισιμότητας στην περίπτωση των μη γραμμικά διαχωρίσιμων προβλημάτων.

  Αυτή η ιδιότητα είναι τόσο σημαντική, πέρα από όρους ταχύτητας και απόδοσης, που μια άλλη πιο βιολογική διατύπωση θα ήταν ίσως καταλληλότερη. Συγκεκριμένα θα μπορούσε να πει κανείς ότι *Το σύστημα αυτό, έτσι όπως εκπαιδεύεται από τον προτεινόμενο αλγόριθμο, έχει την δυνατότητα της αποθήκευσης νέας γνώσης χωρίς όμως να 'ξεχνάει' την παλιά.*

Επειδή πεποίθησή μας, έστω και αναπόδεικτη, είναι ότι η όλη διαδικασία σύγκλισης του **FLF** είναι όχι μόνο πεπερασμένη αλλά και πολυωνυμική θα θέλαμε να μελετηθούν σε βάθος τα παρακάτω θέματα:

- Έχουμε ήδη αποδείξει ότι η προτεινόμενη στρατηγική (**FLF**) οδηγεί σε πεπερασμένο αριθμό βημάτων στον τερματισμό του αλγόριθμου. Για να αποδειχτεί πολυωνυμική σύγκλιση αρκεί να αποδειχτεί ότι ο αλγόριθμος εξαντλεί το κάθε πολύτοπο σε πολυωνυμικό αριθμό βημάτων.

- Ο **FLF** όμως προαπαιτεί τη χρήση ενός αλγόριθμου τετραγωνικού προγραμματισμού για την εύρεση της βέλτιστης αποδεκτής διεύθυνσης. Αν αυτός ο αλγόριθμος (αλγόριθμος διπλής αναζήτησης) δεν είναι πολυωνυμικής πολυπλοκότητας τότε ούτε ο **FLF** είναι. Παρ' όλο που έχουμε ήδη υποδείξει έναν αλγόριθμο, σε αντίθεση με τον αλγόριθμο διπλής αναζήτησης, πεπερασμένων βημάτων (βλ. 3.3.5) γεγονός που κάνει όλη την διαδικασία πεπερασμένη με μαθηματική αυστηρότητα, δεν έχουμε ακόμα πρόταση για αλγόριθμο πολυωνυμικής πολυπλοκότητας.

Κεφάλαιο 5

Προεπεξεργασία

5.1 Εισαγωγή

Τα τελευταία χρόνια, εκτεταμένες θεωρητικές έρευνες αλλά και ειδικευμένες μελέτες σε συγκεκριμένες εφαρμογές από πολλούς ερευνητές έχουν βοηθήσει στο να εδραιωθεί ο ρόλος των νευρωνικών δικτύων σαν αποδοτικές και αξιόπιστες μηχανές πληροφορικής επεξεργασίας. Αυτές οι μελέτες δείχνουν, ότι τα νευρωνικά δίκτυα αποτελούν συμφέρουσες επιλογές που μπορούν να λύσουν ένα ευρύ φάσμα προβλημάτων ταξινόμησης, προσέγγισης συναρτήσεων, αλλά και ελέγχου (control), και βελτιστοποίησης. Πιο συγκεκριμένα, τα πολυεπίπεδα νευρωνικά δίκτυα εμπρόσθιας διάδοσης έχουν συγκεντρώσει την περισσότερη προσοχή εξ' αιτίας τόσο των γενικευμένων προσεγγιστικών, όσο και των πρόσφατων σχετικά επιτυχημένων αλγόριθμων εκπαίδευσης.

Η εμπειρία που έχει αποκομισθεί από τα νευρωνικά δίκτυα σε μια πλειάδα εφαρμογών δείχνει ότι σε πολλές περιπτώσεις, η χρήση των νευρωνικών δικτύων είναι χρήσιμη μόνο σαν επιμέρους αρθρωμάτων (modules) ενός γενικότερου συστήματος. Σε πάρα πολλές περιπτώσεις επίσης η προεπεξεργασία ή μετεπεξεργασία του προβλήματος βελτιώνει σημαντικά την παρατηρούμενη απόδοση είτε σε όρους ταχύτητας είτε σε όρους γενικευτικής ικανότητας. Στο στάδιο της προεπεξεργασίας ιδιαίτερα είναι σημαντικό να επιλεγεί ένα μικρό σύνολο χαρακτηριστικών για κάθε πρότυπο, έτσι ώστε να αποφευχθούν χαρακτηριστικά (features) τα οποία είναι κυρίως θόρυβος, και απλώς αυξάνουν τη διάσταση του προβλήματος.

Αυτό είναι μια ιδιαίτερα επώδυνη αλήθεια στην περίπτωση προβλημάτων φυσικού κόσμου (real world problems) στα οποία τα πρότυπα απαρτίζονται από πολυδιάστατα διανύσματα. Προς αυτή τη κατεύθυνση έχουν προταθεί πολλές τεχνικές είτε εξαγωγής είτε επιλογής χαρακτηριστικών, οι οποίες προέρχονται τόσο από τον χώρο της στατιστικής, όσο και από την έρευνα στον τομέα των νευρωνικών δικτύων. Αυτές οι τεχνικές επικεντρώνονται στο να διαλέξουν τις πιο σημαντικές για το πρόβλημα συνιστώσες των διανυσμάτων, ή να συνδυάσουν τις συνιστώσες με τέτοιο τρόπο ώστε να παράγουν νέα χαρακτηριστικά, λιγότερα στο πλήθος, αλλά πιθανόν πιο σημαντικά.

Μια μέθοδος επιλογής χαρακτηριστικών κατάλληλη για χρήση νευρωνικών δικτύων εμπρόσθιας διάδοσης αναπτύχθηκε από τον Ruck [52], ο οποίος όρισε μια μετρική σημαντικότητας που εξαρτάται από την ευαισθησία των εξόδων του δικτύου σε σχέση με τις εισόδους του. Το νευρωνικό δίκτυο αρχικά εκπαιδεύεται έχοντας σαν εισόδους όλα τα διαθέσιμα χαρακτηριστικά των οποίων και η σημαντικότητα υπολογίζεται. Μόνο τα χαρακτηριστικά εκείνα που ξεπερνούν ένα ορισμένο κατώφλι σημαντικότητας συμμετέχουν στην τελική διαδικασία εκπαίδευσης, μειώνοντας έτσι, πιθανόν δραστικά, τον αριθμό των διαστάσεων του χώρου των χαρακτηριστικών. Πρόσφατα, η μέθοδος του Ruck βελτιώθηκε με χρήση στατιστικών τεχνικών για τον υπολογισμό του βέλτιστου κατωφλίου σημαντικότητας (saliency), έτσι ώστε να είναι δυνατή η ακριβής εκτίμηση του τελικού αριθμού των σημαντικών χαρακτηριστικών. Το πρόβλημα με τη μέθοδο του Ruck είναι ότι προσπαθεί να πετύχει τη μείωση των διαστάσεων του αρχικού χώρου, διαλέγοντας χαρακτηριστικά

απ' αυτόν, αλλά δεν προσπαθεί να επιτύχει επιπλέον μείωση, συνδυάζοντας χαρακτηριστικά του αρχικού χώρου για την εξαγωγή νέων.

Αυτό το μειονέκτημα δεν υπάρχει σε μια άλλη εξαιρετικά δημοφιλή και ήδη πλατειά εφαρμοσμένη μέθοδο, πιο γνωστή ως ανάλυση στους πρωτεύοντες άξονες (Principal Component Analysis, PCA, [53, 54]. Αυτή η πολύ γνωστή τεχνική στο χώρο της πολυμεταβλητής στατιστικής ανάλυσης [55] παράγει ένα πιθανά πολύ μικρότερο σύνολο γραμμικών συνδυασμών των αρχικών χαρακτηριστικών, βασισμένη στη μεγιστοποίηση της διασποράς των προτύπων. Επιπλέον υπάρχουν πολλές τόσο απλές όσο και αποδοτικές παραλλαγές της μεθόδου PCA προσαρμοσμένες στα νευρωνικά δίκτυα [10, 11, 56, 57, 58], γεγονός που συντείνει στην ελκυστικότητα της μεθόδου. Παρ' όλα αυτά, η μέθοδος αυτή δεν λαμβάνει υπόψη την πληροφορία της επιθυμητής κλάσης, που είναι διαθέσιμη σε κάθε πρόβλημα εκμάθησης με επίβλεψη (supervised learning). Επιπλέον η μέθοδος είναι επιρρεπής σε σφάλματα αν τυχόν τα δεδομένα είναι κατανεμημένα σε πολλαπλά, ισοτροπικά διανεμημένα συσσωματώματα (clusters) [59].

Στο παρόν, προτείνουμε μια μέθοδο για εξαγωγή χαρακτηριστικών βασισμένη στην εύρεση των διευθύνσεων εκείνων, στο χώρο των χαρακτηριστικών, κατά τις οποίες η απόκριση του νευρωνικού δικτύου γίνεται μέγιστη. Έτσι, ουσιαστικά υλοποιούμε μια γενίκευση της μεθόδου του Ruck, για τον υπολογισμό σημαντικών νέων χαρακτηριστικών που αποτελούνται από γραμμικό συνδυασμό των αρχικών, και όχι απλώς υποσύνολο αυτών. Η προτεινόμενη μέθοδος θυμίζει αρκετά τη μέθοδο PCA, αλλά λαμβάνει υπόψη της και την πληροφορία για την κλάση που ανήκουν τα πρότυπα (Supervised PCA, SPCA). Συνήθως οδηγεί σ' έναν αριθμό σημαντικών χαρακτηριστικών, ο οποίος είναι κατά κανόνα μικρότερος από τον αριθμό των χαρακτηριστικών που προκύπτουν από τη μέθοδο του Ruck και έτσι ελαχιστοποιεί ακόμα περισσότερο τα προβλήματα που προκύπτουν από σύνολα εκπαίδευσης μεγάλης διάστασης (curse of dimensionality), ενώ ταυτόχρονα επιτυγχάνει καλύτερη απόδοση σε όρους γενικευτικής ικανότητας σε πολλά προβλήματα, γεγονός που την κάνει ιδιαίτερα ελκυστική και σε προβλήματα οπτικοποίησης. Η επιτυχία και η πιθανή χρησιμότητα της μεθόδου επιδεικνύεται από μια ομάδα τεχνητών αλλά και πραγματικών προβλημάτων στα οποία έγιναν δοκιμές και με άλλες μεθόδους για σύγκριση.

Το παρόν κεφάλαιο είναι οργανωμένο ως εξής: Στο κεφάλαιο 5.2 κάνουμε μια σύντομη επισκόπηση της μεθόδου του Ruck, αλλά και άλλων μεθόδων, καθώς και παραλλαγών τους, που επιτρέπουν τον υπολογισμό του βέλτιστου αριθμού σημαντικών χαρακτηριστικών. Στο κεφάλαιο 5.3 διατυπώνουμε μια πρωτότυπη πρόταση που αντιμετωπίζει το πρόβλημα της εξαγωγής των χαρακτηριστικών. Σ' αυτό το κεφάλαιο υπογραμμίζονται οι ομοιότητες και οι διαφορές με τη μέθοδο του Ruck αλλά και τη μέθοδο PCA. Το πόσο καλά η προτεινόμενη μέθοδος αντιμετωπίζει το πρόβλημα της εξαγωγής χαρακτηριστικών είναι το θέμα του κεφαλαίου 5.4 όπου γίνεται σύγκριση σε πραγματικά και τεχνητά προβλήματα της προτεινόμενης μεθόδου με άλλες από τη διεθνή βιβλιογραφία. Τέλος στο κεφάλαιο 5.5 παρουσιάζεται μια άλλη πιθανή εφαρμογή της προτεινόμενης μεθόδου σε προβλήματα οπτικοποίησης.

5.2 Προηγούμενες εργασίες

5.2.1 Συμβάσεις και συμβολισμοί

Ας θεωρήσουμε ένα πολυεπίπεδο νευρωνικό δίκτυο εμπρόσθιας διάδοσης, μ' ένα στρώμα εισόδου, L στρώματα από κρυμμένα επίπεδα, και φυσικά ένα στρώμα με τους κόμβους εξόδου. Οι κόμβοι (nodes) επικοινωνούν μέσω των συνάψεων (weights) με όλους τους κόμβους του προηγούμενου επιπέδου. Οι είσοδοι στο πρώτο επίπεδο θα συμβολίζονται με $x_i, i = 1, \ldots, N$ όπου N είναι ο συνολικός αριθμός χαρακτηριστικών με τα οποία το δίκτυο καλείται να εκπαιδευτεί. Οι κόμβοι εξόδου θα συμβολίζονται με $O_i^{(l)}$, όπου ο άνω δείκτης l υποδηλώνει το επίπεδο στην έξοδο του οποίου αναφερόμαστε ($l = 1, 2, \ldots, L$ για όλα τα κρυμμένα επίπεδα, $l = L + 1$ για το επίπεδο εξόδου), ενώ ο δείκτης i δηλώνει τον ακριβή κόμβο στο συγκεκριμένο επίπεδο.

Τα βάρη τώρα θα συμβολίζονται με $w_{i_{l-1} i_l}^{(l)}$ όπου i_l δηλώνει ταυτόχρονα και το επίπεδο και τον κόμβο τον οποίον η σύναψη συνδέει στο εμπρόσθιο άκρο της, ενώ το i_{l-1} δηλώνει τον κόμβο από τον οποίο η σύναψη ξεκινάει. Οι σταθεροί όροι κατωφλίου κάθε επιπέδου (bias) θα αναφέρονται σαν κανονικά συναπτικά βάρη,

με τη διαφορά ότι οι κόμβοι από τους οποίους ξεκινούν έχουν σταθερή απόκριση ίση με ένα. Ως συνήθως η λογιστική συνάρτηση χρησιμοποιείται σαν η απαιτούμενη μη γραμμική συνάρτηση απόκρισης, δηλαδή $f(s) = 1/(1 + \exp(-s))$, τόσο για το στρώμα εξόδου, όσο και για τα ενδιάμεσα στρώματα.

5.2.2 Η μέθοδος του Ruck

Ο Ruck και οι συνεργάτες του έχουν προτείνει μια μέθοδο, για αναδιάταξη των εισόδων του προβλήματος με το οποίο εκπαιδεύουμε ένα νευρωνικό δίκτυο. Η αναδιάταξη αυτή είναι βασισμένη στη φθίνουσα σειρά του μέτρου της σημαντικότητας όπως αυτό ορίζεται από τη μέθοδο του Ruck [52]. Η μέθοδος συνίσταται στην πρώιμη εκπαίδευση ενός νευρωνικού δικτύου με το πλήρες σύνολο χαρακτηριστικών, διαθέσιμο στο δίκτυο. Αφού το δίκτυο μάθει το επιτηρούμενο (supervised) πρόβλημα, είναι δυνατόν να υπολογιστεί μια μετρική σημαντικότητας (saliency metric) S_j, που συσχετίζεται με κάθε ανεξάρτητο χαρακτηριστικό. Η προεκπαίδευση μπορεί να επαναληφθεί αρκετές φορές, με διαφορετικά αρχικά βάρη ή με διαφορετικές διαμερίσεις του συνόλου εκπαίδευσης, έτσι ώστε το υπολογιζόμενο μέτρο της σημαντικότητας να αποκτήσει στατιστική αξιοπιστία.

Το μέτρο της σημαντικότητας S_j στη μέθοδο του Ruck είναι σχεδιασμένο να εκφράζει την ευαισθησία των εξόδων του προεκπαιδευμένου δικτύου, αν θεωρήσουμε μικρές μεταβολές στο συγκεκριμένο χαρακτηριστικό j, αφήνοντας όλα τα άλλα αδιατάρακτα. Ο μαθηματικός ορισμός της μετρικής είναι λοιπόν:

$$S_j = \sum_{\{\boldsymbol{x}\}} \sum_i \left| \frac{\partial O_i^{(L+1)}}{\partial x_j} \right| \tag{5.1}$$

όπου το πρώτο άθροισμα σημαίνει ότι συμπεριλαμβάνουμε την πληροφορία απ' όλα τα πρότυπα και απ' όλες της προεκπαιδεύσεις που έχουν λάβει χώρα. Η μερική παράγωγος υπολογίζεται ως εξής:

$$\frac{\partial O_{i_{L+1}}^{(L+1)}}{\partial x_{i_0}} = \sum_{i_1, i_2, \ldots, i_L} \prod_{l=1}^{L+1} O_{i_l}^{(l)} \left(1 - O_{i_l}^{(l)}\right) w_{i_{l-1} i_l}^{(l)} \tag{5.2}$$

Σαν αποτέλεσμα, το μέτρο της σημαντικότητας του κάθε χαρακτηριστικού είναι άθροισμα μιας μετρικής της παραγώγου της εξόδου, που αντιστοιχεί σε όλα τα πιθανά πρότυπα εισόδου, ως προς με το συγκεκριμένο χαρακτηριστικό. Ο Ruck διάλεξε τη μετρική τύπου 1 (απόλυτη τιμή). Από τη στιγμή που οι τιμές της σημαντικότητας για κάθε χαρακτηριστικό έχουν υπολογισθεί, το νευρωνικό δίκτυο επανεκπαιδεύεται, μόνο που αυτή τη φορά χρησιμοποιούνται τα χαρακτηριστικά των οποίων η σημαντικότητα ξεπερνά ένα συγκεκριμένο κατώφλι.

Μια ενδιαφέρουσα μέθοδος για τον υπολογισμό του πιο κατάλληλου κατωφλίου είναι η ένεση θορύβου (noise injection) η οποία προτάθηκε από τους Belue και Bauer [60]. Σύμφωνα μ' αυτή τη τεχνική ένα επιπλέον χαρακτηριστικό (συνιστώσα), προστίθεται σε κάθε διάνυσμα εισόδου, κατά τη φάση της προεκπαίδευσης, το οποίο σχηματίζεται από τυχαίες τιμές, στο διάστημα $(0, 1)$, ομογενούς κατανομής. Στη συνέχεια, το νευρωνικό δίκτυο εκπαιδεύεται πολλαπλές φορές, με διαφορετικά αρχικά βάρη ή και διαμερίσεις του συνόλου εκπαίδευσης. Αν υποθέσουμε ότι η μέση σημαντικότητα ακολουθεί την κανονική κατανομή, ένα χαρακτηριστικό μπορεί να θεωρηθεί αρκετά σημαντικό αν η δική του μέση τιμή σημαντικότητας υπερβαίνει τη μέση τιμή του θορύβου κατά μια απόσταση ασφαλείας. Κατόπιν το νευρωνικό δίκτυο επανεκπαιδεύεται μόνο με τα χαρακτηριστικά εκείνα που κρίθηκαν επαρκώς σημαντικά.

5.3 Μια καινούρια προσέγγιση

Ο Ruck χρησιμοποιεί τη μετρική της απόλυτης τιμής πάνω στην παράγωγο των εξόδων του νευρωνικού δικτύου ως προς τις εισόδους του. Η τάξη όμως της μετρικής δεν φαίνεται να επηρεάζει ούτε την ταξινόμηση ούτε την σχετική σημαντικότητα των χαρακτηριστικών, Πράγματι, μια παρόμοια, αν και πολύ απλούστερη

μετρική, που προτάθηκε από τον Tarr [61] είναι περισσότερο κοντά στην ευκλείδεια μετρική (2-norm). Από εδώ και στο εξής θα αναφερόμαστε μόνο στην ευκλείδεια μετρική, μια και είναι πιο βολική για τους σκοπούς μας.

Θεωρούμε τον διανυσματικό χώρο $\mathcal{V}$ ο οποίος δημιουργείται απ' όλα τα πιθανά χαρακτηριστικά $\boldsymbol{x}$. Το μέγεθος S_j αντιπροσωπεύει τη σημαντικότητα κατά μήκος της διεύθυνσης j. Ας θεωρήσουμε τώρα μια αυθαίρετη διεύθυνση στον χώρο $\mathcal{V}$, η οποία ορίζεται από ένα μοναδιαίο διάνυσμα $\hat{\boldsymbol{u}}$. Δεδομένου ενός διανύσματος $\boldsymbol{x}$, μπορούμε να ορίσουμε το $x_{\hat{u}}$ σαν την προβολή του στη διεύθυνση $\hat{u}$, δηλαδή $x_{\hat{u}} = \boldsymbol{x} \cdot \hat{\boldsymbol{u}}$. Τότε το μέτρο της σημαντικότητας κατά τη διεύθυνση $\hat{\boldsymbol{u}}$ ορίζεται σαν:

$$S_{\hat{u}} = \sum_{\{\boldsymbol{x}\}} \sum_i \left(\frac{\partial O_i^{(M+1)}}{\partial x_{\hat{u}}} \right)^2 \tag{5.3}$$

Επιθυμούμε την εύρεση των διευθύνσεων εκείνων $\hat{\boldsymbol{u}}$, για τις οποίες το αντίστοιχο μέτρο σημαντικότητας $S_{\hat{u}}$ είναι μέγιστο, κάτω από τον περιορισμό $\hat{\boldsymbol{u}} \cdot \hat{\boldsymbol{u}} = 1$. Θα δείξουμε ότι το πρόβλημα αυτό μετασχηματίζεται στο κλασικό πρόβλημα ιδιοτιμών ενός πραγματικού και συμμετρικού πίνακα, ακριβώς όπως συμβαίνει και στο φορμαλισμό της μεθόδου PCA. Πράγματι, αν εφαρμόσουμε τη βασική ιδιότητα της παραγώγου κατά κατεύθυνση, έχουμε:

$$\frac{\partial O_i^{(M+1)}}{\partial x_{\hat{u}}} = \sum_k \hat{u}_k \frac{\partial O_i^{(M+1)}}{\partial x_k} \tag{5.4}$$

Το μέτρο της σημαντικότητας τώρα γράφεται:

$$S_{\hat{u}} = \sum_{j,k} R_{jk} \hat{u}_j \hat{u}_k \tag{5.5}$$

όπου

$$R_{jk} = \sum_{\{\boldsymbol{x}\}} \sum_i \frac{\partial O_i^{(L+1)}}{\partial x_j} \frac{\partial O_i^{(L+1)}}{\partial x_k} \tag{5.6}$$

είναι πραγματικός και συμμετρικός πίνακας.

Θέλουμε να μεγιστοποιήσουμε λοιπόν την έκφραση 5.5 του μέτρου της σημαντικότητας ως προς $\hat{u_k}$ κάτω από τον περιορισμό $\sum_k \hat{u_k}\hat{u_k} = 1$. Έτσι εισάγουμε τον συντελεστή Lagrange μ, ο οποίος θα μας βοηθήσει να βελτιστοποιήσουμε το $S_{\hat{u}}$ λαμβάνοντας υπόψη το δεσμό. Έτσι έχουμε:

$$S'_{\hat{u}} = \sum_{j,k} R_{jk} \hat{u}_j \hat{u}_k + \mu(1 - \sum_k \hat{u_k}\hat{u_k}) \tag{5.7}$$

Η ακρότατη τιμή για το μέτρο της σημαντικότητας $S_{\hat{u}}$, κάτω από δεσμούς, βρίσκεται στα σημεία όπου $\partial S'_{\hat{u}}/\partial \hat{u}_j = 0$, έτσι ώστε:

$$\sum_k R_{jk} \hat{u_k} = \mu \hat{u}_j \tag{5.8}$$

Απ' αυτή τη διατύπωση λοιπόν είναι προφανές ότι ακρότατες τιμές, για το πρόβλημα με δεσμό, έχουμε όταν το διάνυσμα $\hat{\boldsymbol{u}}$ είναι ένα ιδιοδιάνυσμα του πίνακα $\boldsymbol{R}$.

Χρησιμοποιώντας τη σχέση 5.7, αντικαθιστώντας στη σχέση 5.5 και λαμβάνοντας υπόψη τον δεσμό, συνάγεται ότι πρέπει $S_{\hat{u}} = \mu$, συνεπώς η μέγιστη τιμή της σημαντικότητας ισούται με τη μέγιστη ιδιοτιμή του $\boldsymbol{R}$ και λαμβάνεται όταν το $\hat{\boldsymbol{u}}$ είναι το ιδιοδιάνυσμα του $\boldsymbol{R}$ που αντιστοιχεί στη μέγιστη ιδιοτιμή.

Σαν άμεσο αποτέλεσμα των παραπάνω, προτείνουμε την ακόλουθη μέθοδο για εξαγωγή χαρακτηριστικών: Το νευρωνικό δίκτυο προεκπαιδεύεται χρησιμοποιώντας όλα τα διαθέσιμα χαρακτηριστικά, κατά προτίμηση έναν ικανό αριθμό φορών με διαφορετικά αρχικά βάρη ώστε να επιτευχθεί στατιστική αξιοπιστία. Αφού η φάση της προεκπαίδευσης τελειώσει, υπολογίζονται τα στοιχεία του πίνακα $\boldsymbol{R}$ χρησιμοποιώντας την

εξίσωση 5.6. Αν θεωρήσουμε ένα κατώφλι σημαντικότητας S_T, ας υποθέσουμε ότι υπάρχουν K ιδιοτιμές οι οποίες ξεπερνούν το κατώφλι αυτό S_T. Τα K ιδιοδιανύσματα $\hat{\boldsymbol{u}}^r$, $r = 1, \ldots K$ του $\boldsymbol{R}$ που αντιστοιχούν σ' αυτές τις ιδιοτιμές, υπολογίζονται και τα K πιο σημαντικά χαρακτηριστικά που εξάγονται από τη μέθοδο μας δίνονται από:

$$\boldsymbol{x} \cdot \hat{\boldsymbol{u}}^r,\ r = 1, \ldots K \tag{5.9}$$

για καθένα από τα K ιδιοδιανύσματα.

Τελικά το νευρωνικό δίκτυο επανεκπαιδεύεται χρησιμοποιώντας μόνο τα K σημαντικότερα χαρακτηριστικά. Σύμφωνα με τους Belue και Bauer, είναι δυνατόν να υπολογιστεί το S_T, συμπεριλαμβάνοντας ένα επιπλέον χαρακτηριστικό, που είναι απλώς θόρυβος, στη φάση της προεκπαίδευσης. Το μέτρο της σημαντικότητας γι' αυτό το επιπλέον 'ιδιαίτερο' χαρακτηριστικό μπορεί να υπολογισθεί χρησιμοποιώντας την σχέση του Ruck 5.1 (με ευκλείδεια μετρική εννοείται), οπότε και έχουμε μια μέθοδο αυτόματου υπολογισμού του S_T. Σαν επιπλέον διάστημα ασφαλείας από το μέσο όρο της σημαντικότητας του θορύβου, θα μπορούσαμε να χρησιμοποιήσουμε τη διασπορά της σημαντικότητας του θορύβου, αφού υποθέτουμε επιπλέον ότι η σχετική κατανομή είναι κανονική.

Η σύζευξη του αλγόριθμου SPCA, όπως τον έχουμε παρουσιάσει μέχρι τώρα, με τα νευρωνικά δίκτυα είναι ιδιαίτερα ισχυρή. Είναι εύλογο λοιπόν το ερώτημα αν ο αλγόριθμος μπορεί να χρησιμοποιηθεί με άλλου τύπου ταξινομητές. Στην πραγματικότητα υπάρχουν δύο διαφορετικά σκέλη, στα οποία θα μπορούσαμε να εφαρμόσουμε έναν άλλο ταξινομητή:

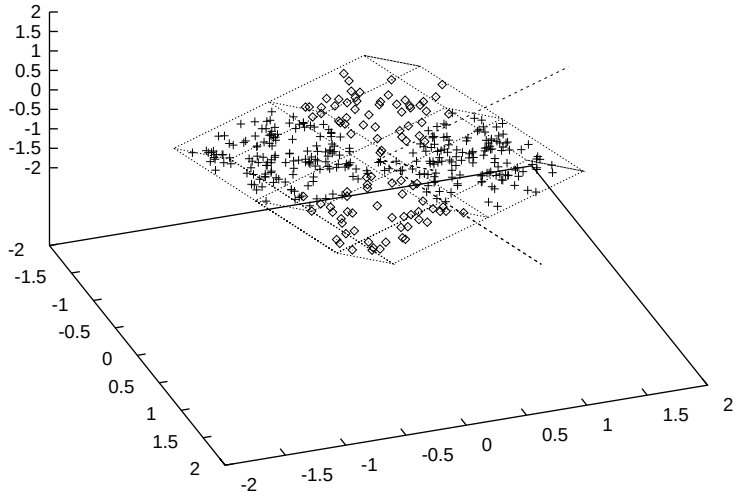
1. Για τον υπολογισμό της ευαισθησίας του ταξινομητή, προς μια συγκεκριμένη διεύθυνση, δηλαδή το βαθμό σημαντικότητας αυτής της διεύθυνσης. Εφ' όσον τα νευρωνικά δίκτυα ουσιαστικά επιστρέφουν την συνάρτηση κλάσης, ή την πιθανότητα να ανήκει το διάνυσμα εισόδου σε μια κλάση, με τρόπο συνεχή, είναι εύκολο να υπολογιστεί η παράγωγος με τρόπο αναλυτικό. Μια τέτοια διαδικασία δεν είναι δυνατή μ' έναν μη συνεχή ταξινομητή όπως είναι ο knn, όχι χωρίς τουλάχιστον σημαντική αναμόχλευση του αλγορίθμου. Πράγματι μια τέτοια εργασία έχει ήδη δρομολογηθεί από το εργαστήριο Νευρωνικών Δικτύων σαν συνέχεια αυτής της διατριβής. Ο βασικός στόχος είναι να μελετηθούν οι διαφορές στην απόδοση διαφορετικών ταξινομητών (knn, RBF κλπ.) για τον υπολογισμό των πιο σημαντικών κατευθύνσεων.

2. Ο ταξινομητής τύπου knn θα μπορούσε απλά να εφαρμοστεί στα δεδομένα, μετά την διαδικασία προεπεξεργασίας που προτείνουμε. Με αυτόν τον τρόπο είναι δυνατό να αξιολογήσουμε την απόδοση της μεθόδου SPCA, με ένα στατιστικό ταξινομητή όπως είναι ο knn, που είναι ξένος ως προς τη φύση της. Πράγματι στη παρούσα διατριβή, στο κεφάλαιο 5.4, δοκιμάζουμε μια τέτοια διαρρύθμιση.

5.4 Πειραματικά αποτελέσματα

Για να μπορέσουμε να δείξουμε την αποδοτικότητα των προτεινόμενων μεθόδων, θα διεξάγουμε συνολικά τέσσερα τεστ, εκ των οποίων ένα είναι συνθετικό πρόβλημα, και τα υπόλοιπα τρία αποτελούνται από προβλήματα πραγματικού κόσμου. Τα προβλήματα πραγματικού κόσμου βρέθηκαν στην έξοχη βάση δεδομένων προβλημάτων τεχνητής νοημοσύνης και μηχανικής μάθησης UCI [62].

Τεχνητό πρόβλημα: Το στραμμένο XOR Ας θεωρήσουμε P διανύσματα (x_1, x_2) τα οποία είναι ομογενώς κατανεμημένα στο τετράγωνο που ορίζεται από τα διαστήματα $-1 < x_1 < 1$ και $-1 < x_2 < 1$. Στο κλασσικό πρόβλημα XOR υπάρχουν δύο κλάσεις. Τα διανύσματα, των οποίων το γινόμενο των συνιστωσών $x_1 x_2 > 0$ ανήκουν στη κλάση 1, ενώ τα διανύσματα για τα οποία ισχύει $x_1 x_2 < 0$ ανήκουν στην κλάση 2. Προσθέτουμε έξι (6) επιπλέον χαρακτηριστικά, εισόδους δηλαδή στο πρόβλημα (x_3, x_4, x_5, x_6, x_7 και x_8), των οποίων οι τιμές είναι τυχαίες και ακολουθούν την ομογενή κατανομή στο διάστημα από $[-1, 1]$. Στη συνέχεια περιστρέφουμε κάθε πρότυπο εισόδου, στον οκταδιάστατο χώρο, κατά έναν αυθαίρετο πίνακα στροφής $\boldsymbol{A}$. Το πρόβλημα του στραμμένου XOR ορίζεται ως εξής: Ένα στραμμένο διάνυσμα $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ ανήκει στην τάξη 1 αν $x_1 x_2 > 0$ και στη τάξη 2 αν $x_1 x_2 < 0$.



Σχήμα 5.1: Το στραμμένο XOR με μια επιπλέον είσοδο θορύβου. Τα σημεία σημειώνονται από κύκλους και σταυρούς, ανάλογα σε ποια κλάση ανήκουν.

Το πρόβλημα εικονίζεται στο σχήμα 5.1, αλλά έχουμε προσθέσει μόνο μια επιπλέον μεταβλητή αντί για έξι για λόγους οπτικοποίησης. Πρέπει να υπογραμμιστεί το γεγονός ότι σ' αυτό το πρόβλημα όλα τα χαρακτηριστικά $y_i, i = 1, \ldots, 8$ παίζουν κάποιο ρόλο στο τελικό αποτέλεσμα. Αλλά μόνο δύο γραμμικοί συνδυασμοί απ' αυτά είναι πραγματικά σημαντικοί. Για την σύνθεση του προβλήματος του στραμμένου XOR επιλέξαμε να χρησιμοποιήσουμε 200 πρότυπα εισόδου.

Ιονοσφαιρικά δεδομένα Σ' αυτό το πρόβλημα τα δεδομένα προέρχονται από εκπομπή ραντάρ στην ιονόσφαιρα. Το πρόβλημα προτάθηκε και μελετήθηκε από τον Sigillito [48]. Αποτελείται από 351 διανύσματα εισόδου, το καθένα με 33 συνιστώσες.

Διαταραχές ήπατος BUPA Εδώ το πρόβλημα συνιστάται στο να ξεχωρίσει το δίκτυο τις περιπτώσεις με πιθανές διαταραχές ήπατος. Το πρόβλημα αποτελείται από 6 χαρακτηριστικά που έχουν συλλεγεί από αποτελέσματα εξετάσεων αίματος και δεδομένα καθημερινής κατανάλωσης οινοπνευματωδών ποτών. Το πρόβλημα αποτελείται συνολικά από 345 πρότυπα εισόδου, με 6 χαρακτηριστικά το καθένα.

Αναγνώριση διαβήτη σε ινδιάνους Pima Αυτό το πρόβλημα [63] αποτελείται από 768 πρότυπα εισόδου, τα οποία πάρθηκαν από ασθενείς που είχαν πιθανότητες να πάσχουν από διαβήτη. Κάθε διάνυσμα αποτελείται από 8 χαρακτηριστικά, και κατατάσσεται σε δύο κατηγορίες, ανάλογα με το αν ο ασθενής πάσχει από διαβήτη ή όχι.

Αναγνώριση στόχων sonar Αυτό είναι ένα πολύ γνωστό πρόβλημα διαχωρισμού ανακλώμενων σημάτων sonar σε δύο κλάσεις, δηλαδή σε μεταλλικούς κυλίνδρους[1] και βράχους. Χρησιμοποιούμε το αρχικό σύνολο δεδομένων που μελετήθηκε από τους Gorman και Sejnowski [29, 30], το οποίο αποτελείται από 208 διανύσματα εισόδου, το καθένα με 60 συνιστώσες. Σ' αυτό το πρόβλημα οι Gorman και

[1] νάρκες

Sejnowski ανέφεραν μόνο 85% επιτυχία για το μονοστρωματικό perceptron επί του συνόλου εκπαίδευσης, το οποίο μπορεί να γίνει 100% μόνο με την εισαγωγή 12 κρυμμένων κόμβων, σ' ένα ενδιάμεσο επίπεδο, στην αρχιτεκτονική του νευρωνικού δικτύου.

Για λόγους αξιοπιστίας των αποτελεσμάτων αλλά και για ευκολότερη σύγκριση, εκτός από τις μεθόδους που παρουσιάζονται σ' αυτό το κεφάλαιο, δοκιμάζουμε και άλλες μεθόδους επιλογής και εξαγωγής χαρακτηριστικών στα ίδια προβλήματα. Πιο συγκεκριμένα:

1. Τη μέθοδο του Ruck
2. Τη μέθοδο του Tarr
3. Μια μέθοδο βασισμένη στο γνωστό κριτήριο t-test του Student. Το κριτήριο αφορά στη διαφορά των δύο μέσων όρων ανά κατηγορία και ανά χαρακτηριστικό [64]. Σαν σημαντικότητα του κάθε ανεξάρτητου χαρακτηριστικού θεωρείται το t-score. Έτσι τα πιο σημαντικά χαρακτηριστικά είναι αυτά που έχουν πιο υψηλό t-score.
4. Ανάλυση σε πρωτεύοντες άξονες (PCA).

Για όλες τις προεκπαιδεύσεις του νευρωνικού δικτύου[2], τα αρχικά χαρακτηριστικά κανονικοποιήθηκαν έτσι ώστε να βρίσκονται στο διάστημα $[0, 1]$. Για να μπορέσουμε να υπολογίσουμε τη γενικευτική ικανότητα του δικτύου, το κάθε σύνολο δεδομένων διαμερίστηκε κατά 80% σε σύνολο εκπαίδευσης, και κατά 20% σε σύνολο δοκιμών. Το αρχικό πρόβλημα κατατμήθηκε μ' αυτές τις αναλογίες τριάντα (30) φορές, με τυχαίο τρόπο, έχοντας έτσι τριάντα διαφορετικές διαμερίσεις για σύνολα εκπαίδευσης και δοκιμής. Τα αποτελέσματα που αφορούν στη γενικευτική ικανότητα δίνονται σαν μέσος όρος αυτών των τριάντα φορών στο σύνολο δοκιμών.

Πρέπει να διευκρινιστεί ότι για όλες τις μεθόδους είτε επιλογής είτε εξαγωγής χαρακτηριστικών των οποίων τα αποτελέσματα παρουσιάζουμε, ο υπολογισμός της σημαντικότητας βασίζεται μόνο στην πληροφορία από το σύνολο εκπαίδευσης. Αυτό άλλωστε είναι και το προτιμότερο, γιατί στην περίπτωση που χρησιμοποιούνταν ολόκληρο το σύνολο, η υπολογιζόμενη σημαντικότητα των χαρακτηριστικών θα ήταν ισχυρά επηρεασμένη (biased) από τα σημεία του συνόλου δοκιμών.

Όλες οι προεκπαιδεύσεις, και όλες οι τελικές εκπαιδεύσεις, διεκπεραιώθηκαν χρησιμοποιώντας μια πιο αποδοτική παραλλαγή[3] του αλγορίθμου οπίσθιας διάδοσης (back - propagation), βασισμένη στην προσαρμοζόμενη χρήση της επιτάχυνσης μέσω της ορμής (momentum acceleration) [65]. Οι τιμές $\delta P = 0.3$ και $\xi = 0.5$ χρησιμοποιήθηκαν για όλα τα προβλήματα. Σ' όλες τις δοκιμές αφήσαμε την εκπαίδευση να προχωρήσει μέχρι 400 επαναλήψεις ή να ρίξει το σφάλμα μέσης τιμής κάτω από την τιμή $2 \cdot 10^{-3}$. Τα δίκτυα που χρησιμοποιήθηκαν ήταν πολυεπίπεδα, εμπρόσθιας διάδοσης (feedforward), και μ' ένα στρώμα ενδιάμεσων κόμβων (1 hidden layer). Για το στραμμένο XOR χρησιμοποιήθηκαν τέσσερις κρυμμένοι κόμβοι, για τα υπόλοιπα προβλήματα 10.

Για να υπολογιστεί η σημαντικότητα κάθε χαρακτηριστικού, με στατιστική αξιοπιστία, για κάθε διαφορετική διαμέριση γίνονται 10 διαφορετικές προεκπαιδεύσεις, ξεκινώντας κάθε φορά από διαφορετικά βάρη. Για τον υπολογισμό του πίνακα σημαντικότητας R_{jk}, προτάθηκαν (και υλοποιήθηκαν) διαφορετικές μέθοδοι για τον σχηματισμό του αθροίσματος 5.6. Μια απ' αυτές τις πιθανότητες ήταν η χρησιμοποίηση τυχαίων σημείων στον μοναδιαίο υπερκύβο[4], ενώ μια άλλη επιλογή αποτελούσαν τα ίδια τα πρότυπα εισόδου που ανήκαν στο σύνολο εκπαίδευσης. Το ενδιαφέρον είναι ότι αυτές οι δύο διαφορετικές μέθοδοι δίνουν παρόμοια αποτελέσματα, τουλάχιστον στα πειράματα που εμείς εκτελέσαμε. Έτσι σ' αυτήν εδώ την εργασία παρουσιάζουμε τα αποτελέσματα από την χρησιμοποίηση των διανυσμάτων που αποτελούν τον σύνολο εκπαίδευσης για τον σχηματισμό του αθροίσματος.

[2] Η προεκπαίδευση του δικτύου, χρησιμοποιείται και για την εξαγωγή της αρχικής γενικευτικής ικανότητας, και συμβολίζεται με τη λέξη 'Αρχικό' στους πίνακες που ακολουθούν

[3] **ALECO**: Algorithm for Learning Efficiently Constrained Optimization

[4] Ας μη ξεχνάμε ότι οι είσοδοι είναι ήδη κανονικοποιημένες.

	RXOR	Iono	BUPA	PIMA	Sonar
SPCA	91.3	95.4	72.4	75.30	79.20
Ruck	86.2	93.5	70.4	73.57	79.02
Tarr	84.6	92.8	69.4	73.57	79.02
t-test	86.0	92.71	69.32	73.57	79.35
PCA	86.1	95.4	69.6	75.22	83.44
Αρχικό	87.3	94.5	70.5	73.57	79.02

Πίνακας 5.1: Γενικευτική ικανότητα στο σύνολο δοκιμών. Εδώ παρουσιάζονται οι μέσοι όροι που προκύπτουν από 30 διαμερίσεις και 10 επανεκκινήσεις με διαφορετικά αρχικά βάρη για κάθε διαμέριση.

	RXOR	Iono	BUPA	PIMA	Sonar
SPCA	3.40	2.21	5.70	2.56	3.93
Αρχικό	3.21	2.63	5.44	2.23	5.83
p-value	$1.66 \cdot 10^{-9}$	0.013	0.110	0.004	0.445

Πίνακας 5.2: Η τυπική απόκλιση από το μέσο όρο της γενικευτικής ικανότητας. Η τυπική απόκλιση αποτελεί πάντα ένα αξιόπιστο κριτήριο για το μέσο όρο της κατανομής. Η τελευταία γραμμή παρουσιάζει την τιμή p-value η οποία αποτελεί ένα μέτρο ελέγχου για το κατά πόσο όντως ανέβηκε η γενικευτική ικανότητα, ή η αύξηση οφείλεται σε στατιστικό θόρυβο. Όσο μικρότερη η τιμή του, τόσο πιο αξιόπιστα τα αποτελέσματα.

Για να μπορέσουμε να καθορίσουμε τώρα τον αριθμό των σημαντικών χαρακτηριστικών που πρέπει να κρατήσουμε, εφαρμόσαμε την τεχνική των Belue και Bauer. Η μέθοδος αυτή δεν είχε καθόλου άσχημα αποτελέσματα όταν χρησιμοποιούνταν σε συνδυασμό με τις μεθόδους Tarr, Ruck και t-test. Σ' αυτές τις περιπτώσεις ο αριθμός των σημαντικών χαρακτηριστικών που προέκυπτε χρησιμοποιώντας σαν κατώφλι τη μέση τιμή της σημαντικότητας του θορύβου, επαυξημένο με την απόκλιση της κανονικής κατανομής, ήταν συγκρίσιμος με το βέλτιστο αριθμό χαρακτηριστικών ο οποίος έδινε μέγιστη απόδοση σε όρους ταξινόμησης όσον αφορά πάντα στο σύνολο δοκιμών. Για την προτεινόμενη μέθοδο όμως, τα πειράματα έδειξαν ότι η τεχνική των Belue και Bauer έτεινε να υπερεκτιμά τον βέλτιστο αριθμό χαρακτηριστικών. Ήταν απαραίτητο να χρησιμοποιούμε ένα κατώφλι το οποίο απείχε τέσσερις με έξι φορές την απόκλιση πάνω από την μέση τιμή για να έχουμε αποδεκτά αποτελέσματα. Είναι σαφές ότι μια τέτοια συμπεριφορά δεν είναι άμεσα εξηγήσιμη και σαφώς χρήζει περισσότερης και πιο εκτεταμένης μελέτης.

Τα αποτελέσματα των πειραμάτων συγκεντρώνονται στους πίνακες 5.1, 5.2 και 5.3. Η γενικευτική ικανότητα παρουσιάζεται στον πίνακα 5.1, με τη μορφή μέσων όρων επιτυχούς ταξινόμησης στο σύνολο δοκιμών. Στον πίνακα 5.2 εικονίζεται η τυπική απόκλιση της γενικευτικής ικανότητας από το μέσο όρο. Επειδή είναι φανερό ότι οι κατανομές επικαλύπτονται, παραθέτουμε και την p-value ως τεκμήριο ότι υπάρχει, με στατιστική αξιοπιστία, βελτίωση στη γενικευτική ικανότητα με την προτεινόμενη μέθοδο. Η p-value υπολογίζεται με την υπόθεση ότι οι παραδοχές που κάνει η μέθοδος t-test ισχύουν. Στην πραγματικότητα η τιμή της p-value είναι ένα μέτρο για την πιθανότητα δύο κατανομές να έχουν τον ίδιο μέσο όρο. Στον πίνακα 5.3 παραθέτουμε τον βέλτιστο αριθμό χαρακτηριστικών που επιλέχθηκε ή εξάχθηκε από την κάθε μέθοδο για το κάθε πρόβλημα.

Σ' όλα τα προβλήματα, εκτός του sonar, είναι σαφές ότι η μέθοδος μας δίνει, με στατιστική αξιοπιστία, αυξημένη γενικευτική ικανότητα σε σχέση με το σύνολο των αρχικών χαρακτηριστικών. Σε τρία προβλήματα (RXOR, BUPA, PIMA) η προτεινόμενη μέθοδος παρουσίασε την καλύτερη γενικευτική ικανότητα απ' όλες τις άλλες μεθόδους, ενώ στο πρόβλημα με τα ιονοσφαιρικά δεδομένα ήρθε δεύτερη πολύ κοντά στη μέθοδο του Tarr. Φυσικά τα καλύτερα αποτελέσματα τα είχε το RXOR, το οποίο είναι βέβαια ένα τεχνητό πρόβλημα που στόχος του ήταν να αναδείξει την χρησιμότητα της μεθόδου, και είναι γνωστό εκ κατασκευής ότι τα πραγματικά σημαντικά χαρακτηριστικά του είναι γραμμικοί συνδυασμοί ενός υποσύνολου των αρχικών χαρακτηριστικών.

	RXOR	Iono	BUPA	PIMA	Sonar
SPCA	3	4	2	3	18
Ruck	7	12	4	8	60
Tarr	7	8	5	8	60
t-test	7	12	6	8	30
PCA	7	18	5	6	20
Αρχικό	8	33	6	8	60

Πίνακας 5.3: Ο αριθμός των εξαγόμενων σημαντικών χαρακτηριστικών, όπως προκύπτει από την κάθε μέθοδο σ' όλα τα προβλήματα.

	RXOR	Iono	BUPA
SPCA	92.5	94.3	70.9
Ruck	67.5	91.4	65.1
Tarr	67.5	90.3	59.3
PCA	70.0	93.7	63.4
Αρχικό	65.0	93.1	61.6

Πίνακας 5.4: Γενικευτική ικανότητα όπως προκύπτει από τη μέθοδο πλησιέστερου γείτονα (Knn).

Γνωρίζοντας ότι το sonar πρόβλημα είναι γραμμικά διαχωρίσιμο [51, 13], είναι προφανές ότι μόνο ένας γραμμικός συνδυασμός των εισόδων είναι κατάλληλος για να επιτύχει πλήρη διαχωρισμό των κλάσεων. Πράγματι, το πιο σημαντικό χαρακτηριστικό, σύμφωνα με τη μέθοδό μας, έχει πολύ μεγάλη διαφορά από το δεύτερο. Συγκεκριμένα, η ιδιοτιμή του αποτελεί το 85% του αθροίσματος των ιδιοτιμών όλων των χαρακτηριστικών, ενώ η δεύτερη μεγαλύτερη ακολουθεί με 2.5%. Ακόμα όμως και αν αφήναμε το δίκτυο να επανεκπαιδευτεί μ' ένα χαρακτηριστικό, τότε θα είχε 78.80% απόδοση στη γενικευτική του ικανότητα. Η βέλτιστη απόδοση (79.20%) επιτεύχθηκε φυσικά, όπως φαίνεται και στους σχετικούς πίνακες, κρατώντας συνολικά 18 χαρακτηριστικά.

Επίσης, και στα πέντε υπό δοκιμή προβλήματα, η προτεινόμενη μέθοδος έχει επιτύχει να αποδώσει αρκετά καλά έχοντας εξάγει ένα σχετικά μικρό αριθμό χαρακτηριστικών, σε σχέση με άλλες μεθόδους, γεγονός που την κάνει κατάλληλη για εφαρμογές οπτικοποίησης όπως θα δούμε και πιο αναλυτικά στο κεφάλαιο 5.5. Είναι φανερό από τον πίνακα 5.3, ότι αυτό δεν συμβαίνει και με τις άλλες μεθόδους που πολλές φορές ο προτεινόμενος αριθμός χαρακτηριστικών ισούται με το πλήθος των αρχικών χαρακτηριστικών. Έτσι μείωση στα επιλεγμένα ή με άλλο τρόπο εξαγόμενα χαρακτηριστικά, οδηγεί μοιραία και σε μείωση της γενικευτικής ικανότητας.

Ο αλγόριθμος SPCA αποδίδει καλά και στην περίπτωση που χρησιμοποιηθεί ταξινομητής τύπου knn για την τελική κατάταξη των διανυσμάτων, όπως φαίνεται στον πίνακα 5.4. Ένα σημαντικό πλεονέκτημα της προτεινόμενης μεθόδου είναι ότι με την σημαντική μείωση των διαστάσεων που επιτρέπει, ανοίγει τον δρόμο για τεχνικές γρήγορων υλοποιήσεων knn (fast knn). Αυτές οι τεχνικές χωρίζονται στις ακριβείς, με την έννοια ότι βρίσκουν πάντα τους κοντινότερους γείτονες [66], και στις μη ακριβείς, στις οποίες θυσιάζεται ένα συγκεκριμένο ποσοστό της ακρίβειας των κοντινότερων γειτόνων με αντάλλαγμα καλύτερη απόδοση [67]. Αν ο αριθμός των διαστάσεων είναι μεγάλος και οι δύο κατηγορίες γρήγορων knn χάνουν τα πλεονεκτήματα τους [68]. Οι ακριβείς μέθοδοι είναι δυνατόν να έχουν χειρότερη απόδοση από την αρχική υλοποίηση του knn λόγω εκτεταμένης οπισθοδρόμησης (backtracking), ενώ στις άλλες το ποσοστό πιθανής αστοχίας ανεβαίνει σε μη αποδεκτά επίπεδα. Είναι σαφές λοιπόν ότι η προτεινόμενη μέθοδος είναι ιδιαίτερα ελκυστική και στην περίπτωση που ο ταξινομητής είναι knn, αφού μπορεί να εξοικονομήσει μερικές τάξεις μεγέθους σε ταχύτητα, χωρίς αντίστοιχη πτώση σε ακρίβεια και γενικευτική ικανότητα.

5.5 Οπτικοποίηση

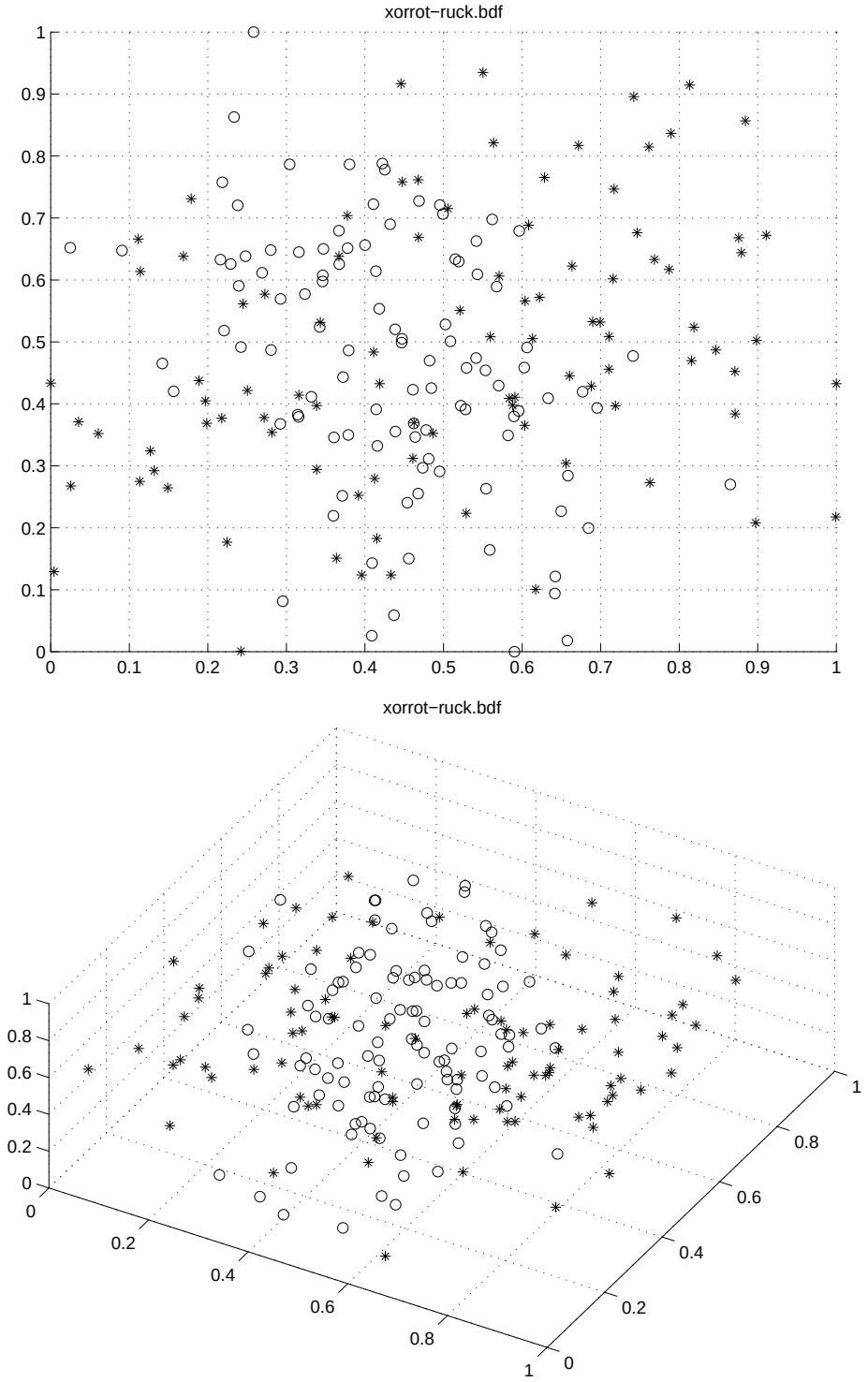
Με τον όρο 'οπτικοποίηση' εννοούμε συνήθως την γραφική αναπαράσταση αριθμητικών δεδομένων. Εφ' όσον τα μέσα αποθήκευσης και επίδειξης είναι από τη φύση τους δισδιάστατα , είναι φανερό ότι επαρκούν για οπτικοποίηση δισδιάστατων δεδομένων. Στην πράξη με μια ιδιότυπη προβολή, μπορούμε να ξεγελάσουμε το μάτι του πρόθυμου να συνεργαστεί αναγνώστη, και να δημιουργήσουμε την ψευδαίσθηση της προοπτικής σε τρισδιάστατα δεδομένα. Για οπτικοποίηση δεδομένων περισσότερων διαστάσεων συνήθως καταφεύγει κανείς σε ορθογραφικές προβολές, επιλέγοντας ουσιαστικά ένα τμήμα των δεδομένων για να απεικονίσει.

Ένα μεγάλο μέρος της αποδοχής των νευρωνικών δικτύων έχει να κάνει με το γεγονός ότι μπορούν να χρησιμοποιηθούν σαν γενικού τύπου ταξινομητές (Univeral Classifiers) χωρίς να είναι γνωστός ο τρόπος λειτουργίας τους (black box). Ένας τέτοιος τρόπος λειτουργίας είναι επιθυμητός όταν το πρόβλημα είναι πολυδιάστατο και δεν παρέχεται γ' αυτό κανένας τρόπος οπτικής επισκόπησης. Αν όμως υπήρχε τρόπος επισκόπησης πολυδιάστατων δεδομένων τότε θα υπήρχε και τρόπος αξιολόγησης της μαθησιακής διαδικασίας των νευρωνικών δικτύων.

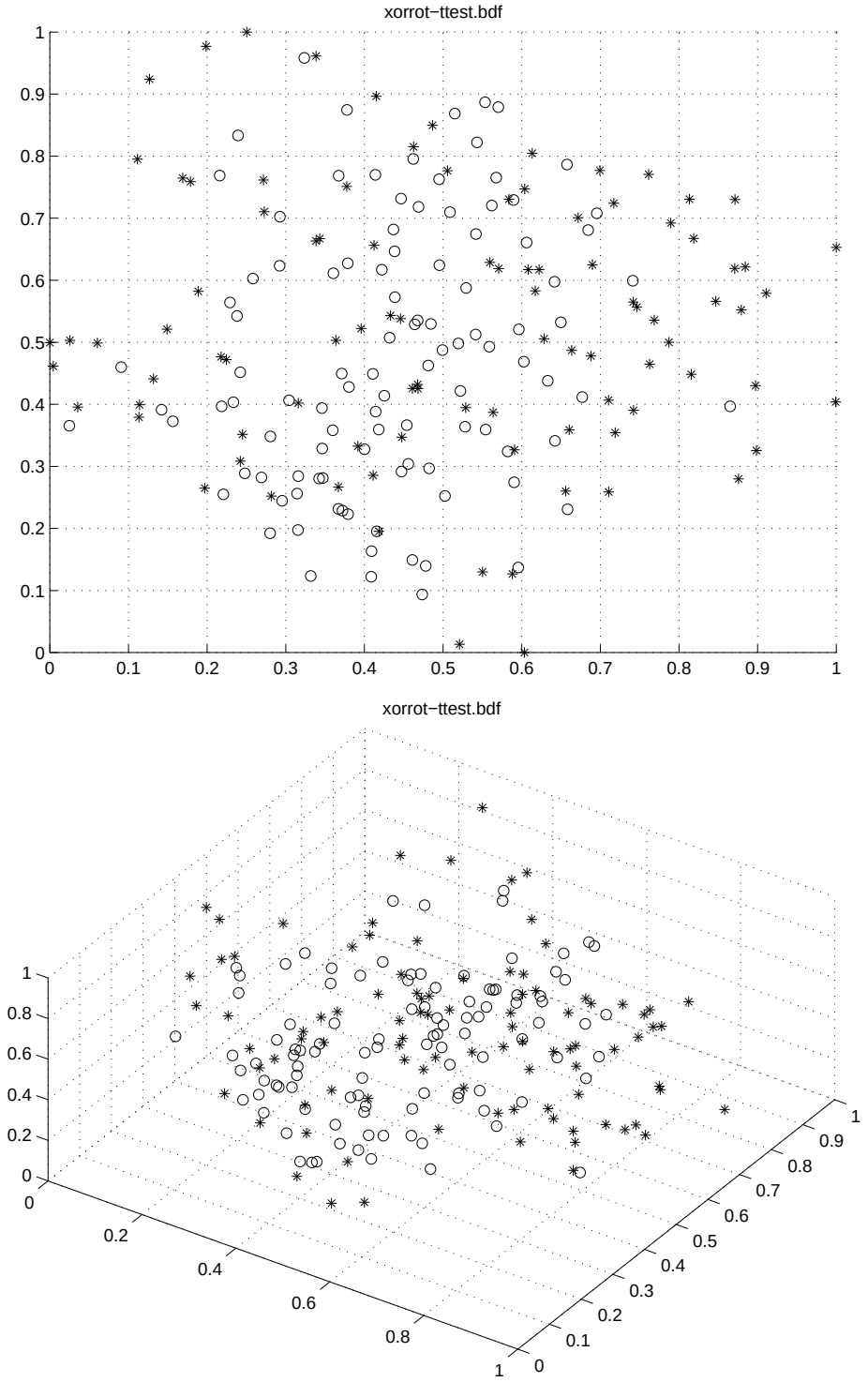
Μέχρι σήμερα έχουν γίνει αρκετές προσπάθειες οπτικοποίησης των νευρωνικών δικτύων, αλλά καμμία δεν φαίνεται να έχει τύχει της ευρύτερης αποδοχής της επιστημονικής κοινότητας [69, 70, 71, 72]. Αντίθετα γίνεται περισσότερο σαφές ότι είναι απαραίτητη μια ταυτόχρονη αντιμετώπιση του προβλήματος, [73, 74], οπτικοποιώντας ταυτόχρονα και τα δεδομένα, αλλά και το αντίστοιχο εκπαιδευμένο δίκτυο.

Με την προτεινόμενη μέθοδο προεπεξεργασίας ο αριθμός των εισόδων του προβλήματος μειώνεται δραματικά, χωρίς να υπάρχει αντίστοιχη μείωση στη γενικευτική ικανότητα του δικτύου. Αυτό σημαίνει ότι η αρχική πληροφορία διατηρείται σε μεγάλο ποσοστό, σε σχέση με το αρχικό πρόβλημα. Αν όμως το πρόβλημα δεν έχει αλλάξει, και ο αριθμός των εισόδων του δεν ξεπερνάει τις δύο ή τρεις, τότε μπορούμε να επιστρατεύσουμε τις συνήθεις πρακτικές για την απεικόνισή του.

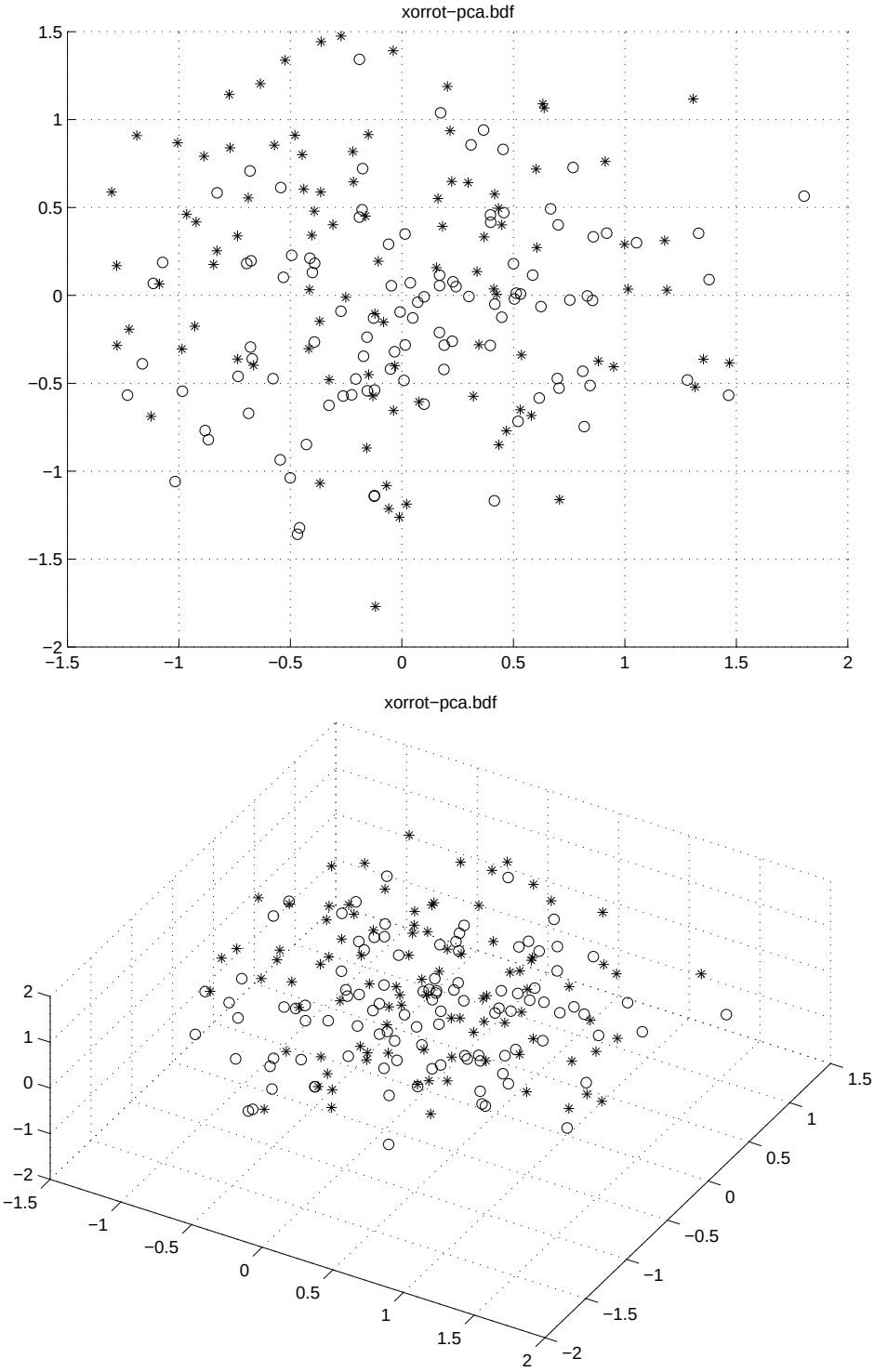
Παρακάτω απεικονίζουμε τις δύο, και τρεις πιο σημαντικές συνιστώσες όπως αυτές προκύπτουν από την κάθε μέθοδο. Οι μέθοδοι που συγκρίνονται είναι οι Ruck, PCA, t-test και η προτεινόμενη SPCA. Οι μέθοδοι του Ruck και η μέθοδος t-test μπορεί κανείς να θεωρήσει ότι αποτελούν ορθογραφική προβολή των δεδομένων στις επιλεγμένες συνιστώσες, ενώ οι PCA και SPCA χρησιμοποιούν γραμμικούς συνδυασμούς των αρχικών συνιστωσών σαν άξονες. Όπως μπορεί να δει κανείς, η μέθοδος SPCA είχε πολύ καλά αποτελέσματα στα σύνολα δεδομένων RXOR, IONO και sonar και όχι χειρότερα από τις άλλες μεθόδους στα υπόλοιπα προβλήματα (BUPA, PIMA). Για το RXOR η καλή απόδοση είναι αναμενόμενη αφού είναι ένα πρόβλημα που δημιουργήθηκε ειδικά για δοκιμή με τη συγκεκριμένη μέθοδο. Στα άλλα προβλήματα όμως είχαμε μη αναμενόμενη εξαιρετική απόδοση αφού όπως μπορούμε να δούμε στην περίπτωση του sonar ο παρατηρητής σχηματίζει την εντύπωση ότι είναι γραμμικά διαχωρίσιμο, πράγμα που είναι απολύτως αληθές, ενώ στην περίπτωση των ιονοσφαιρικών δεδομένων τα διανύσματα που ανήκουν σε διαφορετικές κατηγορίες έχουν ξεκάθαρα διαχωριστεί σε διαφορετικά συσσωματώματα.



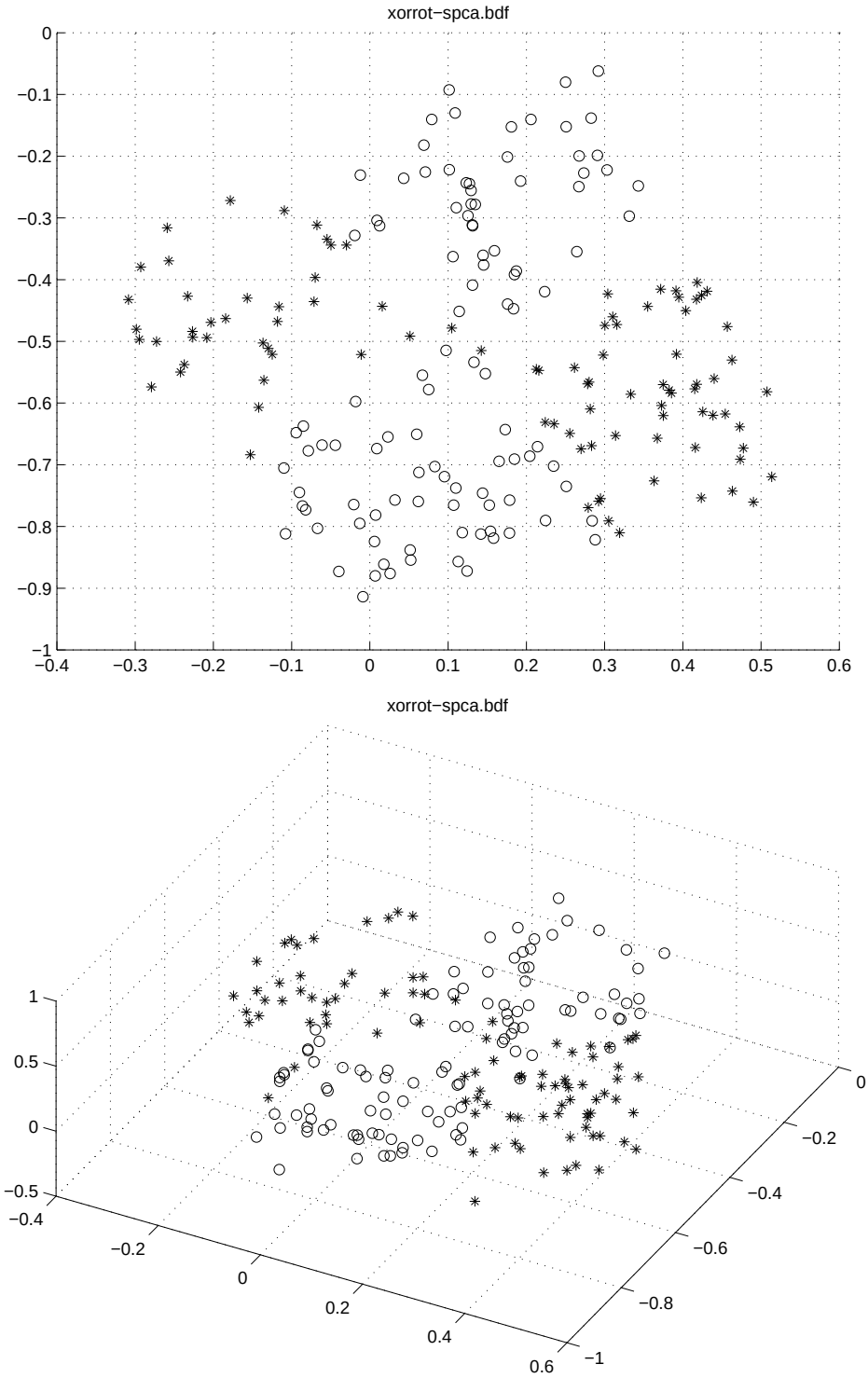
Σχήμα 5.2: Το RXOR οπτικοποιημένο με τη μέθοδο Ruck.



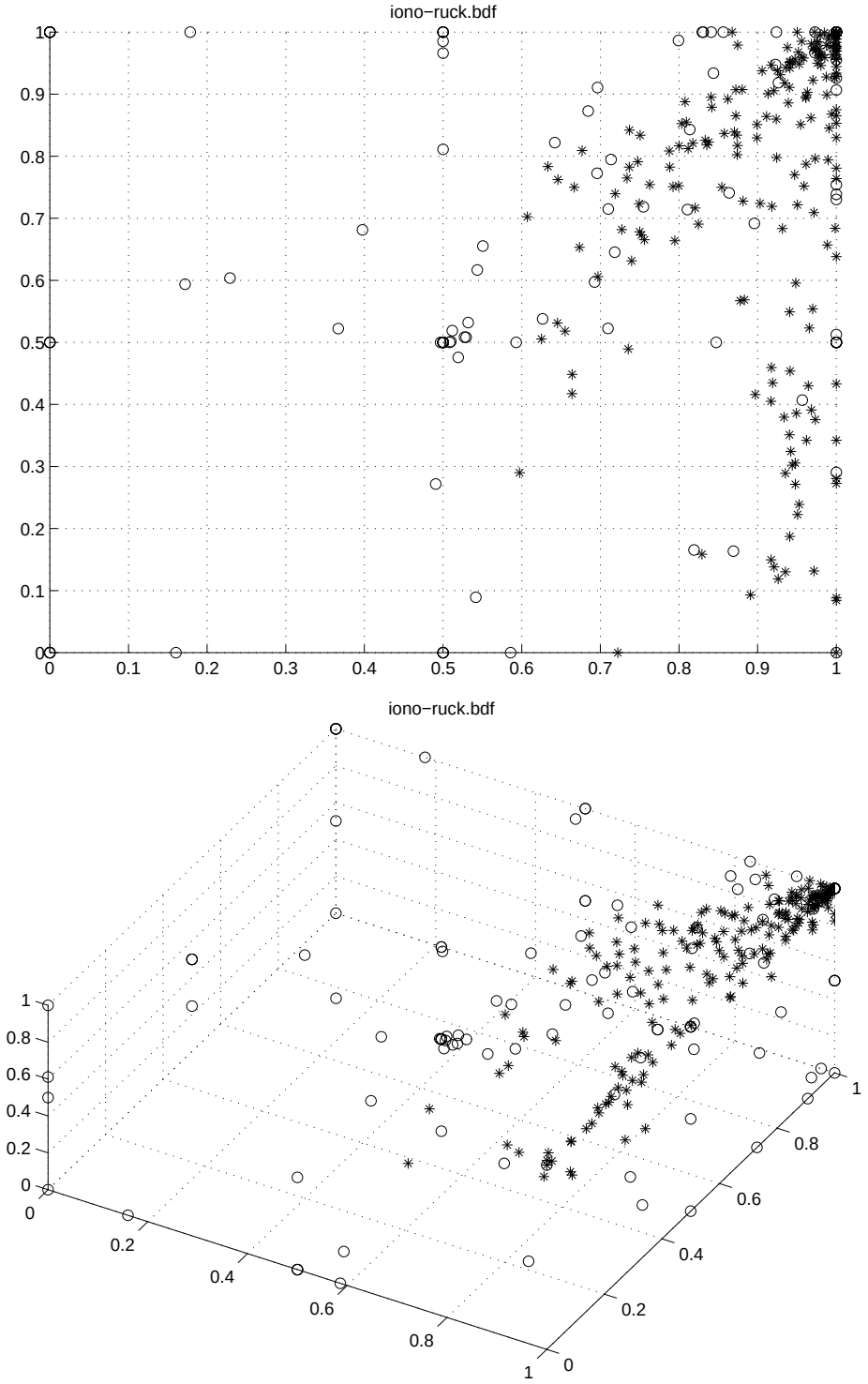
Σχήμα 5.3: Το RXOR οπτικοποιημένο με τη μέθοδο t-test.



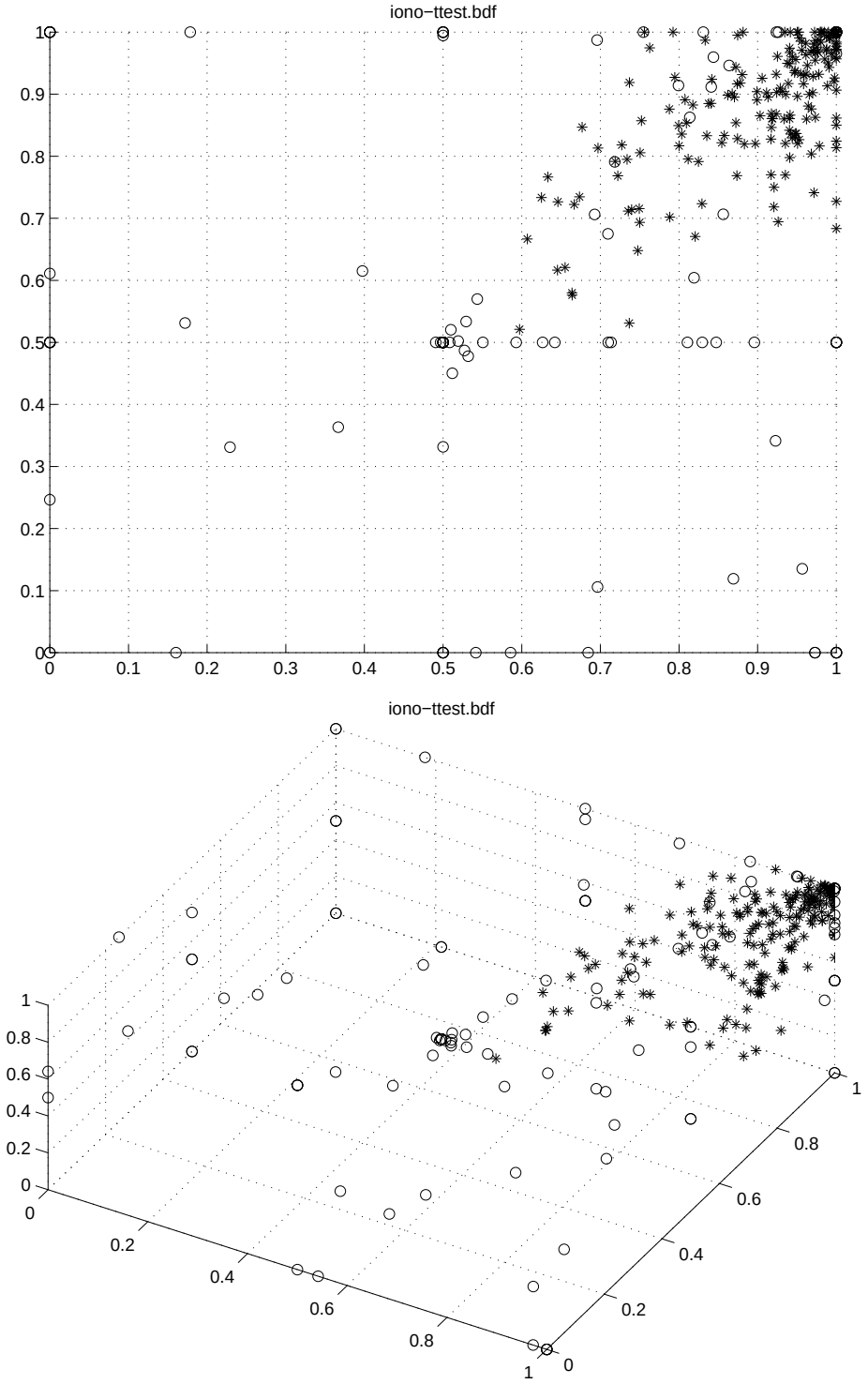
Σχήμα 5.4: Το RXOR οπτικοποιημένο με τη μέθοδο PCA.



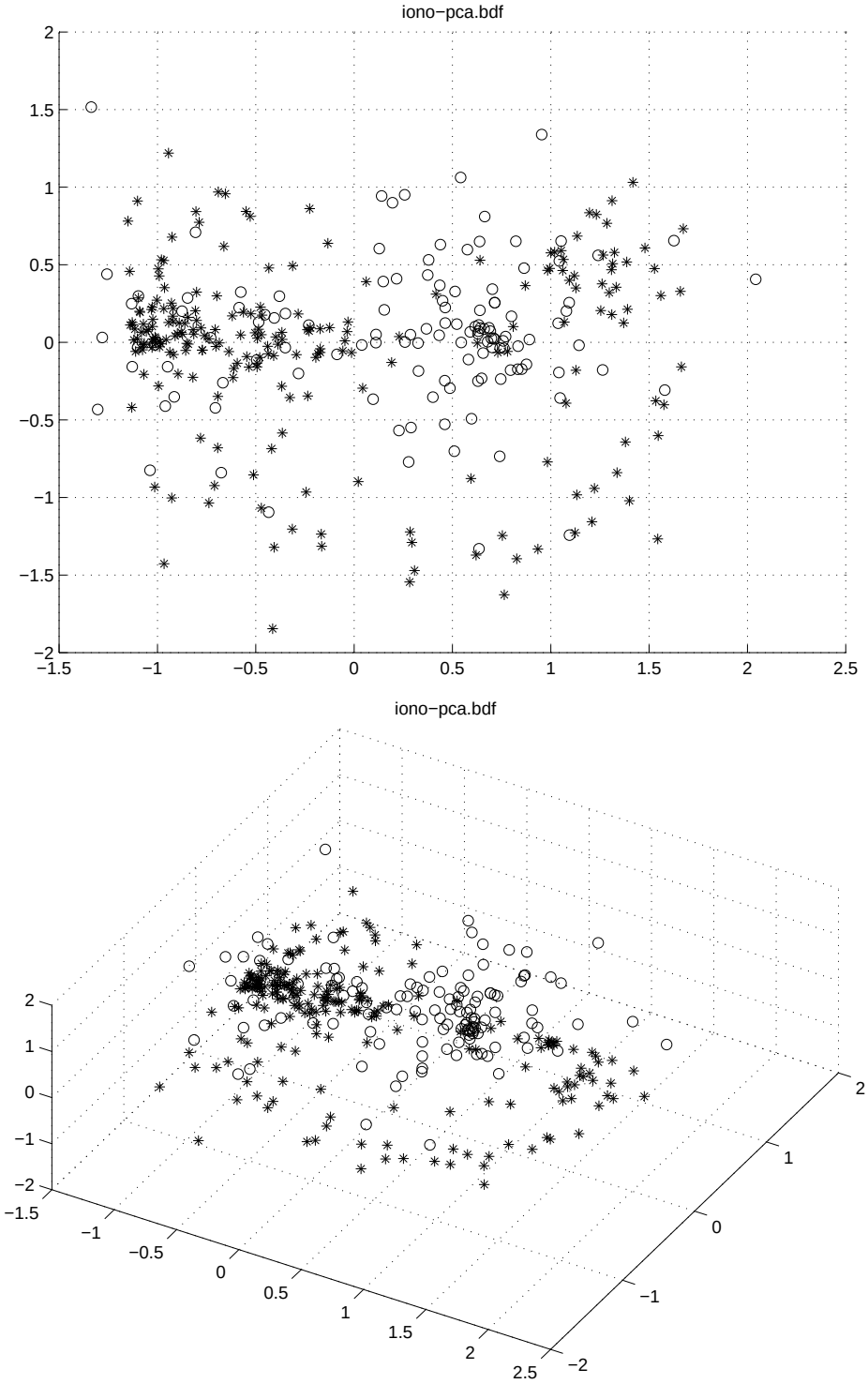
Σχήμα 5.5: Το RXOR οπτικοποιημένο με τη μέθοδο SPCA.



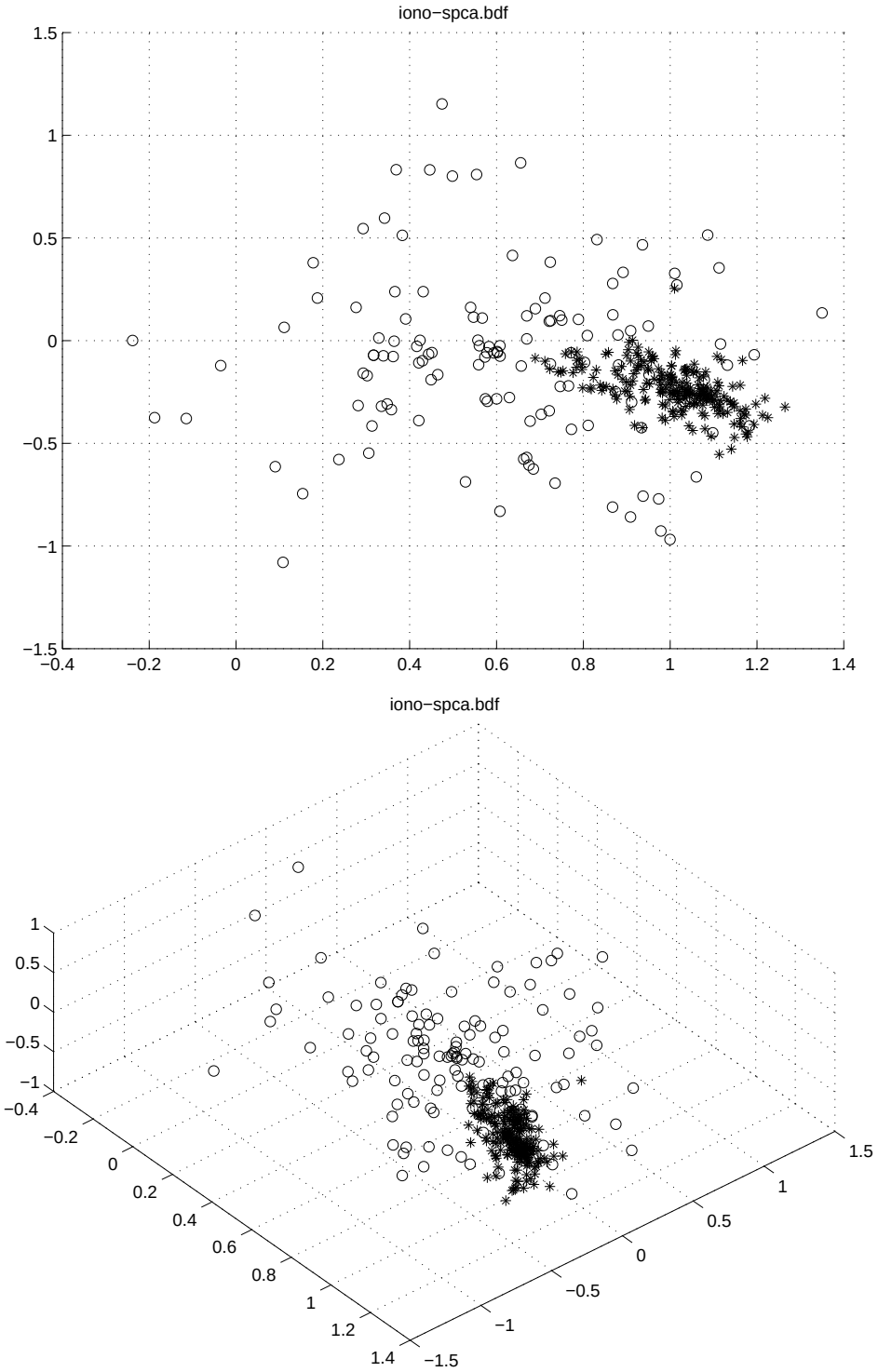
Σχήμα 5.6: Το IONO οπτικοποιημένο με τη μέθοδο Ruck.



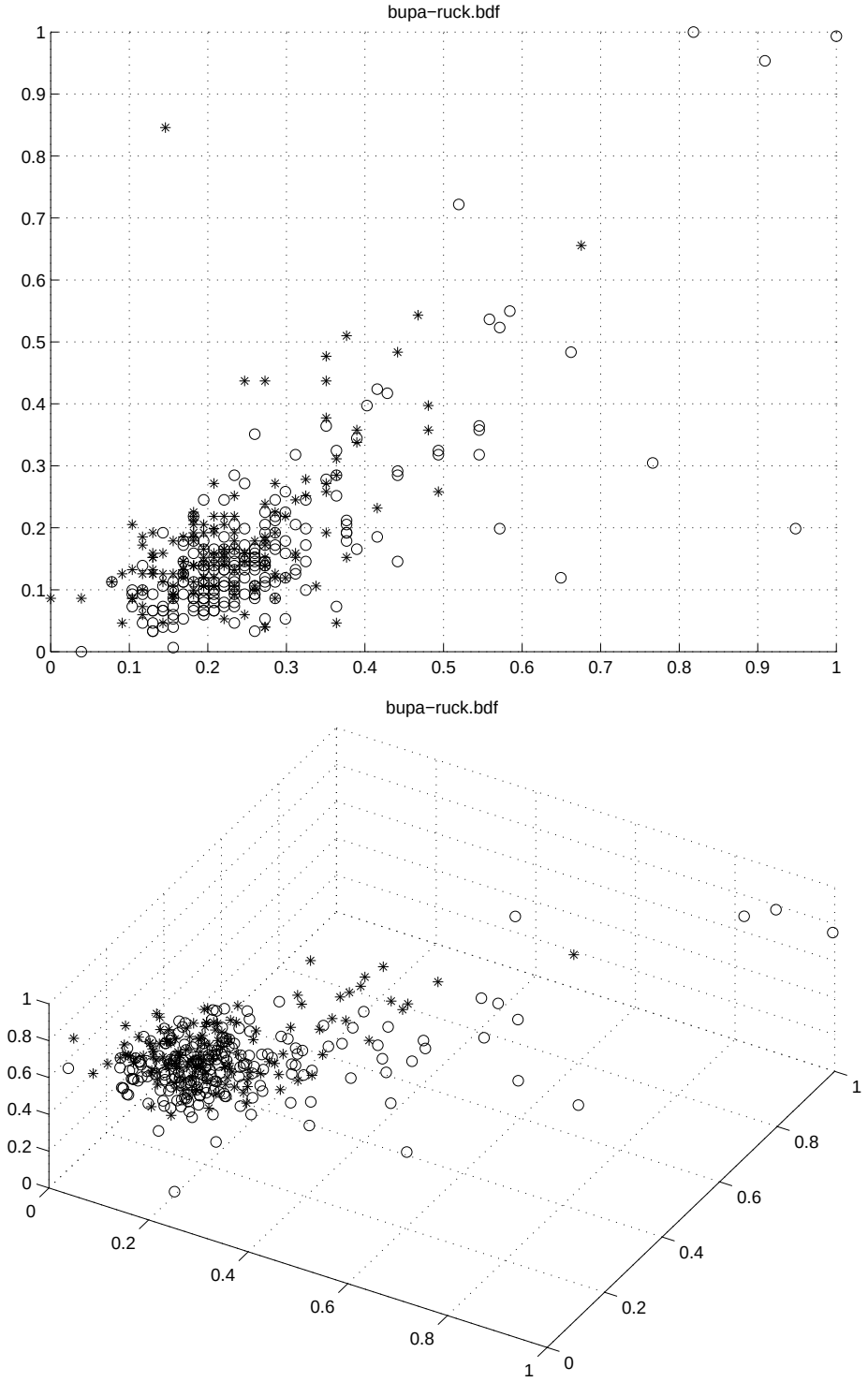
Σχήμα 5.7: Το IONO οπτικοποιημένο με τη μέθοδο t-test.



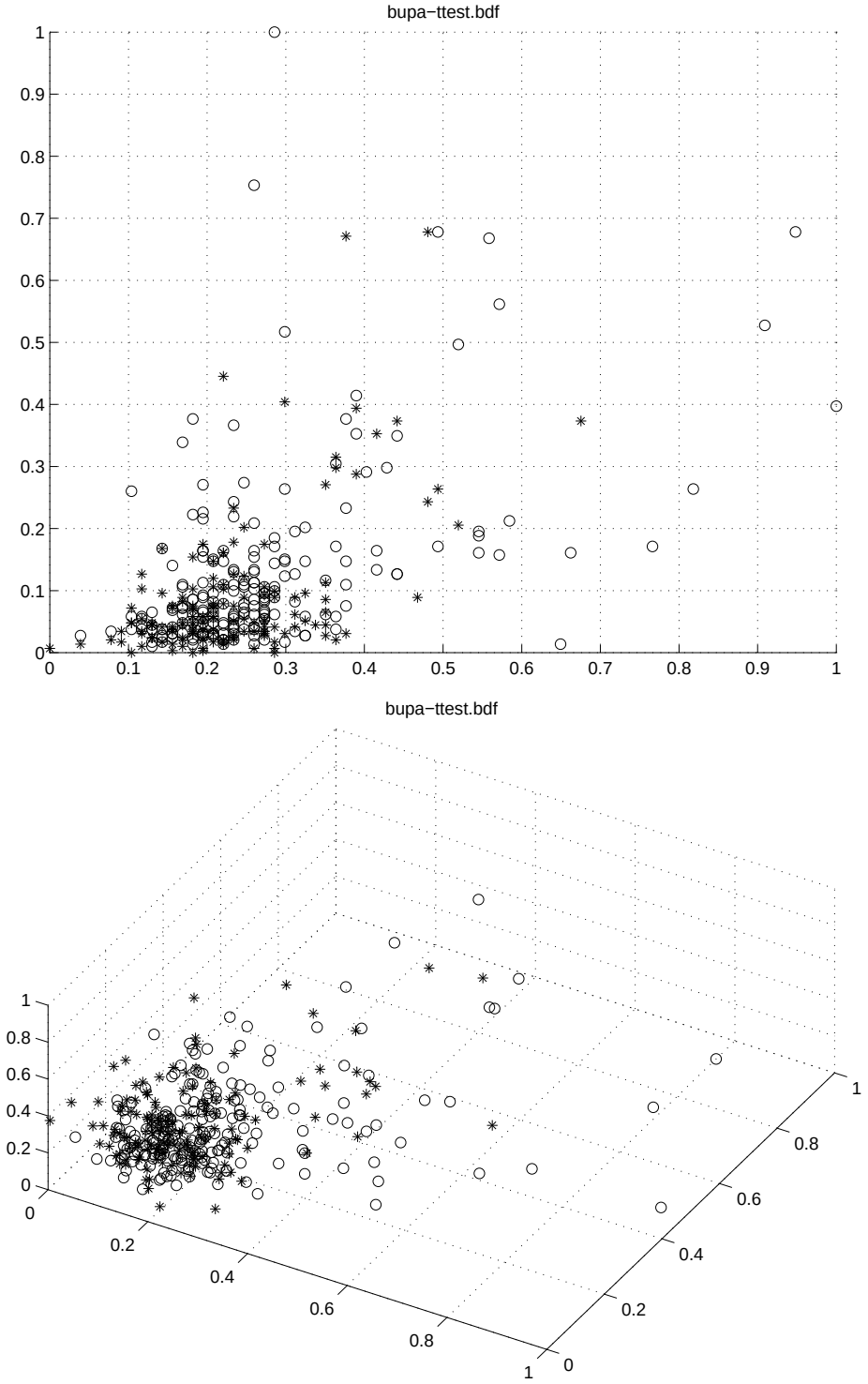
Σχήμα 5.8: Το IONO οπτικοποιημένο με τη μέθοδο PCA.



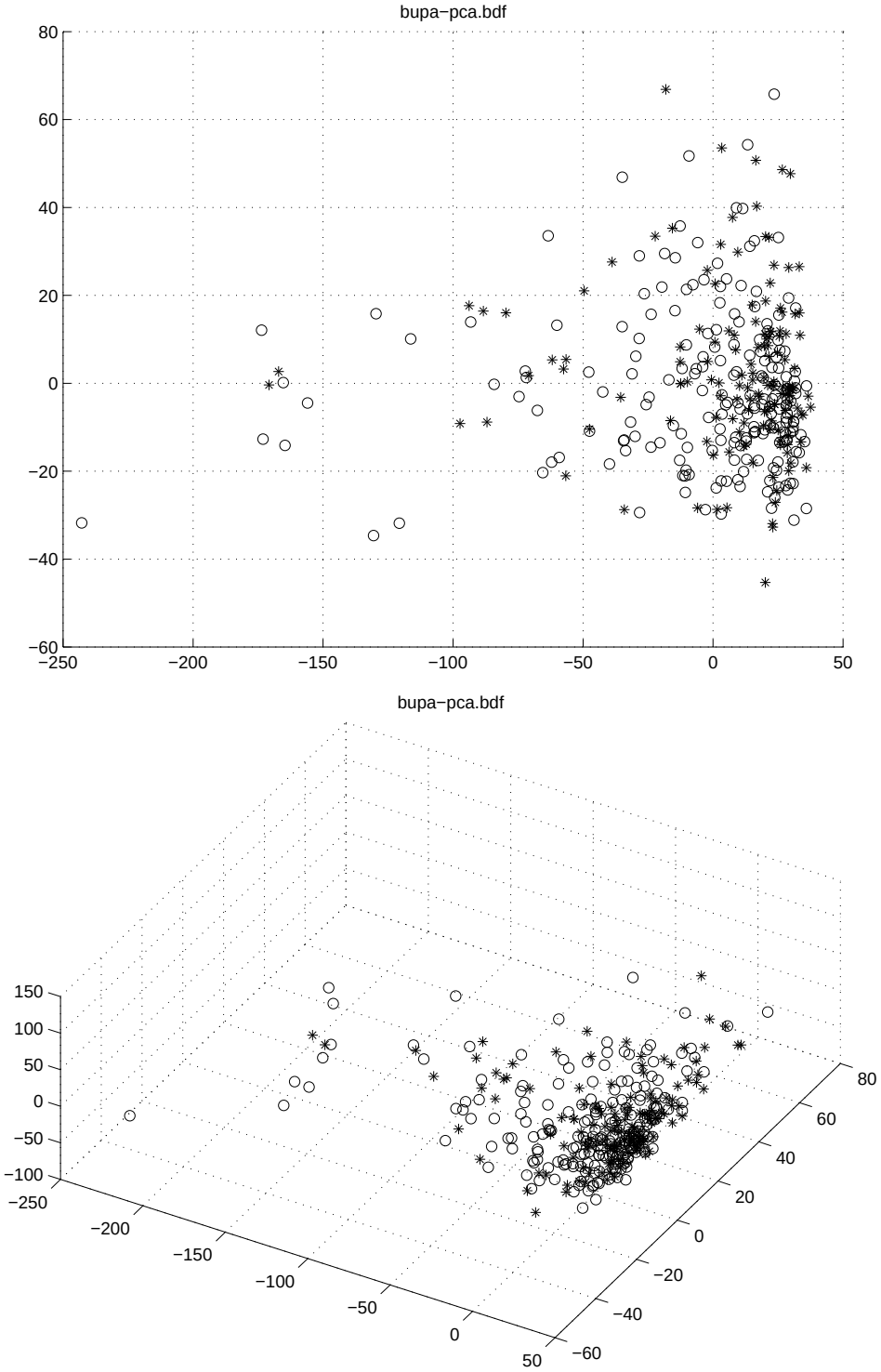
Σχήμα 5.9: Το IONO οπτικοποιημένο με τη μέθοδο SPCA.



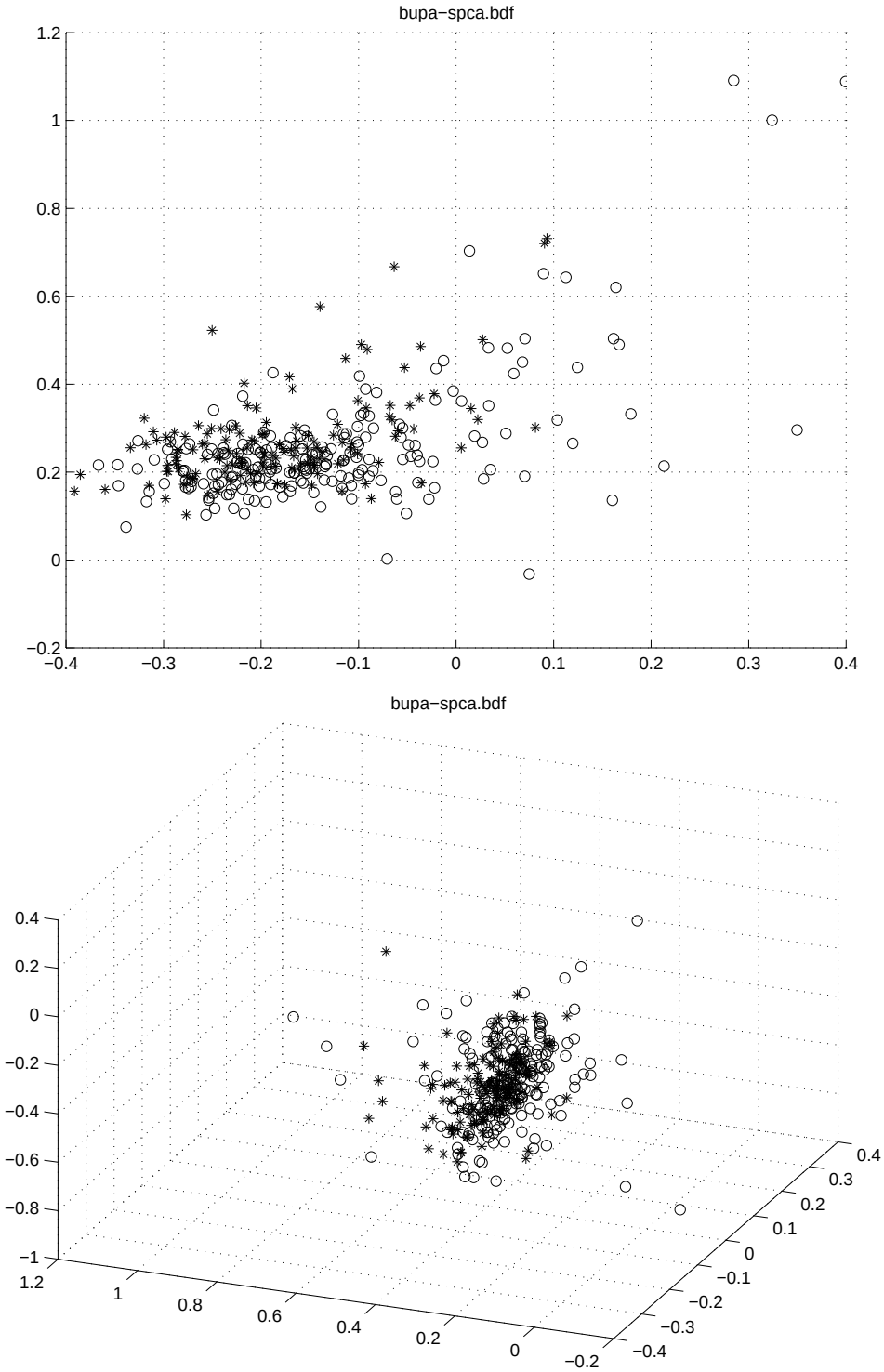
Σχήμα 5.10: Το BUPA οπτικοποιημένο με τη μέθοδο Ruck.



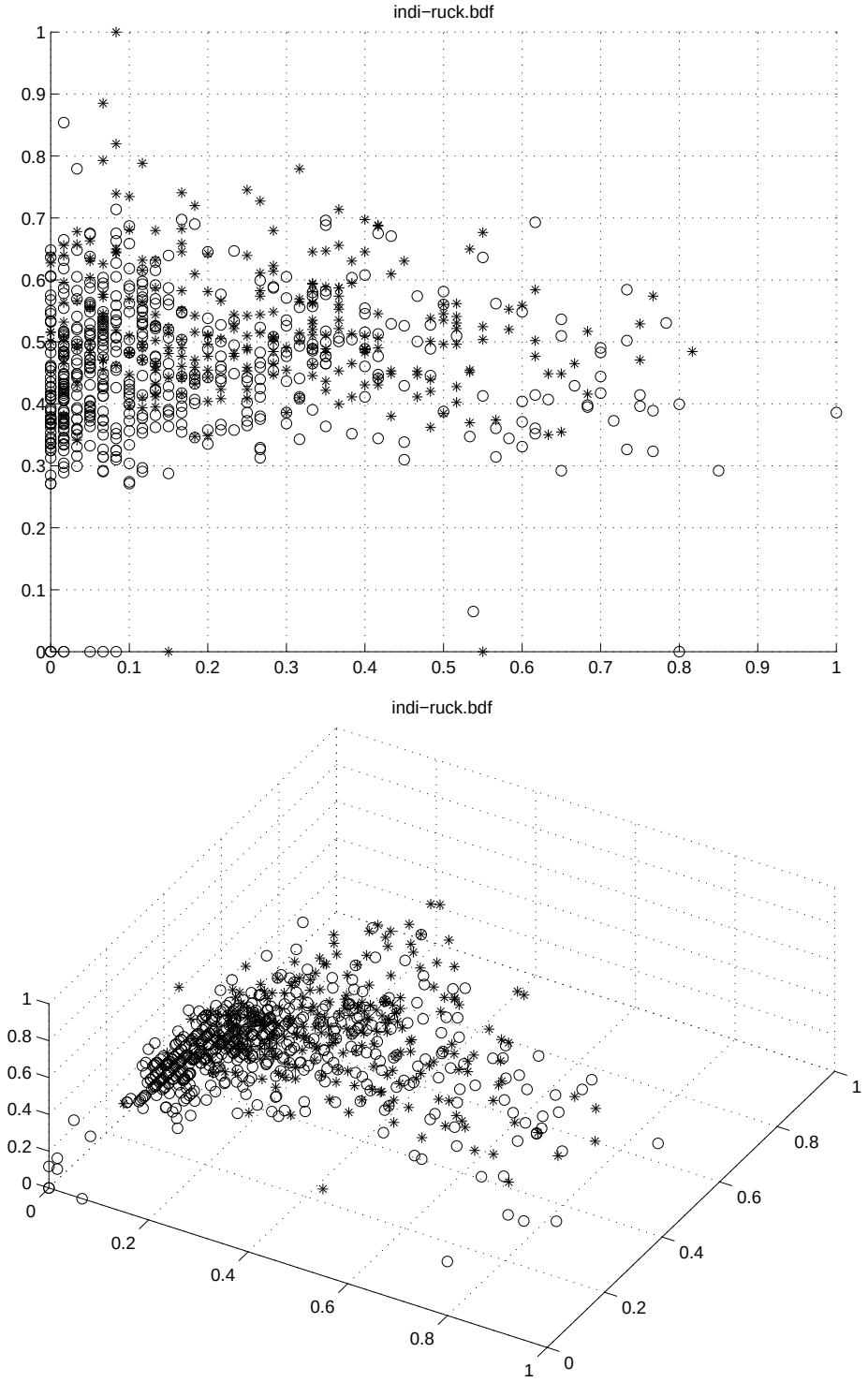
Σχήμα 5.11: Το BUPA οπτικοποιημένο με τη μέθοδο t-test.



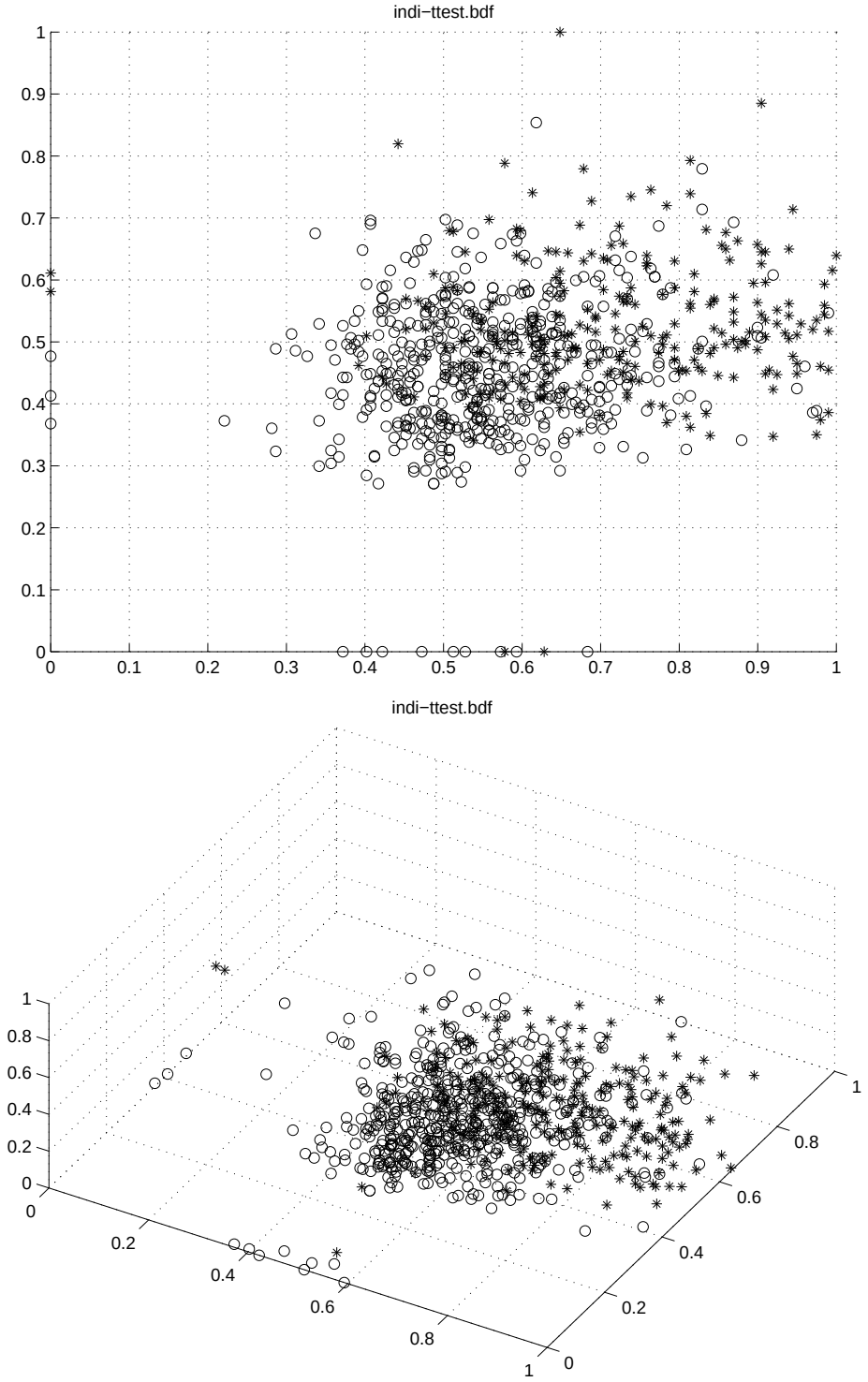
Σχήμα 5.12: Το BUPA οπτικοποιημένο με τη μέθοδο PCA.



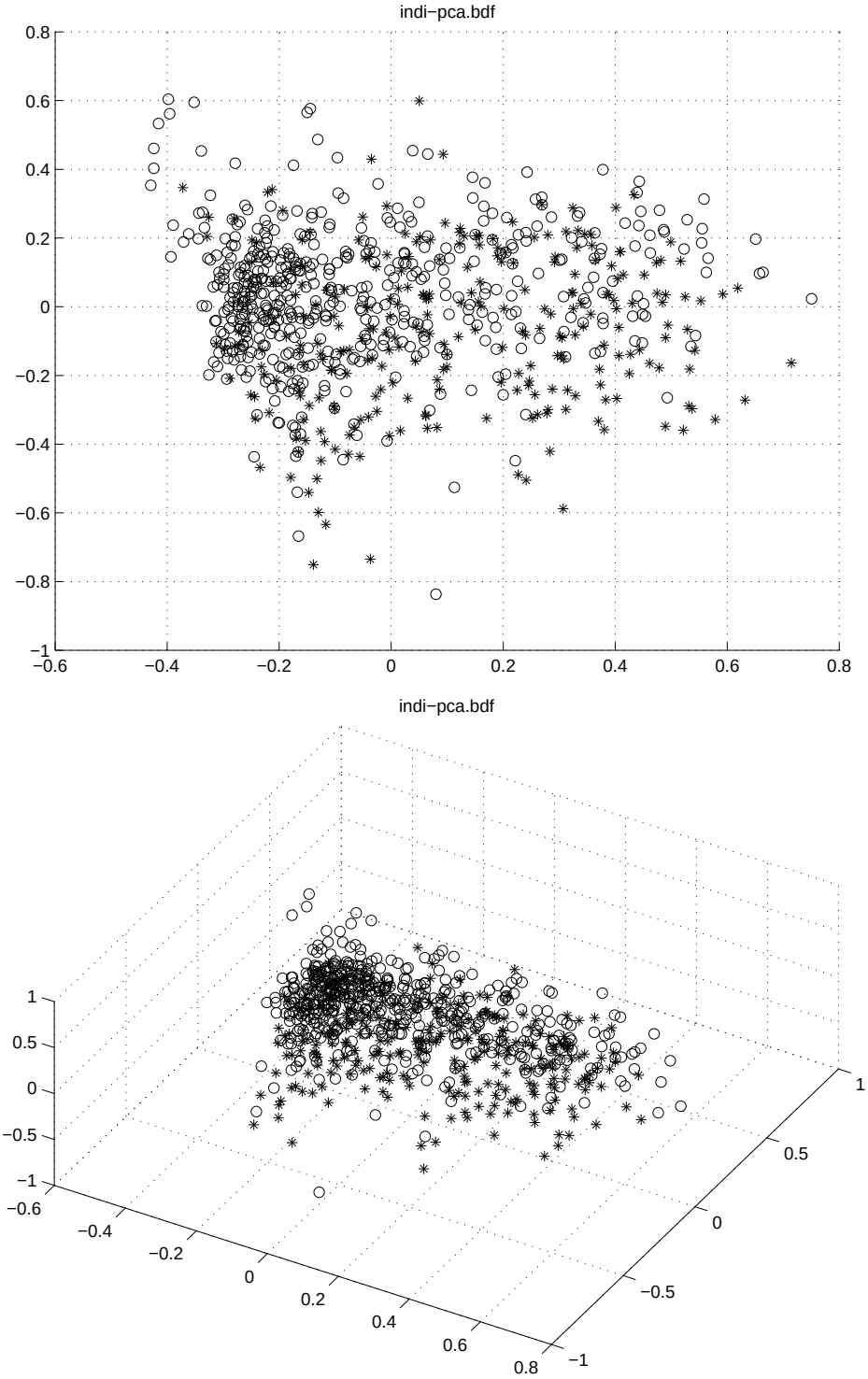
Σχήμα 5.13: Το BUPA οπτικοποιημένο με τη μέθοδο SPCA.



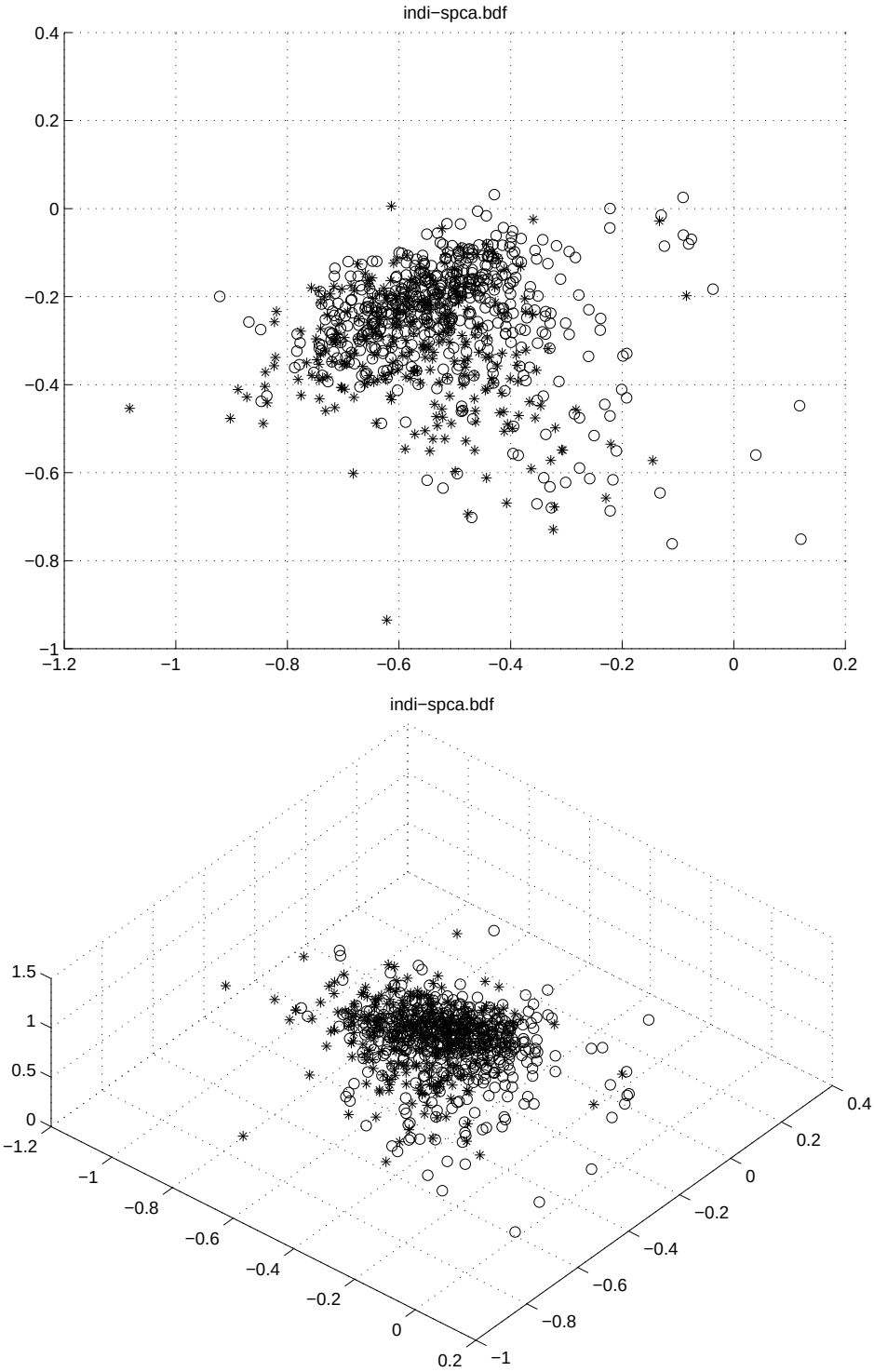
Σχήμα 5.14: Το PIMA οπτικοποιημένο με τη μέθοδο Ruck.



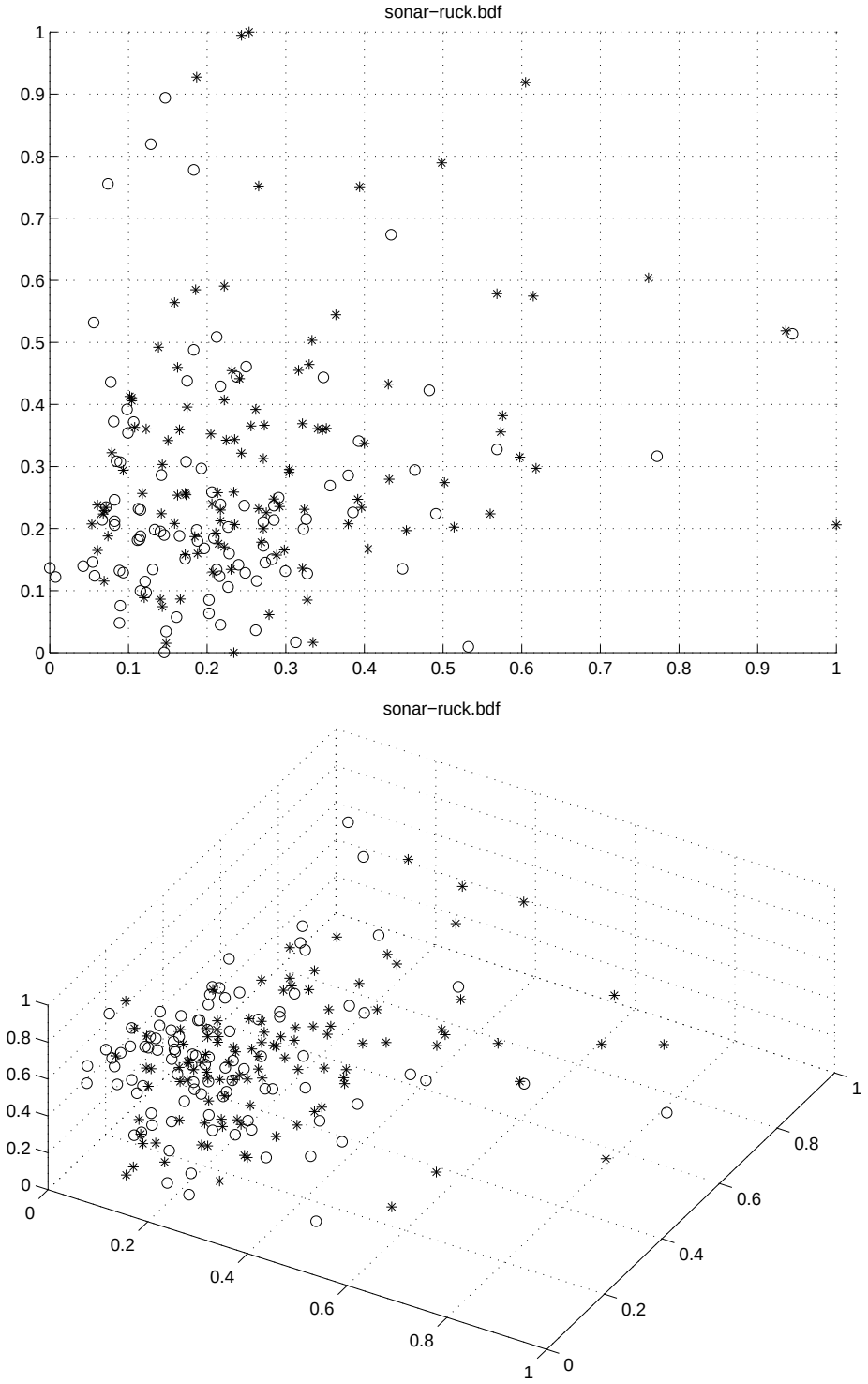
Σχήμα 5.15: Το PIMA οπτικοποιημένο με τη μέθοδο t-test.



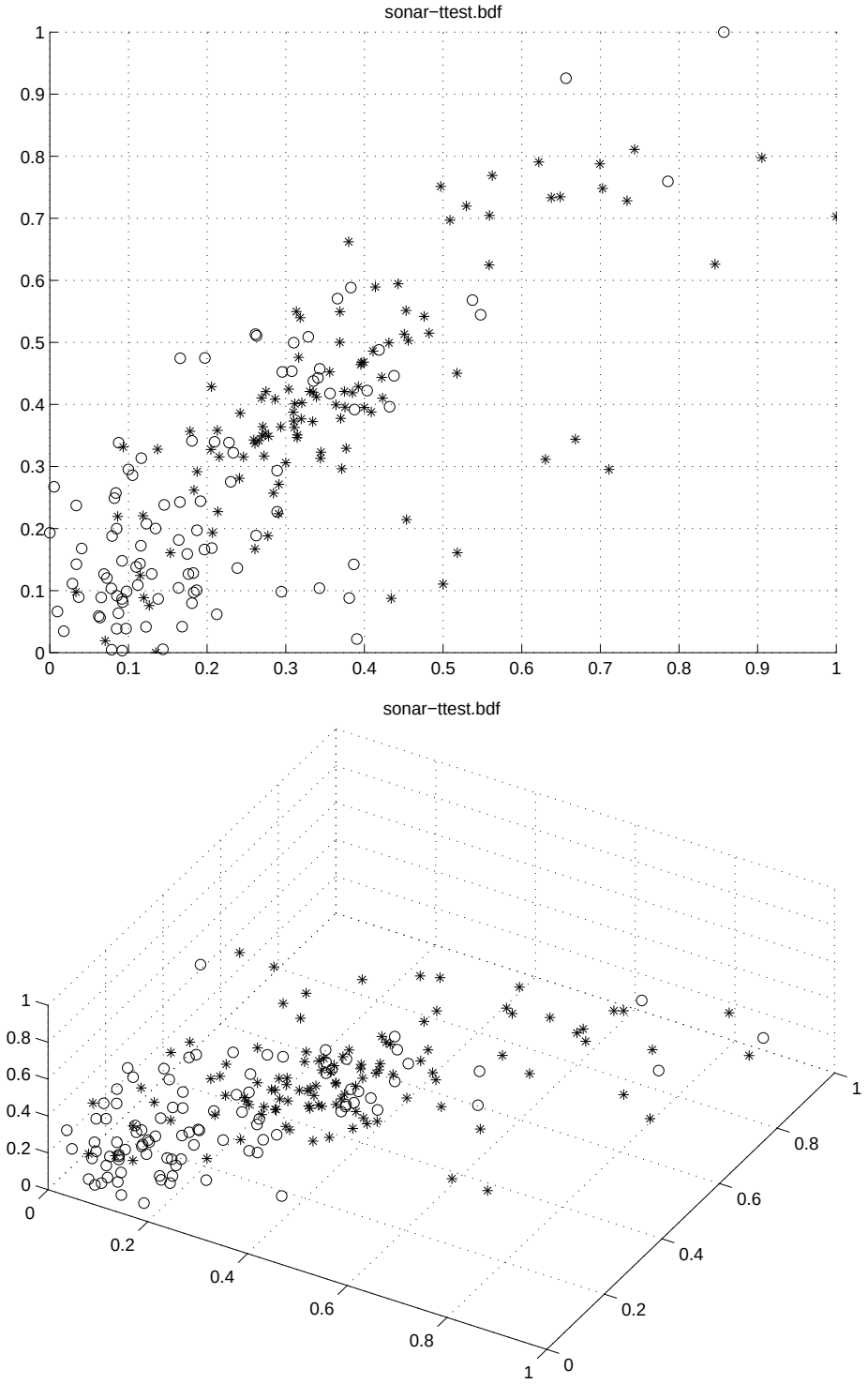
Σχήμα 5.16: Το PIMA οπτικοποιημένο με τη μέθοδο PCA.



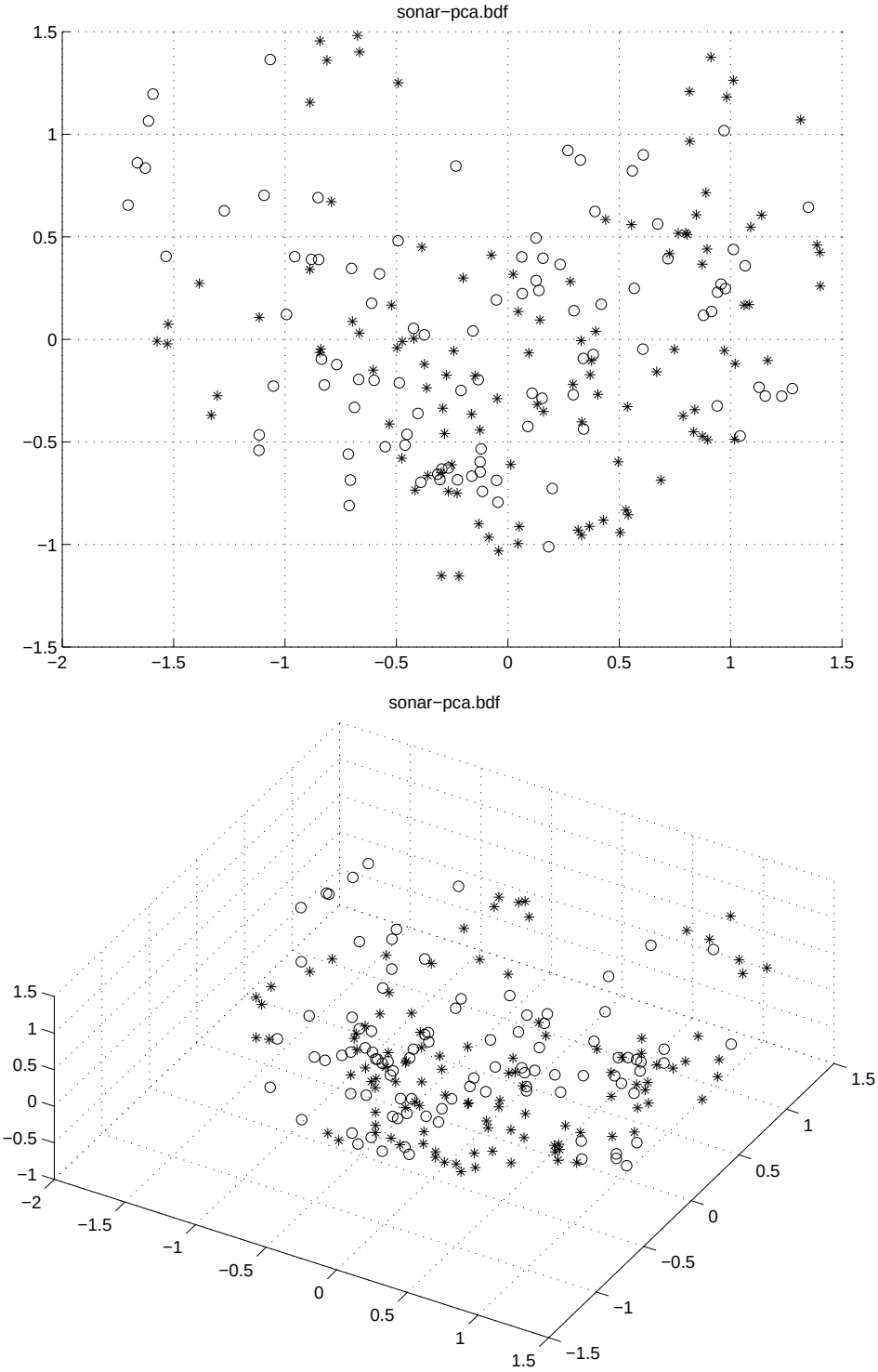
Σχήμα 5.17: Το PIMA οπτικοποιημένο με τη μέθοδο SPCA.



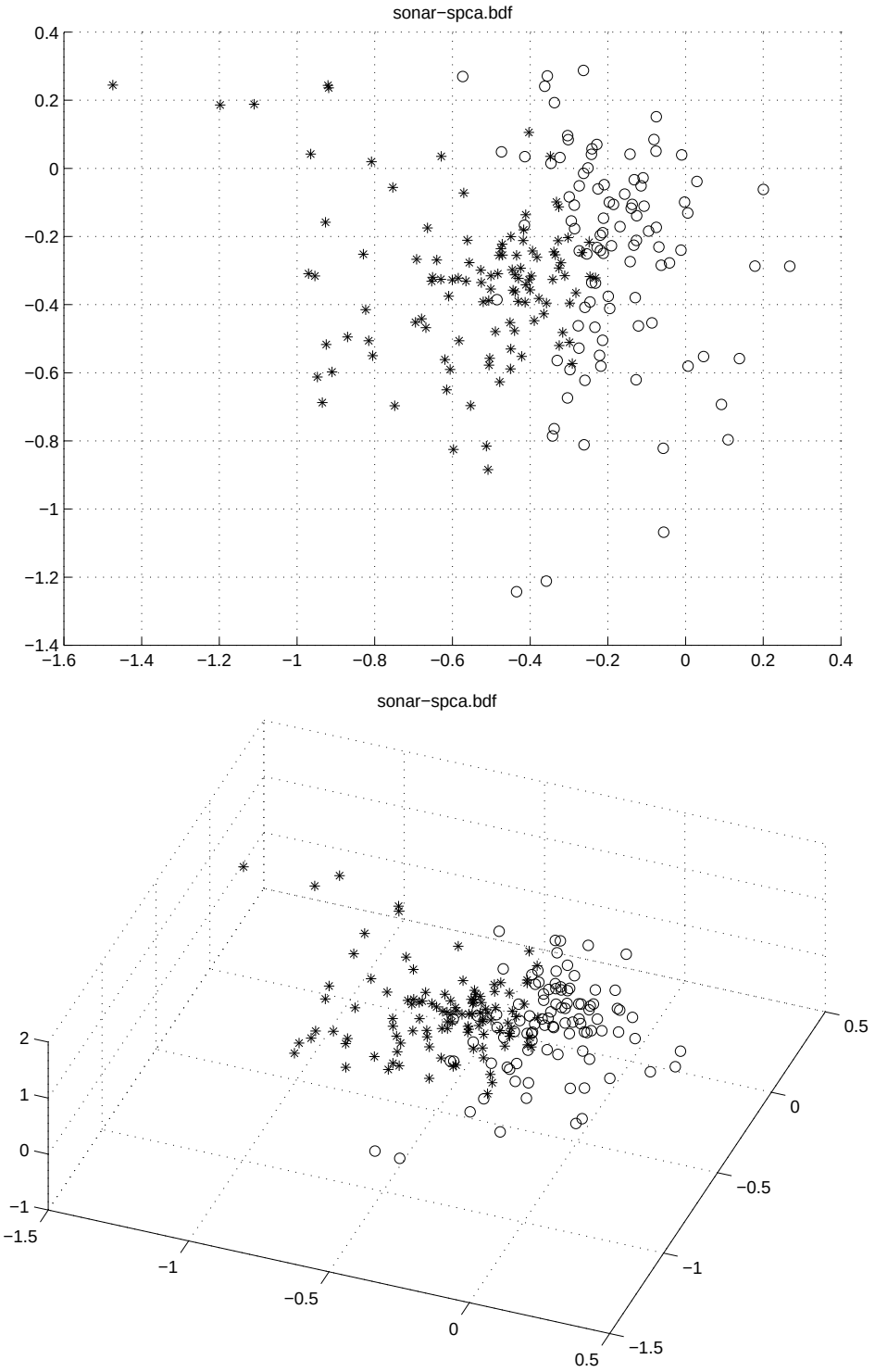
Σχήμα 5.18: Το sonar οπτικοποιημένο με τη μέθοδο Ruck.



Σχήμα 5.19: Το sonar οπτικοποιημένο με τη μέθοδο t-test.



Σχήμα 5.20: Το sonar οπτικοποιημένο με τη μέθοδο PCA.



Σχήμα 5.21: Το sonar οπτικοποιημένο με τη μέθοδο SPCA.

5.6 Συμπεράσματα

Σ' αυτό το κεφάλαιο προτείνουμε μια πρωτότυπη μέθοδο για εξαγωγή χαρακτηριστικών από ένα σύνολο προτύπων το οποίο χρησιμοποιείται για εποπτευόμενη μάθηση από ένα νευρωνικό δίκτυο. Η διαδικασία περιλαμβάνει προεκπαίδευση ενός νευρωνικού δικτύου με το αρχικό σύνολο χαρακτηριστικών. Στη συνέχεια, δημιουργούνται νέα χαρακτηριστικά που αποτελούν γραμμικούς συνδυασμούς των παλιών, τα οποία έχουν την ιδιότητα να μεγιστοποιούν την ευαισθησία του νευρωνικού δικτύου σε σχέση με μικρές μεταβολές των εισόδων (παράγωγος κατά κατεύθυνση). Η προτεινόμενη μέθοδος θυμίζει αρκετά στο φορμαλισμό της τη μέθοδο PCA αλλά λαμβάνει υπόψη της και την πληροφορία της εποπτευόμενης εκπαίδευσης. Γι αυτό το λόγο η μέθοδος ονομάζεται εποπτευόμενη ανάλυση σε κύριους άξονες (Supervised PCA ή SPCA).

Η μέθοδος εφαρμόστηκε σ' έναν αριθμό από τεχνητά και πραγματικού κόσμου προβλήματα και παρείχε ικανοποιητική βελτίωση στη γενικευτική ικανότητα των χρησιμοποιούμενων δικτύων, αλλά και σε αλγορίθμους τύπου πλησιέστερου γείτονα (Knn). Συγκεκριμένα, είναι γνωστό στην επιστημονική κοινότητα ότι οι βελτιστοποιημένες τεχνικές Knn (Fast Knn) απαιτούν μικρό αριθμό διαστάσεων (συχνά κάτω από 15), και η μέθοδος μας καταφέρνει να διατηρεί αρκετή πληροφορία περιορίζοντας δραστικά τη διάσταση του προβλήματος. Επίσης παρουσιάζονται συγκριτικά αποτελέσματα και από άλλες μεθόδους που έχουν παραπλήσιους στόχους.

Το ερευνητικό πεδίο στη συγκεκριμένη κατεύθυνση όμως παραμένει ανοιχτό αφού εκτός από περισσότερες δοκιμές θα μπορούσαν πολλές από τις ιδέες που παρουσιάσαμε εδώ να επεκταθούν και σε άλλου τύπου τεχνικές μάθησης (π.χ. RBF ή Knn), οι οποίες μπορούν επιπλέον να οδηγήσουν σε περαιτέρω μείωση των δεδομένων και αύξηση της γενικευτικής ικανότητας.

Παράρτημα Α΄

billnet

Ο σκοπός αυτού του παραρτήματος είναι χρησιμεύσει σαν εισαγωγή στο πακέτο billnet, το οποίο αποτελείται από έναν αρθρωτό νευρωνικό εξομοιωτή και συνοδευτικά προγράμματα που μπορούν να επιτελέσουν ένα ευρύ φάσμα προεπεξεργασίας δεδομένων. Το billnet μπορεί να χρησιμοποιηθεί για εκπαίδευση τεχνητών νευρωνικών δικτύων, για εφαρμογές πραγματικών προβλημάτων, και για ανάπτυξη νέων αλγορίθμων. Μπορεί δηλαδή να χρησιμοποιηθεί για εκπαιδευτικούς, βιομηχανικούς, αλλά και ερευνητικούς σκοπούς. Το γεγονός ότι επιτρέπει πλήθος ρυθμίσεων αφήνει την απόφαση όλων των πιθανών βελτιστοποιήσεων στον τελικό χρήστη. Η μεταφερσιμότητά του (τρέχει σε μηχανές UNIX αλλά και WIN32), η συμμόρφωση του με το πρότυπο ANSI, και το μικρό του μέγεθος, το κάνει κατάλληλο για embedded συστήματα. Η ταχύτητά του το κάνει καλό υποψήφιο για προβλήματα μεγάλης κλίμακας. Τέλος το billnet διανέμεται δωρεάν, αλλά και ελεύθερα, κάτω από τους όρους της μάλλον δημοφιλούς GPL (General Public Licence).

Α΄.1 Εισαγωγή

Το billnet είναι ένας αρθρωτός νευρωνικός εξομοιωτής. Ο στόχος του είναι η προσομοίωση όσο το δυνατόν περισσότερων αρχιτεκτονικών και αλγορίθμων εκπαίδευσης με τρόπο αποδοτικό και όχι απαιτητικό σε πόρους συστήματος, έτσι ώστε να είναι δυνατή η ταυτόχρονη εφαρμογή του σε εφαρμογές ερευνητικής, βιομηχανικής και εκπαιδευτικής φύσεως.

Το billnet αναπτύχθηκε κυρίως κατά τη διάρκεια της εκπονήσεως της διδακτορικής διατριβής του Βασίλη Βιρβίλη στο Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών στο Δημόκριτο, αλλά τώρα αποτελεί ένα εσωτερικό project του εργαστηρίου Νευρωνικών Δικτύων. Πολλά μέλη του εργαστηρίου έχουν ήδη συνεισφέρει στο billnet με κυριότερη την συμβολή του Δρ. Περαντώνη ο οποίος εκτός από τη χάραξη των βασικών κατευθυντήριων γραμμών, προγραμμάτισε την οικογένεια αλγορίθμων ALECO (Algorithm for Learning Efficiently with Constrained Optimization) σαν μονάδα του billnet. Ο ALECO του billnet παραμένει η μοναδική ελεύθερη υλοποίηση του αλγορίθμου που είναι γνωστή μέχρι σήμερα. Επίσης το billnet είναι δυνατόν να περιέχει μερικούς αλγορίθμους οι οποίοι δεν διατίθενται με μορφή πηγαίου κώδικα, αλλά μόνο σε μορφή εκτελέσιμων δυαδικών αρχείων. Αυτό γίνεται είτε διότι οι σχετικές δημοσιεύσεις που εισάγουν τους αλγόριθμους στην παγκόσμια βιβλιογραφία δεν έχουν ακόμα εκδοθεί, είτε διότι ο αλγόριθμος βασίζεται και σε ιδιόκτητο κώδικα, που η άδεια λειτουργίας του δεν επιτρέπει την αναδιανομή.

Α΄.2 Χαρακτηριστικά

Οι προγραμματιστές σ' ένα ερευνητικό εργαστήριο συχνά είναι αναγκασμένοι να επιδεικνύουν αντιφατική ή στην καλύτερη περίπτωση διττή συμπεριφορά. Στα λεγόμενα προβλήματα πραγματικού κόσμου, τα οποία

θα πρέπει σημειωτέον να λυθούν από απλούς χρήστες και όχι από μύστες της βιβλιογραφίας των νευρωνικών δικτύων, ένας σχετικά υψηλός βαθμός ολοκλήρωσης με ένα γραφικό περιβάλλον (GUI) απαιτείται. Από την άλλη πλευρά, κατά τη διάρκεια ανάπτυξης, σχεδιασμού, και πειραμάτων ο κώδικας που κάνει την εξομοίωση του νευρωνικού δικτύου είναι στην καλύτερη περίπτωση σε 'ρευστή' κατάσταση. Ακόμα χειρότερα, απαραίτητη προϋπόθεση δημοσίευσης σε οποιοδήποτε επιστημονικό περιοδικό είναι οι εκτεταμένες προσομοιώσεις και η σύγκριση με προγενέστερους ή παραπλήσιους αλγορίθμους. Αυτές οι προσομοιώσεις είθισται να συμπεριλαμβάνουν τόσο συνθετικά προβλήματα, όσο και προβλήματα πραγματικού κόσμου.

Για να είναι δυνατόν για έναν νευρωνικό εξομοιωτή να μπορέσει να ανταποκριθεί στις παραπάνω, κάπως αντιφατικές, απαιτήσεις απαιτούνται κάποια χαρακτηριστικά όπως:

Μεταφερσιμότητα (Portability) Η μεταφερσιμότητα είναι αναγκαία σε περιπτώσεις στα οποία η τελική μηχανή που θα τρέξει το πρόγραμμα δεν είναι η ίδια με τη μηχανή όπου αναπτύσσεται το software. Αυτό ισχύει ιδιαίτερα για την περίπτωση εξειδικευμένων μηχανημάτων (embedded systems). Επίσης προγραμματίζοντας με βάση της αρχές της μεταφερσιμότητας είναι συχνά ένας καλός τρόπος για να βρίσκει κανείς λάθη εν τη γενέσει τους (bugs). Η τρέχουσα έκδοση του billnet επιτυγχάνει πλήρη συμμόρφωση με το πρότυπο ANSI της C. Παρ' όλα αυτά το billnet μπορεί να χρησιμοποιήσει ιδιαίτερα χαρακτηριστικά του συστήματος εφ' όσον αυτά υπάρχουν.

Ευελιξία Οι περισσότερες από τις ρυθμίσεις και τις λειτουργίες του billnet δεν χρησιμοποιούνται όταν το billnet χρησιμοποιείται σ' ένα ευρύτερο πλαίσιο ολοκλήρωσης μ' ένα γραφικό περιβάλλον σαν τελικό προϊόν. Η τρέχουσα έκδοση του billnet υποστηρίζει πάνω από 150 παραμέτρους που ρυθμίζουν τη συμπεριφορά του προγράμματος κατά τη διάρκεια της εκτέλεσής του. Είναι φανερό ότι πρέπει να υπάρχει ένας εύχρηστος τρόπος για ένα γραφικό πρόγραμμα να εμπλέκει το billnet με ένα υποσύνολο δυναμικών μεταβλητών και όχι με το πλήρες σύνολο παραμέτρων που υποστηρίζει το billnet.

Το billnet είναι δυνατόν να παραμετροποιηθεί μέσω πολλαπλών αρχείων ρυθμίσεων (configuration files) τα οποία ενδέχεται να είναι και φωλιασμένα (nested), μέσω της γραμμής εντολών (command line) ενώ επιτρέπει και περιορισμένη μάκρο αντικατάσταση μεταβλητών περιβάλλοντος. Το πλήθος των δυνατοτήτων που παρέχει το billnet για τις ρυθμίσεις της λειτουργίας του το κάνει πολύ ευέλικτο και δυναμικό εργαλείο σε συνδυασμό με αρχεία δέσμης (scripts, batch processing) για την εκτέλεση σεναρίων που αποτελούν εκτεταμένες εξομοιώσεις. Πρέπει να υπογραμμιστεί ότι το billnet έχει πολύ λογική αρχικοποίηση των παραμέτρων του έτσι ώστε ο χρήστης να μην υποχρεούται να δηλώνει ευθέως παραμέτρους λειτουργίας που δεν χρειάζεται.

Υψηλή απόδοση Το billnet είναι εξ ολοκλήρου γραμμένο σε C έχοντας λάβει υπόψη του προηγούμενες υλοποιήσεις. Ειδικό βάρος έχει δοθεί και σε χειροκίνητη βελτιστοποίηση όπου κρίθηκε απαραίτητο. Το ίδιο το billnet είναι μικρό και ελαφρύ, αφήνοντας τους πόρους του υπολογιστή για το πραγματικό πρόβλημα. Η ανάγκη αυτή γίνεται όλο και πιο επιτακτική στα embedded systems. Έτσι το billnet αντιμετωπίζει άνετα προβλήματα 20000 διανυσμάτων εισόδου των 40 εισόδων, ενώ το εργαστήριο των νευρωνικών δικτύων έχει χρησιμοποιήσει το billnet και σε επεξεργασία δορυφορικών εικόνων.

Στατιστικές πληροφορίες Είναι συνήθης πρακτική στην ερευνητική κοινότητα των νευρωνικών δικτύων, ένας ερευνητής να υποστηρίζει τα λεγόμενα του με στατιστικές πληροφορίες. Αυτό σημαίνει ότι το billnet πρέπει να έχει την ικανότητα επανεκκίνησης της εκπαιδευτικής διαδικασίας από άλλο αρχικό σημείο στο χώρο των βαρών, και πιθανότατα με άλλη διαμέριση (partition) του συνόλου δεδομένων. Το billnet διαθέτει αρκετές ευκολίες για τη συλλογή χρήσιμων στατιστικών μεγεθών όπως γενικευτική ικανότητα, μέσος όρος απαιτούμενων επαναλήψεων, μέσος όρος απαιτούμενου χρόνου κλπ.

Προγραμματιζόμενες αναφορές (logging) Σε διαφορετικές περιπτώσεις απαιτείται από το billnet και καταγραφή διαφορετικής κάθε φορά πληροφορίας. Το billnet διαθέτει ένα προγραμματιζόμενο πολυκαναλικό υποσύστημα που επιτρέπει στο χρήστη κάθε φορά να διαλέξει τι πληροφορία πρέπει

το billnet να καταγράψει και που να την μεταβιβάσει (σε αρχείο ή σε άλλο πρόγραμμα). Κατά την διάρκεια της εκπαιδεύσεως είναι δυνατό ο χρήστης να κρατάει αρχείο σχετικά με την εξέλιξη των βαρών, του μέσου τετραγωνικού σφάλματος, των εσφαλμένα ταξινομημένων διανυσμάτων, και την τρέχουσα εσωτερική αναπαράσταση του δικτύου.

Ευχρηστία σε αρχεία δέσμης (Scriptability) Αρχεία δέσμης σε ισχυρά περιβάλλοντα όπως το Bourne Shell παρέχουν ένα πολύ δυναμικό τρόπο χρήσης του billnet, πάνω σε πολλαπλά σύνολα δεδομένων, με διαφορετικούς αλγόριθμους. Άλλα πιο εξειδικευμένα σενάρια είναι επίσης δυνατά. Οι δυνατότητες αυτές δεν θα ήταν πραγματικότητα αν το billnet ήταν κατά βάση ένα πρόγραμμα γραφικού περιβάλλοντος. Τα scripts δεν αναμένουν είσοδο ή αλληλεπίδραση από τον χρήστη και γι' αυτό το λόγο μπορούν να σχεδιαστούν ώστε να εκτελούνται τη νύχτα ή άλλες ώρες χωρίς να απασχολούν ανθρώπους. Επίσης το billnet μπορεί να χρησιμοποιηθεί σαν κομμάτι μιας σωλήνωσης, αφού οποιοδήποτε κανάλι του μπορεί να ανακατευθυνθεί στη συνήθη έξοδο (standard output).

Αρθρωτή δομή και ορθογωνιότητα Το σημαντικότερο πιθανότατα χαρακτηριστικό του billnet είναι ότι είναι σχεδιασμένο αρθρωτά (modular) και με το κάθε άρθρωμα να είναι ορθογώνιο σε σχέση με τα υπόλοιπα. Το γεγονός ότι κάθε αλγόριθμος κρατάει τις απαραίτητες σ' αυτόν δομές δεδομένων σε ξεχωριστό από τους άλλους αρχείο επιτρέπει στον προγραμματιστή να συγκεντρωθεί στη ανάπτυξη ενός αλγορίθμου. Αυτό συνήθως σημαίνει ιδιαίτερες και ξεχωριστές βελτιστοποιήσεις για κάθε αλγόριθμο, δηλαδή πιο γρήγορο και πιο καθαρό πηγαίο κώδικα. Οι αλγόριθμοι μπορούν εύκολα να προστεθούν ή να αφαιρεθούν κατά τη διάρκεια της μεταγλώττισης. Έτσι μπορεί κανείς να έχει εξειδικευμένα μικρότερα εκτελέσιμα αρχεία αν ο χώρος είναι μια σημαντική παράμετρος.

Α΄.2.1 Τομείς ρυθμίσεων

Το billnet υποστηρίζει την δυνατότητα πολλαπλών τομέων ρυθμίσεων (billnet sections). Με αυτόν τον τρόπο επιτυγχάνεται η ομαδοποίηση πολλών συγγενών εννοιολογικά παραμέτρων κάτω από το ίδιο τμήμα του αρχείου ρυθμίσεων. Παρακάτω παραθέτουμε περιληπτικά τα σημαντικότερα τμήματα του billnet.

Main Σ' αυτό τον τομέα επιλέγονται οι βασικές ρυθμίσεις λειτουργίας του billnet. Καθορίζεται δηλαδή αν το billnet θα προχωρήσει σε εκπαίδευση του δικτύου ή σε On Line λειτουργία, το πλήθος των επαναλήψεων (epochs) που επιτρέπεται να κάνει, ο αλγόριθμος που θα χρησιμοποιηθεί κλπ.

Stop Εδώ δηλώνονται οι παράμετροι που ρυθμίζουν, τον τρόπο με τον οποίο το billnet σταματάει την εκπαίδευση. Το billnet μπορεί να σταματήσει είτε επειδή έμαθε το δίκτυο, είτε επειδή πέρασε ένα συγκεκριμένο χρονικό διάστημα, είτε επειδή διέκοψε την εκπαίδευση ο χρήστης, είτε επειδή εξαντλήθηκε ο καθορισμένος αριθμός των επαναλήψεων.

Data Εδώ ο χρήστης μπορεί να δηλώσει το αρχείο BDF[1] που περιγράφει το πρόβλημα επιβλεπόμενης εκμάθησης, καθώς και τα διάφορα φίλτρα προεπεξεργαστών που συνήθως απαιτούνται.

Generalization Το τμήμα αυτό περιέχει όλες τις απαραίτητες παραμέτρους, για τον υπολογισμό της γενικευτικής ικανότητας του δικτύου, με στατιστική αξιοπιστία. Έτσι εδώ μπορεί να καθοριστεί ο τρόπος που θα χωρίσει το billnet το σύνολο δεδομένων, σε σύνολα εκπαίδευσης, αξιολόγησης και δοκιμής, καθώς και το πλήθος των διαφορετικών επανεκκινήσεων που απαιτείται.

Output Οι έξοδοι του νευρωνικού δικτύου μπορούν να παρακολουθούνται σε κάθε επανάληψη, ή στο τέλος της εκπαίδευσης. Σε κάθε περίπτωση μπορούν να αναδρομολογηθούν (redirect) με τα κατάλληλα φίλτρα στα κατάλληλα προγράμματα ώστε να έχουμε γραφική απεικόνιση, όπου αυτή είναι επιθυμητή.

[1] billnet Data Format

Channels Ο τομέας αυτός του αρχείου ρυθμίσεων περιλαμβάνει τον πίνακα ανακατευθύνσεων του billnet. Διαφορετικά μέρη του billnet στέλνουν την έξοδο του σε διαφορετικά κανάλια του συστήματος εξόδου του billnet, έτσι ώστε να είναι εύκολο για τον χρήστη να παίρνει μόνο την πληροφορία που επιθυμεί χωρίς να κατακλύζεται από όλη την πληροφορία που μπορεί να έχει. Ένα κανάλι του billnet μπορεί να είναι αρχείο ή πρόγραμμα, ή γενικότερα μια οποιαδήποτε σωλήνωση που υποστηρίζεται από το λειτουργικό σύστημα στο οποίο το billnet εκτελείται.

Α΄.2.2 Αλγόριθμοι

Το billnet υλοποιεί ένα μεγάλο υποσύνολο των διαθέσιμων αλγορίθμων για εκπαίδευση δικτύων εμπρόσθιας διάδοσης. Πιο συγκεκριμένα οι παρακάτω αλγόριθμοι είναι διαθέσιμοι:

- perceptron
- back propagation
- ALECO
  - aleco 0
  - aleco 2
- Rprop (Resilient propagation)
- knn
  - Exhaustive search
  - Fast knn (Nene-Nayar)
- conjugate gradient
  - Polak-Ribiere
  - Fletcher-Reeves
- quickprop
- Delta Bar Delta
  - Delta Bar Delta
  - Extended Delta Bar Delta
  - SuperSAB
- BFGS
- adaptive
- dogleg
- Levenberg-Marquardt
- kmeans
- Kohonen Self Organizing Maps (SOM)
- FLF3

- simplex

Στα δίκτυα εμπρόσθιας διάδοσης το billnet υποστηρίζει κάποιες μεταβλητές που είναι κοινές για τους περισσότερους αλγορίθμους:

NodesPerHiddenLayer Επειδή τα δίκτυα που δημιουργεί το billnet είναι πλήρως συνδεδεμένα (fully connected) αρκεί ο αριθμός των κόμβων σε κάθε επίπεδο για να καθοριστεί η αρχιτεκτονική του εκπαιδευόμενου δικτύου. Το billnet δεν υπαγορεύει κανένα περιορισμό σχετικά με τον αριθμό των επιπέδων, αλλά η πρακτική δείχνει ότι δίκτυα με πάνω από δύο κρυμμένα επίπεδα εκπαιδεύονται δύσκολα.

LearningRate Η πιο σημαντική παράμετρος όλων των αλγορίθμων που τελειώνουν σε prop, και είναι ουσιαστικά παράγωγα ή παραλλαγές του gradient descent.

Momentum Χρησιμοποιείται για την επιτάχυνση της εκπαίδευσης σε περιοχές που η παράγωγος είναι μικρή και δεν αλλάζει πολύ. Ίσως η πρώτη ευριστική παράμετρος από καταβολής νευρωνικών δικτύων.

OnLine Επιλέγει την εκπαίδευσης δέσμης (batch) ή τον υπολογισμό της παραγώγου και την ανανέωση των βαρών ανά διάνυσμα εισόδου. Ο πρώτος τρόπος διευκολύνει την εισαγωγή τεχνικών βελτιστοποίησης (conjugate gradient, bfgs, LM) στα νευρωνικά δίκτυα, ενώ ο δεύτερος διατηρεί ζωντανή την καταγωγή των νευρωνικών δικτύων από τα βιολογικά συστήματα. Πρέπει να υπογραμμιστεί εδώ ότι η μεταβλητή OnLine έχει άλλη ερμηνεία στο πλαίσιο ενός αλγορίθμου, και άλλη στο πλαίσιο ενός προεπεξεργαστή.

SigmoidSteepness Αυτή η παράμετρος είναι ευρύτερα γνωστή σαν β και είναι ο συντελεστής της απόκρισης του δικτύου στον εκθετικό όρο της σιγμοειδούς. Στη φυσική αυτός ό όρος θεωρείται ο αντίστροφος της θερμοκρασίας του συστήματος.

SigmoidPrimeOffset Αυτή η πάραμετρος εισήχθηκε πρώτη φορά από τον Scott Fahlman κατά την ανάπτυξη του quickprop. Ο στόχος της είναι να αυξήσει τη παράγωγο και με αυτό τον τρόπο να επιταχύνει την σύγκλιση.

Α΄.2.3 Προεπεξεργασία δεδομένων

Μέσα στο πακέτο του πηγαίου κώδικα του billnet συμπεριλαμβάνονται και κάποια προγράμματα προεπεξεργασίας δεδομένων. Από αυτά τα δύο πιο σημαντικά είναι:

Normalizer Τα νευρωνικά δίκτυα λόγω της σιγμοειδούς συνάρτησης, που εμφανίζει πολύ μικρή παράγωγο για μεγάλες τιμές του ορίσματός της, είναι σχεδόν αδύνατο να εκπαιδευτούν στη λύση προβλημάτων των οποίων τα δεδομένα δεν είναι κανονικοποιημένα. Αυτό ισχύει τόσο για προβλήματα προσέγγισης συναρτήσεων όσο και για προβλήματα ταξινόμησης και αναγνώρισης προτύπων.

Ο Normalizer του billnet περιμένει σαν είσοδο ένα αρχείο δεδομένων τύπου `BDF` και βγάζει στην έξοδο το κανονικοποιημένο αρχείο. Παρ' όλο που μπορεί να οριστεί ευθέως, ο Normalizer χρησιμοποιεί τη συνήθη είσοδο και έξοδο εξ ορισμού, γεγονός που τον κάνει πολύ χρήσιμο σαν μέρος μεγάλων σωληνώσεων, αλλά και σαν ανεξάρτητο πρόγραμμα. Η κανονικοποίηση που κάνει είναι κάθετης μορφής, που σημαίνει ότι δεν διατηρεί το λόγο μεταξύ των διαφόρων συνιστωσών των διανυσμάτων, παραμορφώνοντας έτσι το πρόβλημα στο χώρο των εισόδων. Παρ' όλα αυτά ο μετασχηματισμός που προξενεί είναι γραμμικός και τις παραμέτρους αυτού του μετασχηματισμού τις κρατάει σε ένα αρχείο (`log file`). Χρησιμοποιώντας την πληροφορία που είναι αποθηκευμένη σ' αυτό το αρχείο ο Normalizer μπορεί είτε να συνεχίσει να κανονικοποιεί άγνωστα δεδομένα (OnLine), ή να εκτελέσει τον αντίστροφο μετασχηματισμό (DeNormalize), παράγοντας έτσι αποκανονικοποιημένα δεδομένα.

salience Το salience είναι ένα πρόγραμμα προεπεξεργασίας δεδομένων. Σαν είσοδο έχει ένα αρχείο δεδομένων τύπου BDF και προαιρετικά ένα ή μια ομάδα αρχείων βαρών (BWF) τα οποία είναι αποτέλεσμα εκπαίδευσης ενός νευρωνικού δικτύου στο δοθέν σύνολο δεδομένων. Το salience υπολογίζει ανάλογα με τον ορισμένο αλγόριθμο τη σημαντικότητα του κάθε χαρακτηριστικού και κρατάει τα σημαντικότερα από αυτά, ανάλογα πάντα με το πως ο χρήστης έχει καθορίσει την πλήρως παραμετροποιήσιμη συμπεριφορά του salience. Στην έξοδο το salience βγάζει ένα καινούριο αρχείο BDF που αντιστοιχεί στα πιο σημαντικά χαρακτηριστικά, πάντα κατά τον χρησιμοποιούμενο αλγόριθμο. Στο salience υλοποιούνται οι κάτωθι αλγόριθμοι επιλογής και εξαγωγής χαρακτηριστικών:

- Η μέθοδος Tarr
- Η μέθοδος Ruck
- Η μέθοδος PCA
- Η μέθοδος t-test
- Η μέθοδος Supervised PCA (SPCA)

Οι προεπεξεργαστές του billnet έχουν σχεδιαστεί έτσι ώστε να δρουν σαν φίλτρα σε αρχεία δεδομένων τύπου BDF. Μια ακόμα σημαντική παρατήρηση είναι ότι κάποιοι προεπεξεργαστές προσφέρουν τη δυνατότητα καταγραφής του μετασχηματισμού που επιτελούν στο αρχείο εισόδου. Αυτό σε συνδυασμό με την επιλογή OnLine, επιτρέπει την προχωρημένη χρήση τους.

Το billnet έρχεται με μερικούς κοινούς, και μερικούς όχι τόσο κοινούς προεπεξεργαστές. Ο πιο απλός τρόπος να εμπλέξει κανείς έναν προεπεξεργαστή είναι να το κάνει ανεξάρτητα από το billnet. Στη συνέχεια μπορεί να εκπαιδεύσει το billnet με την έξοδο του προεπεξεργαστή. Αυτή η λογική συνέχεια μπορεί πολύ εύκολα να περιγραφεί και από το αρχείο ρυθμίσεων του billnet. Πράγματι χρησιμοποιώντας την επιλογή DataChannel μπορεί να ορίσει κανείς ότι τα δεδομένα προέρχονται μέσω σωλήνωσης από τον επιθυμητό προεπεξεργαστή. Αυτό σημαίνει όμως ότι όλες οι πιθανές διαμερίσεις του συνόλου δεδομένων προεπεξεργάζονται χρησιμοποιώντας πληροφορία που βρίσκεται στο πλήρες σύνολο. Αυτό δεν είναι πάντα άσχημο, αλλά υπάρχουν περιπτώσεις, που δημιουργεί ανεπιθύμητα αποτελέσματα (biased effect) κατά το στάδιο της προεπεξεργασίας.

Ας πάρουμε για παράδειγμα την περίπτωση που θέλουμε να κανονικοποιήσουμε τα δεδομένα μας προτού αρχίσουμε την εκπαίδευση του δικτύου. Θα θέλαμε επίσης να επαναλάβουμε το πείραμα πολλές φορές με διαφορετικές κάθε φορά διαμερίσεις (training, validation, testing) ώστε να έχουμε στατιστική αξιοπιστία. Αν κανονικοποιήσουμε ολόκληρο το σύνολο δεδομένων εκτελούμε το πείραμα με λάθος τρόπο.

Αυτό που πρέπει να κάνουμε είναι να περάσουμε μόνο το σύνολο εκπαίδευσης μέσα από τον normalizer. Θα πρέπει να κρατηθούν οι παράμετροι της κανονικοποίησης γιατί με αυτές θα πρέπει να κανονικοποιηθούν τα σύνολα δοκιμής και εκτίμησης. Αυτό μπορεί να γίνει αν ορίσουμε στις επιλογές του προεπεξεργαστή την OnLine λειτουργία. Δηλαδή ο προεπεξεργαστής δεν θα προσπαθεί να υπολογίσει τις παραμέτρους λειτουργίας αλλά αυτές θα δίνονται απ' ευθείας στο αρχείο αναφοράς του.

Με την ολοένα αυξανόμενη χρήση του billnet σε περισσότερα πειράματα αλλά και επιδοτούμενα προγράμματα η ανάγκη για προεπεξεργασία ανά διαμέριση (PPP: Per Partition Preprocessing) έγινε επιτακτική. Δυστυχώς αυτό αυτομάτως απαιτεί κάποιο είδος διεργασιακής επικοινωνίας (IPC: Inter Process Communication), η οποία δεν καθορίζεται από το πρότυπο ANSI. Το billnet υλοποιεί αυτό το τόσο επιθυμητό χαρακτηριστικό τόσο σε POSIX (UNIX) συστήματα όσο και σε μηχανές WIN32. Το billnet παρέχει δικλείδες ασφαλείας ώστε να μπορεί να μεταγλωττιστεί ακόμα και αν η υποβόσκουσα πλατφόρμα δεν υποστηρίζει κανένα από τα δύο σύνολα χαρακτηριστικών.

Α΄.3 Εφαρμογές

Το billnet έχει ήδη χρησιμοποιηθεί σε μια πλειάδα επιδοτούμενων ερευνητικών προγραμμάτων αλλά και δημοσιεύσεων [19, 20, 13, 14, 15, 16, 75, 76, 77, 78, 79, 80], αποδεικνύοντας έτσι την δυνατότητα πολυμορφικής

λειτουργίας που παρέχει στον τελικό χρήστη.

Πιο συγκεκριμένα τα ερευνητικά προγράμματα στα οποία έχει χρησιμοποιθεί εξ' ολοκλήρου ή εν μέρει το billnet είναι:

- BRITE-EURAM – BE-1117 (1996-1998), "GeoNickel" Integrated Technologies for Mineral Exploration: Pilot Project for Nickel Ore Deposits. Συμμετοχή με συνεισφορά του κώδικα για τους supervised feed forward τύπους δικτύων (billnet-0.2).
- ΠΕΝΕΔ (1996-1998), 'Ανάπτυξη Αλγορίθμων Νευρωνικών Δικτύων για τη Μελέτη και Πρόβλεψη Παραμέτρων Πλάσματος σε Αντιδραστήρες Θερμοπυρηνικής Σύντηξης'.
- Πρόγραμμα παροχής υπηρεσιών (1998-1999), 'Προηγμένες μέθοδοι οπτικής αναγνώρισης χαρακτήρων'. Απευθείας ανάθεση προς το ΕΚΕΦΕ 'Δ' από τον Δημοσιογραφικό Οργανισμό Λαμπράκη.
- ΠΑΒΕ 97ΒΕ323 (1998-2000), 'Προγραμματιζόμενη Ευφυής Κάμερα για την Αναγνώριση Υφής' (συνεργασία με την εταιρία ανάπτυξης εμφωλιασμένων (εμβεδδεδ) εφαρμογών ΤΕΣΕΙΚ).
- Πρόγραμμα παροχής υπηρεσιών (1999-2002), 'Μοντελοποίηση δεδομένων και εξαγωγή πληροφοριών με προηγμένες τεχνικές μηχανικής μάθησης'. Απευθείας ανάθεση προς ΕΚΕΦΕ 'Δ' από την Βρεταννική Εταιρεία ερευνών αγοράς 'Millward Brown PLC'.

Α΄.4 Άδεια χρήσης

Το billnet διανέμεται 'ελεύθερα' κάτω από τους όρους της GPL (General Public Licence). Οι λόγοι γι' αυτό είναι αρκετοί και παρατίθενται παρακάτω:

- Η ελεύθερη διακίνηση ιδεών και εργαλείων είναι κοινό χαρακτηριστικό γνώρισμα κάθε επιστημονικής κοινότητας, και έχει συνεισφέρει στο να διατηρηθεί η επιστημονική σκέψη ζωντανή και δημιουργική.
- Η ελπίδα ότι η έκθεση του billnet σε πολλούς υποψήφιους χρήστες αλλά και προγραμματιστές θα βοηθήσει στην εξέλιξη του πακέτου είτε με τη μορφή καλόπιστης κριτικής αναφοράς λαθών (bug report), είτε με την συμβολή νέων αλγορίθμων στην αναπτυσσόμενη βιβλιοθήκη του billnet.
- Είναι εύκολο να αναδιανείμει κανείς αυστηρά πειραματικούς αλγόριθμους, έτσι ώστε και άλλα μέλη της επιστημονικής κοινότητας να μπορούν να πειραματιστούν, συχνά σε εμπιστευτικά δεδομένα, πράγμα που συνήθως καταλήγει σε πολύτιμες αναφορές. Με αυτό τον τρόπο ο αλγόριθμος τυγχάνει μεγαλύτερης αποδοχής πολύ πιο γρήγορα, αφού η επιστημονική κοινότητα εξοικειώνεται μ' αυτόν πιο εύκολα.
- Σε κάποια από τα προβλήματα που θέτει ο Lutz Prechelt [81] προσπαθεί να δώσει λύση το billnet. Πιο συγκεκριμένα ο Prechelt αναφέρει:
  - Οι αλγόριθμοι των άλλων ερευνητών συχνά δεν είναι διαθέσιμοι με μορφή προγράμματος, ή οι υλοποιήσεις τους είναι ασταθείς ή βασισμένες σε μη σύνηθες περιβάλλον.
  - Ακόμα και αποτελέσματα που αφορούν στο ίδιο πρόβλημα συχνά δεν μπορούν να συγκριθούν, ή ακόμα και να αναπαραχθούν εξαιτίας των διαφορετικών πιθανών αναπαραστάσεων ή των διαφορετικών πειραματικών διατάξεων.

  Το billnet δίνει λύση και στα δύο αυτά υπαρκτά προβλήματα. Οι πειραματικές διατάξεις εύκολα περιγράφονται από ένα αρχείο ρυθμίσεων του billnet, ενώ οι αλγόριθμοι που εμπεριέχονται στο billnet είναι συνεπείς ως προς την υλοποίηση τους και ντετερμινιστικοί ως προς την συμπεριφορά τους σε όλες τις πλατφόρμες στις οποίες έχει μεταφερθεί το billnet. Έχοντας τα παραπάνω σαν δεδομένα είναι εύκολο για οποιονδήποτε να αναπαράγει τα πειραματικά αποτελέσματα άλλων ερευνητών και να επιβεβαιώσει ή να διαψεύσει τα αποτελέσματα τους.

Αλγόριθμος	Προγραμματιστής	Αναφορά
back_prop	Β. Βιρβίλης	[82]
perceptron	Β. Βιρβίλης	[28]
aleco	Σ. Περαντώνης	[65]
rprop	Σ. Περαντώνης Ν. Αμπαζής	[83]
knn	Β. Βιρβίλης	[84, 67]
conj grad	Ν. Αμπαζής R. Brahimi	[50, 85, 86]
quick prop	Β. Βιρβίλης	[87]
dbd	Β. Βιρβίλης	[88, 89, 90]
bfgs	Σ. Σπύρου	[91]
adaptive	Β. Βιρβίλης	[92]
flf3	Β. Βιρβίλης	[13, 14]
simplex	Β. Βιρβίλης	[31, 32]
dogleg	Σ. Σπύρου	
lev_mar	Σ. Σπύρου Β. Βιρβίλης	
kmeans	Β. Βιρβίλης	
som	Β. Βιρβίλης	

Πίνακας Α΄.1: Οι αλγόριθμοι του billnet και οι προγραμματιστές τους

- Η ελπίδα ότι το billnet θα συνεχίσει να υπάρχει σαν εσωτερικό πρόγραμμα του εργαστηρίου Νευρωνικών Δικτύων. Τα open source προγράμματα έχουν πολύ καλύτερη τύχη από προγράμματα κλειστού τύπου ανάπτυξης.
- Στον χώρο της μοντέρνας πληροφορικής με την τεράστια ανάπτυξη του ελεύθερου λογισμικού (`www.fsf.org`), η υπάρχουσα τάση βοηθά στο να αποφεύγεται ο άσκοπος διπλασιασμός κώδικα, και ενθαρρύνει την επαναχρησιμοποίηση κώδικα.

Α΄.5 Ευχαριστίες

Το billnet δεν θα μπορούσε να γίνει πραγματικότητα αν δεν υπήρχε ένας συνδυασμός ανθρώπων και καταστάσεων που βοήθησαν και στήριξαν αυτό το project.

Ο Δρ. Σ. Περαντώνης πέρα από την προγραμματιστική του συμβολή, σχεδίασε και την βασική κατευθυντήρια γραμμή γύρω από την οποία χτίστηκε όλο το οικοδόμημα. Έχοντας σαν βάση το επιθυμητό σύνολο χαρακτηριστικών είναι πολύ πιο εύκολο να σχεδιάσει κανείς ευέλικτα και αρθρωτά υποσυστήματα που απαρτίζουν τη συνολική λύση.

Ο Δρ. Κ. Κοντοβασίλης με τις πολύτιμες συμβουλές του, σχεδόν σε κάθε περιοχή της πληροφορικής, αποτέλεσε θεμέλιο λίθο σ' όλες τις καλές ιδιότητες της αρχιτεκτονικής του billnet. Χωρίς αυτόν κάθε έννοια μεταφερσιμότητας που το billnet σήμερα απολαμβάνει, πολύ απλά δεν θα υπήρχε.

Επειδή έχουν συνεισφέρει και άλλοι με τη συγγραφή ενός νέου αλγορίθμου, στον πίνακα Α΄.1 παρατίθεται για λόγους πληρότητας το αρχείο AUTHORS όπως διανέμεται μαζί τον πηγαίο κώδικα του billnet.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] B. Widrow and M. Hoff. Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, volume 4, pages 96–104, 1960. Reprinted in J. A. Anderson and Rosenfeld (Eds.) *Neurocomputing: Foundations of Research.* Cambridge: MIT Press, 1988.

[2] H. D. Block. The perceptron: A model for brain functioning. *Reviews of Modern Physics*, 34:123–135, 1962. Reprinted in J. A. Anderson and Rosenfeld (Eds.) *Neurocomputing: Foundations of Research.* Cambridge: MIT Press, 1988.

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

[4] F. Rosenblatt. The perceptrons. A probabilistic model for information storage and organization in the brain. *Psycological Review*, 65, 1958.

[5] B. S. Wittner and J. S. Denker. Strategies for teaching layered networks classification tasks. In *Neural Information Processing Systems*, pages 850–859, Denver, 1997.

[6] S. D. Hunt and J. R. Deller. Selective training of feedforward artificial neural networks using matrix perturbation theory. *Neural Networks*, 8:931–944, 1995.

[7] S. Ergenziger and E. Thompsen. An accelerated learning algorithm for multilayer perceptrons: optimization layer by layer. *IEEE Transactions on Neural Networks*, 6:31–42, 1995.

[8] T. Grossman, R. Meir, and E. Domany. Learning by choice of internal representations. *Advances in Neural Information Processing Systems*, 1:73–80, 1989.

[9] H. Takahashi, E. Tomita, and Kawabata. T. Separability of internal representations in multilayer perceptrons with application to learning. *Neural Networks*, 6:689–703, 1993.

[10] E. Oja. A simplified neuron model as a principal component analyzer. *Journal Mathematical Biology*, 15:267–273, 1982.

[11] T. Sanger. Optimal unsupervised learning in a single-layer-linear feedforward neural network. *Neural Networks*, 2:459–473, 1989.

[12] E. Oja. Neural networks, principle components, and subspaces. *International Journal Neural Networks*, 1:61–68, 1989.

[13] S. J. Perantonis and V. Virvilis. Efficient linear discriminant analysis using a fast quadratic programming algorithm. In *International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling*, pages 164–169, Leuven, Belgium, 1998.

[14] S. J. Perantonis and V. Virvilis. Efficient perceptron learning using constrained steepest descent. *Neural Networks*, 13(3):351–364, 2000.

[15] S. J. Perantonis, V. Virvilis, and N. Ampazis. Recent advances in neural network training using constrained optimization. In *Proceedings of 4th International Conference on Applied Mathematical Programming and Modeling APMOD98*, Limassol, Cyprus, 1998.

[16] S. J. Perantonis, N. Ampazis, and V. Virvilis. A learning framework for neural networks using constrained optimization methods. In *Annals of Operations Research*, volume 99, pages 385–401, 2000.

[17] S. J. Perantonis, V. Virvilis, Ch. Papageorgiou, and A. Rabavilas. Neural network based parameter selection for ERP analysis. In *Proceedings of X World Congress of Psychiatry*, volume I, page 178, 1996.

[18] N. Vassilas, S. J. Perantonis, V. Virvilis, Ch. Papageorgiou, A. Rabavilas, and C. Stefanis. Erp classification using neural network based feature selection and multiple classifier models. Accepted for publication in Technology and Health Care.

[19] S. J. Perantonis and V. Virvilis. Dimensionality reduction using a novel neural network based feature extraction method. Washington, DC, 1999. Presented at International Joint Conference on Neural Networks. Best presentation award.

[20] S. J. Perantonis and V. Virvilis. Input feature extraction for multilayer perceptrons using supervised principal component analysis. *Neural Processing Letters*, 10(3):243–252, 1999.

[21] A. Agogino, J. Ghosh, S. J. Perantonis, V. Virvilis, S. Petridis, and Lisboa P. J. G. The role of multiple linear-projection based visualization techniques in RBF based classification of high dimensional data. In *Proceedings of IEEE-INNS-ENNS IJCNN2000*, volume 3, Como, Italy, 2000.

[22] S. J. Perantonis, S. Petridis, and V. Virvilis. Supervised principal component analysis using a smooth classifier paradigm. In *Proceedings of International Conference on Pattern Recognition*, number 1572 in 0, Barcelona, Spain, 2000.

[23] J. Rourke. *Computational Geometry in C.* Cambirdge University Press, 1993.

[24] B. A. Telfer and D. P. Casasent. Minimum-cost associative processor for piecewise-hyperspherical classification. *Neural Networks*, 6:1117–1130, 1993.

[25] E. Barnard and D. Casasent. A comparison between criterion functions for linear classifiers, with an application to neural nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19:1030–1041, 1989.

[26] E. Barnard. Performance and generalization of the classification figure of merit criterion function. *IEEE Transactions on Neural Networks*, 2:322–325, 1991.

[27] Gilbert Strang. *Linear Algebra and its Applications.* Harcourt Brace Jovanovich, 1988.

[28] F. Rosenblatt. *Principles of Neurodynamics.* Spartan Books, Washington DC, 1962.

[29] R. P. Gorman and T. J. Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1:75–89, 1988.

[30] R. P. Gorman and T. J. Sejnowski. Learned classification of sonar targets using a massively parallel network. *IEEE Transactions on Neural Networks*, 36:1135–1140, 1988.

[31] G. B. Dantzig, A. Orden, and P. Wolfe. Generalized simplex method for minimizing a linear form under linear inequality restraints. *Pacific J. Math*, 5:183–195, 1955.

[32] G. B. Dantzig. *Linear Programming and Extensions.* Princeton University Press, New Jersey, 1963.

[33] V. Klee and G. J. Minty. How good is the simplex algorithm? In O. Shisha, editor, *Inequalities III*, pages 159–175. Academic Press, New York, 1972.

[34] L. Bobrowski and W. Niemiro. A method of synthesis of linear discriminant function in the case of nonseparability. *Pattern Recognition*, 17:205–210, 1984.

[35] J. S. Pang. Methods for quadratic programming: a survey. *Computers and Chemical Engineering*, 7:583–594, 1983.

[36] N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4:373–395, 1984.

[37] G. Zoutendijk. *Methods of Feasible Directions.* Elsevier, 1960.

[38] J. B. Rosen. The gradient projection method for non linear programming. part I: linear constraints. *SIAM Journal Appl. Math.*, 8:181–217, 1960.

[39] J. B. Rosen. The gradient projection method for non linear programming. part II: nonlinear constraints. *SIAM Journal Appl. Math.*, 9:514–532, 1961.

[40] M. Bazaraa and C. Shetty. *Nonlinear Programming.* Wiley and Sons, 1979.

[41] St. Nash and A. Sofer. *Linear and Nonlinear Programming.* McGraw Hill International Editions, 1996.

[42] S. S. Rao. *Optimization: Theory and applications.* Wiley Eastern, 1984.

[43] Ph. Gill, W. Murray, and M .H. Wright. *Practical Optimization.* Harcourt Brace and Company, 1981.

[44] D. A. Pierre. *Optimization Theory with Applications.* Wiley, 1969.

[45] G. Zoutendijk. Non linear programming: A numerical survey. *SIAM Journal Control*, 4:194–210, 1966.

[46] D. J. Volper and S. E. Hampson. Quadratic function nodes: Use, structure and training. *Neural Networks*, 3:93–107, 1990.

[47] L. Bobrowski. Design of piecewise linear classifiers from formal neurons by a basis exchange technique. *Pattern Recognition*, 24:863–870, 1991.

[48] V. G. Sigillito, S. P. Wing, and K. B. Baker. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10:262–266, 1989.

[49] S. J. Perantonis and P. J. G. Lisboa. Invariant pattern recognition using higher-order networks and moment classifiers. *IEEE Transactions on Neural Networks*, 3:241–251, 1992.

[50] E. M. Johansson, F. U. Dowla, and D. M. Goodman. Backpropagation learning for multilayer feedforward networks using the conjugate gradient method. *International Journal of Neural Systems*, 2:291–301, 1992.

[51] J. M. Torres Moreno and M. B. Gordon. Characterization of the sonar signals benchmark. *Neural Processing Letters*, 7:1–4, 1998.

[52] D. W. Ruck, S. K. Rogers, and M. Kabrisky. Feature selection using a multilayer perceptron. *Neural Network Computing*, 2:40–48, 1990.

[53] K. Karhunen. Ueber lineare Methoden in der Wahrscheinlichkeitsrechnung. *Annales Academiae Scientiarum Fennicae*, 37:3–79, 1947.

[54] M. Loéve. *Probability Theory.* Van Nostrand, New York, 1963. 3rd ed.

[55] R. W. Preisendorfer. *Principal component analysis in Meteorology and Oceanography.* Elsevier, New York, 1988.

[56] P. Földiak. Adaptive network for optimal linear feature extractors. In *International Joint Conference on Neural Networks*, volume 1, pages 401–405, Washington, DC, 1989.

[57] S. Y. Kung and K. I. Diamantaras. A neural network learning algorithm for adaptive principal component extraction (APEX). In *International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 861–864, Albuquerque, NM, 1990.

[58] H. Chen and R. W. Liu. Adaptive distributed orthogonalization processing for principal components analysis. In *International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 293–296, San Francisco, CA, 1992.

[59] P. J. Huber. Projection pursuit. *Annals of Statistics*, 13:435–475, 1999.

[60] L. M. Belue and K. W. Bauer Jr. Determining input features for multilayered perceptrons. *Neurocomputing*, 7:111–121, 1995.

[61] G. Tarr. *Multilayered Feedforward Networks for Image Segmentation.* PhD thesis, Air Force Institute of Technology, 1991.

[62] P. M. Murphy and D. W. Aha. UCI repository of machine learning databases, machine readable data repository. Technical report, University of California, Department of Information and Computer Science, Irvine, CA, 1992.

[63] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Symposium on Computer Applications and Medical Care*, pages 261–265. IEEE Computer Society Press, 1988.

[64] M. Kendall and A. Stuart. *The Advanced Theory of Statistics.* Griffin, London, 1977. 4th edition.

[65] S. J. Perantonis and D. A. Karras. An efficient constrained learning algorithm with momentum acceleration. *Neural Networks*, 8(2):237–239, 1995.

[66] K. Fukunaga and M. N. Patrenahalli. A branch and bound algorithm for computing k nearest neighbors. *IEEE Transactions on Computers*, C-24:750–753, 1975.

[67] S. A. Nene and S. K. Nayar. A simple algorithm for nearest neighbor search in high dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9):989–1003, September 1997.

[68] J. McNames. A nearest trajectory strategy for time series prediction. In *International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling*, pages 112–128, Leuven, Belgium, 1998.

[69] G. Hinton. Connectionist learning procedures. *Artificial Intelligence*, 40:185–234, 1986.

[70] H. Bischof, A. Pinz, and W. G. Kropatsch. Visualization methods for neural networks. In *Proceedings of the 11th International Conference on Pattern Recognition*, pages 581–585, Netherlands, 1992. The Hague.

[71] J. Wejchert and G. Tesauro. Visualizing process in neural networks. *IBM Journal of Research and Development*, 35:244–253, 1991.

[72] G. Whittington and T. Spracklen. Applying visualization techniques to the development of real world artificial neural network applications. In *Proceedings of the Applications of Artificial Networks III*, pages 1024–1033, Orlando, FL, 1992.

[73] A. K. Agogino. Interactive visualization of radial basis function networks. Master Thesis, University of Texas at Austin, 1999.

[74] A. K. Agogino, J. Gosh, and M. Cheryl. Visualization of radial basis function networks. Accepted for publication in IJCNN, 1999.

[75] V. Virvilis. *Finite Training in Single Layer Perceptrons and Feature Extraction.* PhD thesis, University of Athens Department of Informatics, 1999. In greek.

[76] N. Ampazis, S. J. Perantonis, and J. G. Taylor. Dynamics of multilayer networks in the vicinity of temporary minima. *Neural Networks*, 12 (1):43–58, 1999.

[77] N. Ampazis, S. J. Perantonis, and J. G. Taylor. A dynamical model for the analysis and acceleration of learning in feedforward networks. Under review in Neural Networks.

[78] N. Ampazis, S. J. Perantonis, and J. G. Taylor. Acceleration of learning in feedforward networks using dynamical systems analysis and matrix perturbation theory. Washington, DC, 1999. Presented at International Joint Conference on Neural Networks.

[79] N. Ampazis and S. J. Perantonis. Levenberg-marquardt algorithm with adaptive momentum for the efficient training of feedforward networks. Como, Italy, 2000. Submitted to IJCNN' 2000.

[80] S. J. Perantonis, N Ampazis, and S. Spirou. Training feedforward neural networks with the dogleg method and bfgs hessian updates. Como, Italy, 2000. Submitted to IJCNN' 2000.

[81] L. Prechelt. Proben1 – a set of benchmarks and benchmarking rules for neural network training algorithms. Technical report, Faculty for Informatics, University of Karlsruhe, 1994.

[82] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, chapter 8, pages 318–362. MIT Press, Cambridge, MA, 1986.

[83] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Proceedings of the International Joint Conference on Neural Networks*, volume 1, pages 586–591, Denver, 1993.

[84] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Informations Theory*, 13:21–27, January 1967.

[85] J. C. Gilbert and J. Nocedal. Global convergence properties of conjugate gradient methods for optimization. *SIAM Journal on Optimization*, 2(1):21–42, 1992.

[86] J. Nocedal. Theory of algorithms for unconstrained optimization. *Acta Numerica*, 1:199–242, 1992.

[87] S. E. Fahlman. Faster learning variations on back propagation: An empirical study. In *Proceedings of the 1988 Connectionist Models Summer School*, pages 38–51, San Mateo, 1988. Morgan Kaufmann.

[88] R. A. Jacobs. Increased rates of convergence through learning rate adaptation. *Neural Networks*, 1:295–307, 1988.

[89] A. A. Minai and R. D. Williams. Back propagation heuristics: A study of the extended delta-bar-delta algorithm. In *Proc. of the IJCNN'90*, volume 1, pages 595–600, San Diego, 1990.

[90] T. Tollenaere. SuperSAB: Fast adaptive back propagation with good scaling properties. *Neural Networks*, 3:561–573, 1990.

[91] J. E. Jr Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM, 1996.

[92] T. P. Vogl, J. K. Mangis, A. K. Rigler, W. T. Zink, and D. L. Alkon. Acceleration of the convergence of the back propagation method. *Biological Cybernetics*, 59:257–263, 1988.